

ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА  
ЧАЧАК

|                     |      |        |          |
|---------------------|------|--------|----------|
| ПРИМЉЕНО 24.03.2025 |      |        |          |
| Орг.јед.            | Број | Прилог | Вредност |
|                     | 649  |        |          |



УНИВЕРЗИТЕТ У КРАГУЈЕВЦУ  
ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА У ЧАЧКУ

Слађана М. Ђурашевић Пешовић

**МОДЕЛ ПРЕДИКЦИЈЕ ОТКАЗА  
МАГНЕТНИХ ДИСКОВА ЗАСНОВАН НА  
ДЕТЕКЦИЈИ АНОМАЛИЈА**

докторска дисертација

Чачак, 2025



UNIVERSITY OF KRAGUJEVAC  
FACULTY OF TECHNICAL SCIENCES ČAČAK

Sladana M. Đurašević Pešović

**PREDICTION MODEL FOR  
HARD DISK DRIVE FAILURES  
BASED ON ANOMALY DETECTION**

Doctoral Dissertation

Čačak, 2025

## ИДЕНТИФИКАЦИОНА СТРАНИЦА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

| <b>Аутор</b>                                                                                                                                                                                                       |  |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|
| Име и презиме: Слађана Ђурашевић Пешовић                                                                                                                                                                           |  |
| Датум и место рођења: 20. 4. 1985. године, Ивањица                                                                                                                                                                 |  |
| Садашње запослење: истраживач сарадник, Универзитет у Крагујевцу, Факултет техничких наука у Чачку                                                                                                                 |  |
| <b>Докторска дисертација</b>                                                                                                                                                                                       |  |
| Наслов: Модел предикције отказа магнетних дискова заснован на детекцији аномалија                                                                                                                                  |  |
| Број страница: 115                                                                                                                                                                                                 |  |
| Број слика: 68                                                                                                                                                                                                     |  |
| Број табела: 20                                                                                                                                                                                                    |  |
| Број библиографских података: 130                                                                                                                                                                                  |  |
| Установа и место где је рад израђен: Универзитет у Крагујевцу, Факултет техничких наука у Чачку                                                                                                                    |  |
| Научна област (УДК): Техничко технолошке науке<br>Електротехничко и рачунарско инжењерство<br>Рачунарска техника<br>Информационе технологије и системи                                                             |  |
| <b>Коментори:</b><br>др Вања Луковић, ванредни професор, Факултет техничких наука у Чачку, Универзитет у Крагујевцу,<br>др Борислав Ђорђевић, виши научни сарадник, Институт Михајло Пупин, Универзитет у Београду |  |
| Број и датум одлуке Већа универзитета о прихватању теме докторске дисертације: Одлука Већа за техничко-технолошке науке, Универзитета у Крагујевцу, број: IV-04-884/11 од 10.11.2021. године.                      |  |

## **Захвалница**

*Докторска дисертација је настала као резултат научно истраживачког рада у Лабораторији за напредне рачунарске и информационе технологије, на Факултету техничких наука у Чачку. Захвалност дугујем руководиоцу пројекта др Горану Димићу са Института „Михајло Пупин“ који ми је омогућио да постанем део истраживачког тима на пројекту TR32043 „Развој и моделовање енергетско ефикасних, адаптивних, вишепроцесорских и вишесензорских електронских система мале снаге“ финансираног од стране Министарства просвете, науке и технолошког развоја Републике Србије. Овом приликом желим да изразим захвалност Факултету техничких наука у Чачку на којем су спроведена истраживања и ресорном Министарству које је финансијски подржало мој научно истраживачки рад свих претходних година.*

*Захвалност упућујем менторима др Вањи Луковић и др Бориславу Ђорђевићу, на свим саветима, коментарима и сугестијама који су допринели да ова дисертација буде што квалитетније написана.*

*Такође, захваљујем се свим члановима комисије који су издвојили драгоцено време и учествовали у поступку пријаве теме, као и комисији за оцену и одбрану докторске дисертације на свом доприносу. Захвална сам и свим професорима и колегама, који су ме подржали у току докторских студија.*

*Највећу и неизмерну захвалност дугујем супругу др Урошу Пешовићу на несебичној подршци, разумевању и љубави којом ми је помогао да остварим свој циљ, коме и посвећујем ову дисертацију!*

*Слађана Ђурашевић Петровић*

## Сажетак

Човечанство својим активностима генерише огромне количине дигиталних података чија се величина из године у годину убрзано повећава. Већина ових података данас специјализованим објектима центара података у којима се налазе рачунарски системи који се користе за дељење апликација и података. У њима се за складиштење података првенствено користе магнетни дискове на којима је ускладиштено око 90% укупне количине података, док се преосталих 10% података чува на полупроводничким дисковима. Главна предност магнетних дискова је њихов велики капацитет и ниска цена ускладиштених података, што их чини идеалним за опште складиштење података и резервне копије података. Електромеханички дизајн магнетних дискова чини их подложнијим отказима у односу на остале компоненте рачунарског система. Као последица отказа магнетног диска најчешће долази до губитка корисничких података, чија економска вредност значајно надмашује цену самог диска.

Употребом SMART (енг. *Self-Monitoring, Analysis and Reporting Technology*) технологије, рачунарски систем је у стању да упозори корисника ако је неки од радних параметара диска одступи од унапред дефинисане вредност прага. Методе машинског учења користе предност зависности између више SMART атрибута како би се побољшала стопа предвиђања отказа дискова. У овој докторској дисертацији представљен је модел предикције отказа дискова који је заснован на методи детекције аномалија. Рад диска може бити представљен у вишедимензионалном простору као низ тачака чија је позиција дефинисана вредностима његових SMART атрибута. Тачке података које описују регуларни рад диска ће имати тенденцију да се групишу око одређене тачке дефинисане као средња вредност или центар масе, док ће тачке података са неисправних дискова обично бити расуте даље око овог центра масе. Употребом Махаланобисовог растојања, могуће је измерити удаљеност, изражену у стандардним одступањима, тачке података у вишедимензионалном простору у односу на центар масе, чиме се уклања утицај скалирања и корелације између атрибута. Коришћењем подесиве границе одлучивања обучен је модел за детекцију аномалија тако да предвиди отказе са највећом могућом стопом детекције, при чему минимизира број лажних детекција.

Реализовани модел тестиран је на CMRR скупу података тако што је методом рекурзивне елиминације атрибута креиран оптимални скуп од седам најзначајнијих атрибута. На поменутом скупу података на десет насумичних тестова остварена је просечна стопа детекције отказа од 96.11% , при чему није било погрешних детекција отказа. Предложени модел је предвидео више од 80% кварова, 24 сата пре њиховог стварног настанка, што омогућава благовремену израду резервне копије података што оправдава његову примену у пракси. Модел за детекцију аномалија тестиран је и на полупроводничким дисковима на којима је такође постигао висок ниво предиктивних перформанси.

**Кључне речи:** магнетни дискови, SMART атрибути, детекција аномалија, предикција отказа, Махаланобисово растојање

## Summary

Humanity generates enormous amounts of digital data through its activities, the size of which is increasing rapidly from year to year. Most of this data is now stored in specialized data center facilities that house computer systems used to share applications and data. Magnetic disks are primarily used for data storage, with about 90% of the total amount of data stored, while the remaining 10% of the data is stored on semiconductor disks. The main advantage of magnetic disks is their large capacity and low cost of stored data, which makes them ideal for general data storage and data backup. The electromechanical design of magnetic disks makes them more susceptible to failure than other components of a computer system. The failure of a magnetic disk most often leads to the permanent loss of user data, the economic value of which significantly exceeds the cost of the magnetic disk itself.

Using SMART (Self-Monitoring, Analysis and Reporting Technology) technology, a computer system can alert the user if any of the disk's operating parameters deviate from a predefined threshold value. Machine learning methods take advantage of the dependencies between multiple SMART attributes to improve the disk failure prediction rate. This doctoral dissertation presents a disk failure prediction model based on the anomaly detection method. The operation of a disk can be presented in a multidimensional space with a series of points whose positions are defined by the values of its SMART attributes. Data points describing regular disk operation will tend to cluster around a certain point defined as the mean value or center of mass, while data points from failed disks will usually be scattered further around this center of mass. Using the Mahalanobis distance, it is possible to measure the distance of data points in a multidimensional space from a point representing the center of mass expressed in standard deviations, thereby removing the influence of scaling and correlation between attributes. Using an adjustable decision boundary, an anomaly detection model was trained to predict failures with the highest possible detection rate, while minimizing the number of false detections.

The implemented model was tested on the CMRR dataset by creating an optimal set of the seven most significant attributes using the recursive attribute elimination method. On the aforementioned dataset, an average failure detection rate of 96.11% was achieved in ten random tests, with no false detections of failures. The proposed model predicted more than 80% of failures 24 hours before their actual occurrence, which allows for timely data backup, which justifies its application in practice. The anomaly detection model was also tested on semiconductor disks, where it also achieved a high level of predictive performance.

**Keywords:** magnetic disks, SMART attributes, anomaly detection, failure prediction, Mahalanobis distance

# САДРЖАЈ

|          |                                                                                       |           |
|----------|---------------------------------------------------------------------------------------|-----------|
| <b>1</b> | <b>Увод.....</b>                                                                      | <b>1</b>  |
| 1.1      | Предмет истраживања .....                                                             | 1         |
| 1.2      | Циљ истраживања .....                                                                 | 3         |
| 1.3      | Полазне хипотезе.....                                                                 | 3         |
| 1.4      | Методе и методологије истраживања .....                                               | 4         |
| 1.5      | Научни доприноси истраживања .....                                                    | 5         |
| 1.6      | Преглед досадашњих истраживања.....                                                   | 5         |
| 1.7      | Преглед дисертације по поглављима .....                                               | 6         |
| <b>2</b> | <b>Алгоритми за предикцију отказа магнетних дискова .....</b>                         | <b>8</b>  |
| 2.1      | Технологије трајног складиштења података .....                                        | 8         |
| 2.2      | Магнетни дискови .....                                                                | 10        |
| 2.3      | Класификација грешака код магнетних дискова .....                                     | 16        |
| 2.4      | Технологија самонадзорне анализе и извештавања SMART .....                            | 20        |
| 2.5      | RAID систем независних редундантних дискова.....                                      | 23        |
| 2.6      | Предвиђање отказа магнетних дискова .....                                             | 24        |
| 2.7      | Машинско учење .....                                                                  | 24        |
| 2.8      | Фазе у процесу машинског учења .....                                                  | 28        |
| 2.9      | Процена перформанси предвиђања .....                                                  | 31        |
| 2.10     | Преглед модела за предикцију отказа магнетног диска .....                             | 35        |
| <b>3</b> | <b>Прикупљање и обрада скупова података .....</b>                                     | <b>39</b> |
| 3.1      | Прикупљање података .....                                                             | 39        |
| 3.2      | Обрада скупа података.....                                                            | 43        |
| 3.3      | Технике за балансирање скупа података .....                                           | 46        |
| <b>4</b> | <b>Селекција значајних атрибута .....</b>                                             | <b>47</b> |
| 4.1      | Алгоритми за селекцију атрибута.....                                                  | 47        |
| 4.2      | Алгоритам рекурзивне елиминације атрибута .....                                       | 49        |
| <b>5</b> | <b>Трансформација непараметарски дистрибуираних података и мерење удаљености.....</b> | <b>58</b> |
| 5.1      | Дистрибуција података.....                                                            | 58        |
| 5.2      | Трансформација података .....                                                         | 60        |

|          |                                                                                           |            |
|----------|-------------------------------------------------------------------------------------------|------------|
| 5.3      | Мерење растојања у вишедимензионалном простору .....                                      | 62         |
| <b>6</b> | <b>Детекција аномалија .....</b>                                                          | <b>67</b>  |
| 6.1      | Класификација аномалија.....                                                              | 67         |
| 6.2      | Алгоритми за детекцију аномалија.....                                                     | 69         |
| <b>7</b> | <b>Модел за предикцију отказа магнетних дискова заснован на детекцији аномалија .....</b> | <b>74</b>  |
| 7.1      | Структура предложеног модела.....                                                         | 74         |
| 7.2      | Примена реализованог модела у другим областима истраживања .....                          | 80         |
| <b>8</b> | <b>Представљање резултата истраживања и дискусија .....</b>                               | <b>82</b>  |
| 8.1      | Приказ резултата предикције отказа магнетних дискова.....                                 | 82         |
| 8.2      | Поређење добијених резултата са постојећим истраживањима .....                            | 94         |
| 8.3      | Резултати примене модела за предикцију отказа на полупроводничке SSD дискове.....         | 96         |
| <b>9</b> | <b>Закључна разматрања .....</b>                                                          | <b>99</b>  |
|          | <b>Литература .....</b>                                                                   | <b>102</b> |
|          | <b>Списак слика.....</b>                                                                  | <b>112</b> |
|          | <b>Списак табела .....</b>                                                                | <b>115</b> |

# 1 Увод

У уводном поглављу дефинисан је предмет истраживања докторске дисертације у области предикције отказа магнетних дискова. На основу постављених полазних хипотеза дефинисан је циљ истраживања и методе које су се користиле током истраживања. У наставку приказана је анализа досадашњих истраживања везаних за предикцију отказа и дат је кратак осврт на структуру рада.

## 1.1 Предмет истраживања

Савремене дигиталне технологије генеришу огромне количине података чији је раст у последњој деценији прати експоненцијални тренд. Различити сектори људске делатности као што су индустрија, здравство, образовање, уметничко стваралаштво се све више ослањају на складиштење великих количина података. Поред тога, савремени концепти попут IoT-а (енг. *Internet of Things*) [1], рачунарства у облаку (енг. *Cloud Computing*) [2] и технологије „уланчаних блокова“ (енг. *blockchain*) [3] генеришу велике количине података за које су потребни робусни системи за складиштење који захтевају обраду података и чување односно архивирање у реалном времену. Укупна количина података креираних у свету 2024. године, износила је 147 ZB ( $1\text{ ZB} = 10^{21}\text{ B}$ ), док се 2025. године очекује да ће генерисана количина података износити 182 ZB. Поређења ради, то представља десетоструко повећање количине података у односу на 2014. годину у којој је генерисано 12.5 ZB и 2015. годину са 15.5 ZB. Пројекције за 2028. годину предвиђају укупну генерисану количину података од 394 ZB. Ови подаци обухватају све податке који су генерисани, сачувани, копирани и коришћени [4].

Растућа глобална потражња за складиштењем података у различитим секторима ослања се на коришћење специјализованих простора за складиштење података који се називају центри података (енг. *data centers*). Потражња за складиштењем података бележи константан раст. У 2020. години произведено је медијума за складиштење података укупног капацитета 1.4 ZB [5], од чега су магнетни дискови чинили 1 ZB (70%), магнетне траке 0.15 ZB (10%) и полупроводнички дискови 0.25 ZB (20%). На основу пројекција предвиђа се да ће у 2025. години бити произведено дискова укупног капацитета 2 ZB, при чему ће удео магнетних дискова бити 1.3 ZB, што је приближно 60% испорученог капацитета.

Магнетни дискови или тзв. хард дискови HDD (енг. *Hard Disk Drives*) [6] користе ротирајуће плоче пресвучене површинским слојем феромагнетног материјала за трајно складиштење података. С обзиром да су се појавили далеке 1956. године они се и данас најчешће користе у рачунарским системима за трајно чување података. Када се упореде са NAND FLASH меморијама [6], магнетни дискови пружају већи капацитет уз знатно нижу цену по ускладиштеном бајту. Захваљујући развоју нових технологија наставиће се са даљим повећањем густине записа података (количина података ускладиштена по јединици површине диска) и очекује се да ће магнетни дискови задржати своју ценовну предност и у наредних десет година у односу на полупроводничке SSD (енг. *Solid State Drives*) дискове.

Магнетни дискови, иако су изграђени са високим степеном прецизности услед своје електромеханичке конструкције подложнији су отказима у односу на остале компоненте рачунарског система. Отказ диска представља ситуацију у којој магнетни диск не може

остварити захтевану функцију поузданог складиштења података и најчешће доводи до трајног губитка ускладиштених података. Трошкови губитка података који настају услед отказа магнетног диска у већини случајева вишеструко превазилазе цену самог диска. Поступак опоравка података са отказалих дискова је могућ у специјализованим лабораторијама али је изузетно скуп и примењује се у случају губитка јако битних података. Типична стопа отказа магнетних дискова изражена на годишњем нивоу је реда од 1 до 3 % [7]. На основу овог податка може се закључити да ће у популацији од сто магнетних дискова који раде током једне године бити између један и три отказала диска.

Отказ магнетног диска може настати као последица физичких дефеката електромеханичких склопова магнетног диска или услед оштећења слоја феромагнетног материјала који се користи за складиштење података. Физички дефекти могу се јавити као резултат деловања спољашњих сила (вибрације, утицаји средине, неправилно руковање и физички дефекти), чија је појава изненадна. Утицаји наведених сила се не могу предвидети, а самим тим и отказ магнетног диска који настаје као последица њиховог деловања. У савременим центрима података тежи се да се утицај ових спољашњих сила сведе на најмању могућу меру, како би се повећала поузданост складиштења података. Поред тога, физички дефекти настају постепено узроковани хабањем електромеханичких склопова у магнетном диску, који временом деградирају перформансе диска.

Постоје две технике приступа које се користе при сервисирању отказа магнетног диска: реактивни или проактивни приступ [8].

1. Код реактивног приступа долази до реаговања тек када дође до отказа магнетног диска. Овај приступ се примењује када магнетни дискови формирају редундантни низ независних дискова познатији као RAID (енг. *Redundant Array of Independent Disks*) [9]. RAID низови су отпорни на губитак података услед отказа неког диска у низу.
2. Друга техника је проактивни приступ који тежи да предвиди отказ магнетног диска у блиској будућности на основу одређених показатеља перформанси рада диска. Овакав приступ омогућава кориснику да благовремено направи резервну копију пре настанка отказа диска и спречи губитак података ускладиштених на диску. Проактиван приступ се користи код RAID низова, како би се правовременим реаговањем повећала безбедност сачуваних података за разлику од реактивног приступа [10].

С обзиром да се из године у годину повећава количина и важност ускладиштених података, област предикције отказа дискова је актуелна у области рачунарске технике, како у великим центрима података тако и код индивидуалних корисника. Осим код магнетних дискова, предикција отказа може се примењивати и на полупроводничке SSD дискове и друге компоненте рачунарског система. Ова област истраживања подељена је у следеће категорије:

1. предикција отказа магнетног диска,
2. прогнозирање преосталог времена рада магнетног диска RUL (енг. *Remaining Useful Time*) [11] и
3. класификација тренутног статуса диска.

Произвођачи магнетних дискова развили су SMART (енг. *Self-Monitoring, Analysis and Reporting Technology*) [12] технологију која пружа могућност надзора над низом радних параметара. Неки од најчешће праћених параметара су: радни сати диска, број

укључења/искључења, радна температура диска, број „лоших“ сектора (енг. *bad sectors*) и др. SMART технологија се ослања на постепену деградацију перформанси диска које настају као резултат хабања електромеханичких компоненти и магнетног медијума па се њиховим праћењем може предвидети предстојећи отказ диска. Произвођачи дефинишу дозвољене границе, односно прагове за вредности SMART параметара и када дође до прекорачења ових вредности корисник се упозорава да постоји могућност отказа магнетног диска. Ово упозорење би требало да се генерише благовремено, како би корисник имао довољно времена да изради резервну копију података пре него што наступи отказ диска. Како би се спречила појава лажних аларма који сигнализирају отказ, вредности прага SMART параметара се постављају високо од стране произвођача. То за последицу има да ће упозорење бити генерисано непосредно пре отказа, па је чест случај да корисник нема довољно времена да изради резервну копију података.

Приступ заснован на прекорачењу прагова SMART параметара, дефинисаним од стране произвођача је једноставан али поседује ниску стопу детекције отказа која се креће у распону од 3% до 10%. Google-ова студија која је спроведена на популацији од 100.000 магнетних дискова [13], утврдила је да се праћењем већег броја SMART параметара може предвидети око 64% свих отказа, док је остатак отказа узрокован деловањем утицаја који се не могу предвидети. Већина SMART параметара међусобно су зависни, тако да се потенцијални отказ може знатно раније препознати пре него што неки од индивидуалних параметара пређе вредност дефинисаног прага. Међусобну зависност SMART параметара је тешко одредити јер је изузетно сложена и из тог разлога се најчешће користе методе машинског учења. Ове методе омогућавају да одређени алгоритам научи зависност између отказа диска и SMART параметара из скупа података без експлицитног програмирања од стране програмера.

### 1.2 Циљ истраживања

Циљ предложеног истраживање обухвата развој новог модела за предвиђање отказа магнетних дискова, који је заснован на примени метода машинског учења. Развијени модел побољшаће стопу предикције отказа дискова у односу на приступ заснован на прекорачењу прагова SMART параметара. Основни циљ истраживања је развој модела за детекцију аномалија заснованог на мерењу Махаланобисовог растојања (енг. *Mahalanobis distance*) SMART параметара који омогућава предикцију отказа магнетних дискова.

Додатни циљеви истраживања су:

1. Смањење рачунарске комплексности модела елиминацијом SMART параметара који немају утицаја на предикцију отказа;
2. Остваривање високе стопе детекције отказа магнетног диска;
3. Одређивање границе одлучивања којом се избегава појава погрешних детекција отказа магнетног диска, односно лажних аларма;
4. Благовремено предвиђање отказа у интервалу од 12 до 48 часова пре настанка стварног отказа магнетног диска.

### 1.3 Основне хипотезе

Основне хипотезе на којима је засновано истраживање предикције отказа магнетних дискова представљено у овој дисертацији су:

1. Атрибути помоћу којих се прати рад магнетног диска немају исти утицај на механизам отказа диска. Неки од атрибута у већој мери доприносе предикцији отказа у односу на остале, сходно томе могу бити ранжирани методом машинског учења са рекурзивном селекцијом атрибута на основу њиховог коефицијента тежине.
2. Модел предикције отказа заснован на мултиваријабилним атрибутима оствариће бољу стопу детекције од модела који користи појединачне међусобно независне атрибуте. Атрибути који немају утицаја на детекцију отказа диска уклањају из модела из разлога да се избегне претерано уклањање (енг. *overfitting*).
3. Могуће је развити модел заснован на алгоритму за детекције аномалија, који ће са високом стопом детекције отказа FDR (енг. *Failure Detection Rate*) и врло ниском стопом лажних аларма FAR (енг. *False Alarm Rate*) предвидети отказе магнетних дискова;
4. Модел је у стању да благовремено предвиди отказ диска како би корисник на време креирао резервну копију података и тиме спречио њихов губитак.

### 1.4 Методе и методологије истраживања

Општа метода истраживања односи се на прикупљање и проучавање доступне литературе у штампаном и електронском облику која обухвата монографије, научне радове публиковане у часописима и у зборницима конференција. Прегледом литературе анализираће се методе и модели за предвиђање отказа магнетних дискова који су представљени од стране других истраживача. Прикупљање ових информација неопходно је како би се доказала оправданост за развој новог модела за предикцију отказа.

Специфичне методе које се примењују у оквиру овог истраживања ослањају се на употребу математичких, статистичких и рачунарских метода које су неопходне у области предвиђања отказа магнетних дискова. Ове методе обухватају прикупљање податка, статистичку обраду података, селекцију значајних атрибута, нормализацију података и одређивање растојања између различитих атрибута као и примену алгоритама за детекцију аномалија заснованих на удаљености.

1. Прикупљање података захтева праћење популације магнетних дискова у дужем временском периоду који раде у истим условима. Овакав начин прикупљања изискује значајне ресурсе, материјална средства и дуг временски период прикупљања података, што је недоступно већини истраживача. Из тог разлога се у истраживањима најчешће користе подаци из јавно доступних извора који су прикупљени од стране произвођача или центара података. За потребе истраживања спроведених у овој дисертацији коришћен је отворени CMRR (енг. *Center for Magnetic Recording Research*) [14] скуп података, као и у већини претходних истраживања у овој области.
2. Статистичка обрада података подразумева анализу скупа података како би из скупа података биле уклоњене абнормалне вредности података (енг. *outliers*), а након тога допуниле недостајуће вредности података.
3. Селекција значајних атрибута користи методу рекурзивне елиминације атрибута RFE (енг. *Recursive Feature Elimination*) која рангира атрибуте према њиховим коефицијентима тежине применом алгоритма подржавајућих вектора SVM (енг. *Support Vector Machine*).

4. Применом *Box-Cox* трансформације врши се трансформисање непараметарски дистрибуираних вредности атрибута у нормално дистрибуирани облик, како би се применили алгоритми за детекцију аномалија који су засновани на удаљености.
5. Мерење удаљености у вишедимензионалном простору између вредности различитих SMART параметара захтева примену Махаланобисовог растојања које одређује удаљеност у стандардним одступањима.
6. Алгоритми за детекцију аномалија засновани на удаљености примениће се за детектовање аномалија ако су неке тачке података довољно удаљене од већинске дистрибуције података.

## 1.5 Научни доприноси истраживања

Научни доприноси настали као резултат овог истраживања су:

1. Развој модела за предвиђање отказа магнетних дискова који је заснован на алгоритму за детекцију аномалија;
2. Модел остварује високу стопу предикције отказа ( $FDR > 90\%$ );
3. Модел успешно минимизује стопу лажних аларма ( $FAR < 0.1\%$ );
4. Предвиђање отказа диска у периоду од 12 до 48 часова пре настанка стварног отказа;
5. Селекција најзначајнијих атрибута побољшава предиктивне перформансе модела и смањује његову сложеност.

## 1.6 Преглед досадашњих истраживања

Досадашња истраживања везана за предвиђања отказа магнетних дискова заснована су на примени различитих метода машинског учења како би се утврдила зависност SMART атрибута и отказа диска. Овакав приступ захтева дуготрајну аквизицију радних параметара дискова на основу којих се креира одговарајући модел за предвиђање отказа магнетног диска. Главни недостатак ових модела је што захтевају прикупљање података у дужем временском периоду пре него што се започне са практичном применом модела. С обзиром да се на тржишту константно појављују нови модели дискова са бољим карактеристикама, постоји могућност да ће неки модел диска престати да се производи пре него што се довољно добро обучи алгоритам за предикцију отказа. Примена оваквих алгоритама оправдана је у центрима за складиштење података у којима се велика популација једног модела диска користи годинама и може помоћи у случају проактивног сервисирања кварова. Аутори у досадашњим истраживањима користили су ограничен број отворених скупова података од којих је у највећем броју истраживања коришћен CMRR скуп података који је такође коришћен и у овој докторској дисертацији. Алгоритми машинског учења који се користе за детекцију отказа обучавају се са циљем да се оствари стопа погрешних детекција отказа FAR која не прелази 0.1%, а да притом постигну високу стопу детекције отказа FDR.

У овом истраживању прикупљен је и анализиран велики број радова представљених од стране других истраживача у области предикције отказа магнетних дискова. Истраживања у овој области су оријентисана на развој алгоритама који користе мултиваријантне SMART параметре за предвиђање отказа. Истраживачи су на почетку користили основне алгоритме машинског учења за предвиђање отказа магнетних

дискова, као што је Наивни Бајесов (енг. *Naïve Bayes*) модел који је постигао стопу детекције отказа FDR од 30%, при стопи лажних детекција од 0.67% FAR [15].

Истраживања на бази *Wilcoxon* статистичких тестова, постигла су стопу предикције отказа од 32% FDR при 0.2% FAR [16] и 43.1% FDR на 0.6% FAR [17]. У раду [14] аутори су користили SVM алгоритам остваривши стопу предикције отказа од 50.6% FDR при 0% FAR. Коришћењем скривених Марковљевих модела (енг. *Hidden Markov models*) аутори рада [18] су постигли стопу детекције отказа од 52% FDR са 0% FAR. Истраживачи у раду [19] користили су FSMMD (енг. *Feature Selection-based Mahalanobis Distance*) методу која је заснована на мерењу Махаланобисовог растојања, при чему су постигли стопу детекције отказа од 67% FDR са 0% FAR.

Исти аутори су у раду [20] предложили нову двостепену параметарску методу TSP (енг. *Two Step Parametric*) такође засновану на Махаланобисовом растојању и методи клизног прозора, при чему је остварена стопа предикције отказа од 68% FDR, без детектованих лажних аларма. Применом проактивне методе предикције која је заснована на приоритетима [21], остварена је стопа предикције отказа од 86.3% FDR са 0.52% FAR. Аутори рада користили су вештачку неуронску мрежу са селективним параметрима [22] са којом су постигли стопу детекције отказа од 94.62% FDR и 0.48% FAR. Анализом претходних истраживања најбољи резултати постигнути су применом GMFD (енг. *Gaussian Mixture based Fault Detection method*) модела који користи Гаусове и непараметарске статистике [23] са стопом детекције отказа од 92.21% FDR и 0% FAR.

### 1.7 Преглед дисертације по поглављима

Ова теза је организована у девет поглавља.

У првом поглављу представљен је увод у истраживање који укључује предмет и циљ истраживања, хипотезе, методе истраживања као и преглед релевантних истраживања у овој области.

Друго поглавље дисертације описује структуру магнетних дискова и представља узроке који доводе до отказа магнетних дискова. У наставку овог поглавља дата је класификација алгоритама машинског учења и описани су алгоритми за предвиђање отказа магнетних дискова.

У трећем поглављу представљена је анализа и обрада података који су доступни из отворених извора података. У поступку обраде података користи се статистичка метода за уклањање абнормалних вредности података, а затим се приступа попуњавању недостајућих вредности података. У овом поглављу представљене су и технике које се користе за балансирање скупа података.

Четврто поглавље описује поступак селекције значајних атрибута помоћу алгоритма рекурзивне елиминације атрибута. Овим поступком добија се уређена листа атрибута, где први атрибут у листи највише доприноси предикцији отказа, док последњи атрибут на листи најмање утиче на предикцију отказа магнетног диска.

У петом поглављу приказана је *Box-Cox* трансформација за непараметарски дистрибуиране податке и поступак одређивања Махаланобисовог растојања. Трансформација података врши се појединачно за сваку инстанцу из скупа података, тако да вредности атрибута буду трансформисане у нормално дистрибуирани облик. Мерење растојања у вишедимензионалном простору извршено је помоћу Махаланобисовог растојања, које одређује удаљеност између инстанце података и центра дистрибуције података, представљено у стандардним одступањима.

Појам аномалија и њихова класификација представљени су у шестом поглављу. Такође, у овом поглављу представљени су различити типови алгоритама који се користе за детекцију аномалија.

Седмо поглавље приказује оригинални допринос дисертације који се огледа у реализацији модела за предвиђање отказа магнетних дискова који је заснован на алгоритму за детекцију аномалија. Најпре је представљена алгоритамска шема предложеног модела да би се приступило систематичном развоју свих компоненти предложеног модела. У другом делу овог поглавља приказане су метрике предиктивних перформанси модела које се користе за одабир оптималне границе одлучивања на основу које се детектују аномалије као откази диска. Поред тога, дат је приказ предлога имплементације реализованог модела за предикцију отказа у центрима података. Такође, анализирана је могућност примене модела у другим областима и приказан је прилагођени модел за предикцију отказа полупроводничких SSD дискова.

У осмом поглављу биће приказани постигнути резултати предиктивних перформанси реализованог модела. Резултати претходних истраживања остварени применом GMFD, TSP и FSMД модела биће упоређени са резултатима који су добијени тестирањем реализованог модела на CMRR отвореном скупу података. Критеријуми који се користе при поређењу модела су остваривање што веће стопе детекције отказа при чему се тежи избегавању погрешних детекције отказа. Поред стопе детекције отказа, модели су упоређени према могућности да предвиде отказ у интервалу 12 до 48 часова пре појаве стварног отказа магнетног диска. Такође, представљени су резултати примене модела заснованог на детекцији аномалија за предикцију отказа полупроводничких SSD дискова.

Закључна разматрања докторске дисертације, анализа испуњености постављених хипотеза и правци будућих истраживања представљени су у деветом поглављу.

## 2 Алгоритми за предикцију отказа магнетних дискова

У овом поглављу дисертације приказана је структура магнетних дискова и анализирани су узроци услед којих долази до отказа магнетних дискова. У наставку поглавља дата је класификација алгоритама машинског учења и описани су алгоритми који се користе за предвиђање отказа магнетних дискова.

### 2.1 Технологије трајног складиштења података

Рачунарски системи пружају могућност трајног складиштења корисничких података. Актуелне технологије складиштења обухватају магнетно складиштење на магнетним дисковима и магнетним тракама, полупроводничке дискове засноване на NAND FLASH технологији, магнетнорезистивне меморије и оптичко складиштење.

Магнетни дискови су већ дуго времена доминантан медијум за трајно складиштење података у рачунарским системима. Први комерцијално доступан магнетни диск 305 RAMAC [24] представљен је 13. септембра 1956. године, поседовао је 50 магнетних плоча пречника 24 инча које су биле капацитета пет милиона карактера величине 7 бита. Густина записа је била 2.000 битова по квадратном инчу, а брзина преноса је износила за то време импресивних 8.800 бајтова по секунди. Врло брзо су се показали као решење које је пружало знатно већу брзину приступа подацима у односу на магнетне траке и бушене картице. Константним унапређењем технологије капацитет магнетних дискова се повећавао, а димензије смањивале што је омогућило да задрже доминантну позицију као медијум за трајно памћење података дуже од пола века.

Паралелно са магнетним дисковима развијала се и технологија магнетних трака које се користе за архивирање података (терцијално складиштење). Магнетне траке поседују два реда величине мању густину записа у односу на магнетне дискове, али се њихова предност огледа у површини магнетног медијума за снимање која је око хиљаду пута већа од у односу на површину плоче диска од 3.5 инча. Због тога, магнетне траке остварују најнижу цену по ускладиштену биту али њихов недостатак је мала брзина приступа подацима, и из тог разлога се искључиво користе за чување резервних копија односно архивирање података. Очекивана трајност података уписаних на магнетну траку је најмање 30 година када се трака чува у препорученим архивским условима [25].

Полупроводничка FLASH технологија меморијског складиштења заснована на транзистору са пливајућим гејтом (енг. *Floating Gate*) [26] развијена је 80-тих година прошлог века три деценије након развоја магнетних медијума за складиштење података. У зависности од организације меморијских ћелија постоје NOR и NAND FLASH меморије, при чему се за складиштење података користи NAND FLASH меморија. Повећање густине записа код NAND FLASH меморије постигнуто је литографским скалирањем, уписом већег броја битова у меморијску ћелију и вишеслојном тродимензионалном (3D) технологијом. Полупроводнички SSD дискови састоје се од низа меморијских чипова заснованих на NAND FLASH технологији и контролера који се стара о равномерном упису података у флеш ћелијама. SSD дискови остварују боље перформансе у брзини приступа подацима јер немају покретне делове као магнетни дискови. У последњој деценији код персоналних рачунара магнетни дискови су потиснути од стране полупроводничких медијума који поседују боље перформансе, али и вишу цену по ускладиштену биту. Недостатак SSD дискова је ограничен број циклуса

уписа као и осетљивост на појаву битских грешака изазваних цурењем наелектрисања са пливајућег гејта. С обзиром да су SSD дискови вишеструко скупљи у односу на магнетне дискове, центри за складиштења података се ослањају на магнетне дискове на којима се данас складишти више од 90% свих података.

Технологија магнетнорезистивног складиштења користи магнетнорезистивни ефекат који омогућава брзо и поуздано складиштење података у MRAM меморији (енг. *Magnetoresistive Random Access Memory*) [27]. Ова технологија пружа велику брзину приступа подацима, али њен тренутни недостатак је висока цена по ускладиштену биту и из тог разлога није погодна за масовно складиштење података.

Технологија оптичког складиштења користи ласерску технологију за складиштење података у оптичком медијуму [28]. Ова технологија је била значајно заступљена у претходне три деценије у виду CD, DVD, Blue-ray дискова за архивирање и дистрибуцију података, али је последње деценије потпуно потиснута од стране полупроводничких FLASH меморија.

На Слици 1 приказан је тренд производње до 2029. године [29], за три доминантне технологије за складиштење података. SSD дискови су се појавили пре једне деценије али њихов удео на тржишту се до данас није променио у односу на магнетне дискове и магнетне траке. Пад производње у периоду 2022-2023 године, узрокован је застојима у ланцима снабдевања као последица светске пандемије. Тржиште се убрзо опоравило и предвиђања показују све већу потражњу за магнетним дисковима који ће у наредних пет година бити доминантан медијум за складиштење података, са учешћем од 75% од укупно произведеног капацитета. И ако неки аналитичари и произвођачи SSD дискова, попут Pure Storage-а, тврде да ће магнетни дискови убрзо нестати, анализа Coughlin Associates-а показује супротан тренд. Захваљујући континуираном повећању капацитета и смањењу цене по ускладиштену бајту њихова продаја ће наставити да расте након 2025. године а удео у укупном складишном капацитету ће значајно порасти до 2029. године. На основу претходне анализе може се закључити да ће магнетни дискови због ниске цене и трајности чувања података задржати доминацију у центрима за складиштење података бар још једну деценију.

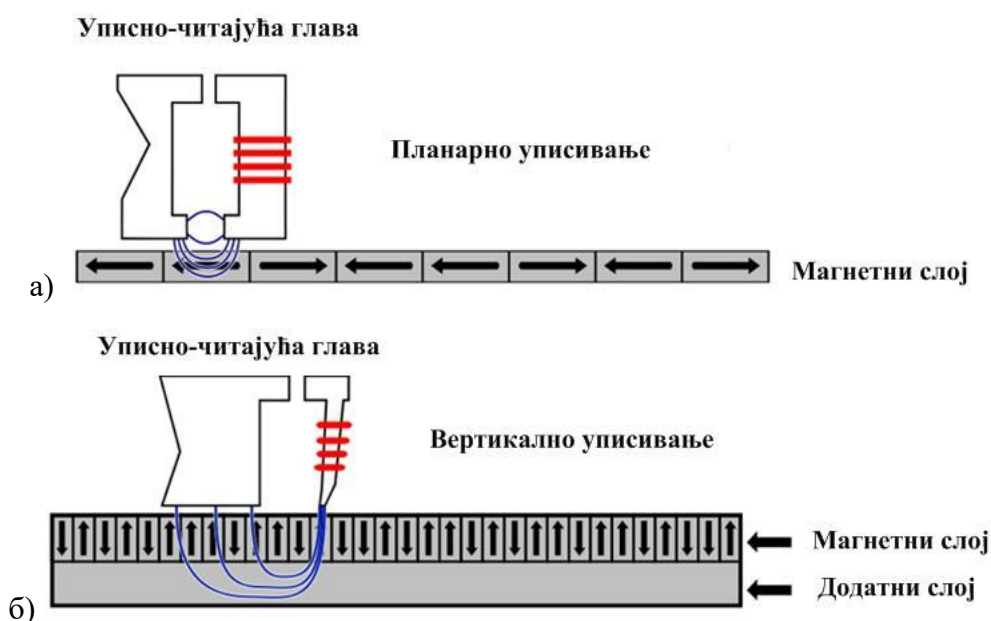


Слика 1. Прогнозирани тренд производње магнетних дискова (HDD), магнетних трака и полупроводничких дискова (SSD) до 2029. године [29]

## 2.2 Магнетни дискови

Магнетни диск се састоји од кружне плоче која се ротира константном угаоном брзином око своје осе помоћу погонског електромотора. Данашњи дискови најчешће се израђују са плочама пречника 3.5 инча, док се преносни дискови израђују са плочама од 2.5 инча. Плоча диска израђена је од стакла или алуминијума и пресвучена је слојем феромагнетног материјала. Подаци на магнетном диску се складиште у форми микроскопских магнетских домена, чији смер магнетизације зависи од логичке вредности ускладиштеног бита. Најчешће коришћени материјал за складиштење података је легура платине и кобалта која поседује магнетно тврду хистерезисну карактеристику са високом вредношћу коерцитивног магнетног поља. Она се наноси у форми танког слоја на површину диска који се састоји од низа магнетних зрна величине неколико нанометара. Магнетна зрна која су намагнетисана у истом смеру чине магнетни домен који је у стању да ускладишти један бит података. Магнетни домен се састоји од неколико магнетних зрна и његова величина зависи од величине уписно-читајуће главе.

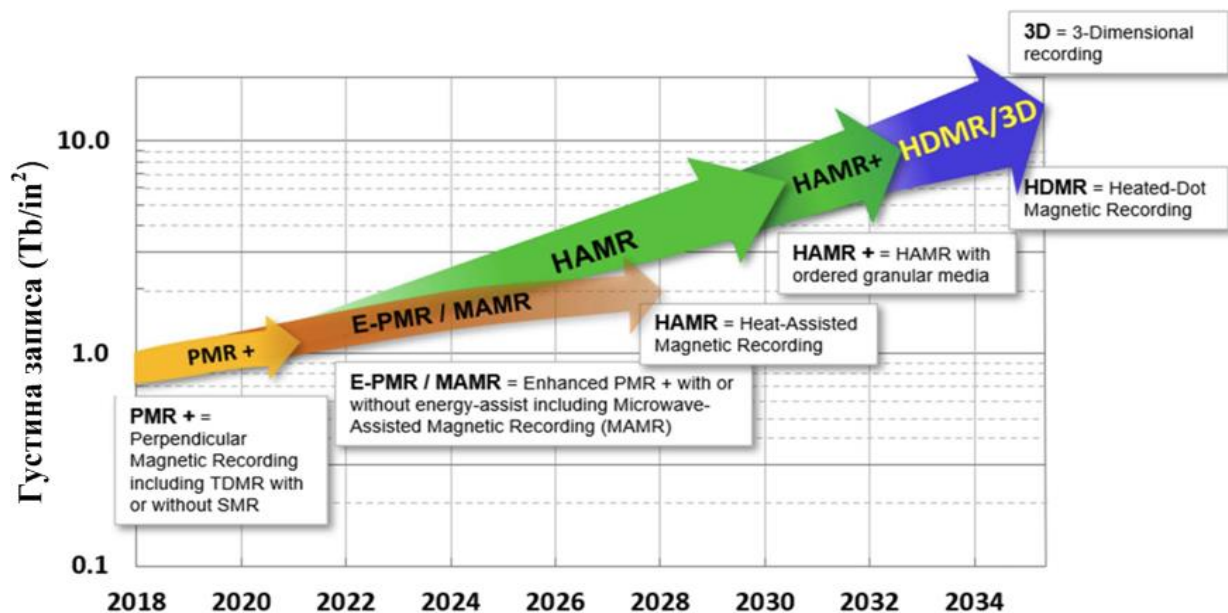
Уписно-читајућа глава састоји се од магнетног кола са намотајем кроз који се пропушта електрична струја чији смер зависи да ли се на диску уписује логичка нула или јединица. Генерисано магнетно поље у намотају главе се затвара кроз магнетни материјал на површини плоче који се трајно магнетише. С обзиром да се магнетна плоча ротира, сав магнетни материјал који прође испод главе ће бити намагнетисан у смеру који дефинише смер струје у намотају главе. Променама смера струје магнетне честице ће бити намагнетисане једна за другом као низ битова. У прошлости магнетни домени су били намагнетисани планарно у односу на површину плоче, тзв. LMR (енг. *Longitudinal Magnetic Recording*) технологијом магнетног записа [30] као што је приказано *Слици 2а*. Временом су се димензије уписно-читајуће главе смањивале што је резултовало повећањем густине записа и повећања капацитета диска. Смањењем димензија магнетних домена опада њихова јачина магнетног поља што ствара проблеме приликом њихове детекције. То се превазилази техникама кодирања при чему се битови кодирају помоћу промене магнетног поља између суседних домена које је много лакше детектовати. Планарно уписивање је ограничавало даљу минијатуризацију уписно-читајуће главе и даље повећање капацитета диска.



Слика 2. Планарно (а) и вертикално (б) уписивање података на магнетни диск

Применом технологије вертикалног уписивања PMR (енг. *Perpendicular Magnetic Recording*) [30] у односу на површину плоче, долази до промене конструкције главе која поседује магнетно коло у облику игле, која је у стању да намагнетише магнетне домене врло малих димензија (Слика 26). Магнетно поље које генерише намотај главе намагнетише циљани домен у вертикалном правцу у односу на раван диска. Ово поље се затвара преко доњег магнетног меког слоја кроз суседне магнетне домene. С обзиром да се магнетно поље дели кроз неколико суседних домена његова вредност у овим доменима је нижа од коерцитивне вредности магнетног поља материјала па неће доћи до њиховог размагнетисавања. Захваљујући технологији вертикалног уписивања која је омогућила даљу минијатуризацију уписно-читајуће главе, знатно је повећана густина записа и она данас достиже 1.028 Gb по квадратном инчу површине код дискова капацитета 16 TB [31].

Проблем који се јавио даљим повећањем густине записа је нестабилност магнетних домена коју су произвођачи решили коришћењем технологије топлотно потпомогнутог магнетног уписивања HAMR (енг. *Heat Assisted Magnetic Recording*) [32]. Ова технологија користи материјале на бази гвожђа и платине са високом вредношћу коерцитивног поља које спречава размагнетисавање суседних домена. Како би уписно-читајућа глава могла да намагнетише одређени домен, он се најпре загрева коришћењем минијатурног ласера на температуру блиској Киријевој температури, како би се лакше намагнетисао. Магнетни дискови капацитета 36 TB засновани на бази HAMR технологији који су тренутно доступни на тржишту постижу густину записа од 1.73 Tb по квадратном инчу површине [33]. Коришћењем ове технологије у блиској будућности достићи ће се густина записа од 10 Tb по квадратном инчу површине за дискове капацитета 120 TB. На Слици 3. приказан је план развоја будућих технологија за магнетно складиштење података у наредних десет година.

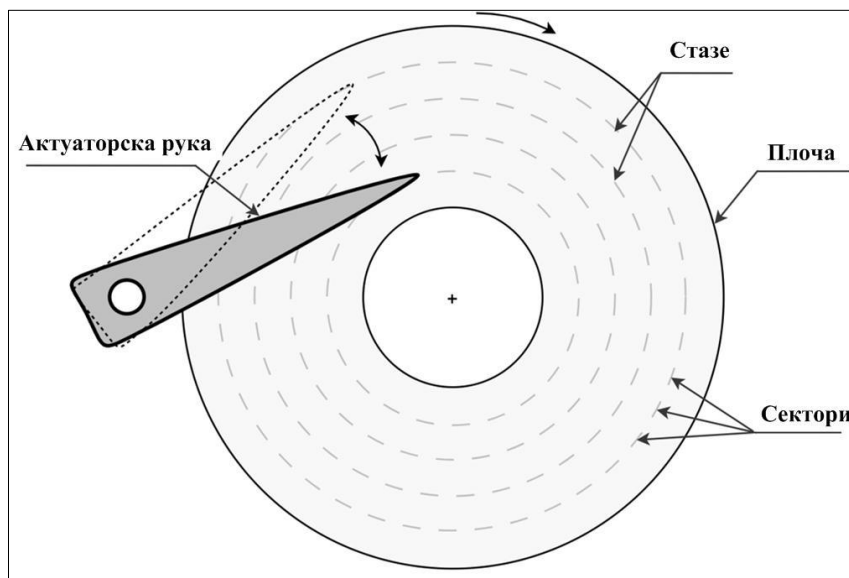


Слика 3. План развоја будућих технологија за магнетно складиштење података [34]

Подацима на површини диска се приступа помоћу уписно-читајуће главе која је причвршћена за актуаторски механизам који позиционира главу дуж целог полупречника плоче диска. Када се глава позиционира на одређену позицију на површини плоче диска, услед ротирања плоче, она ће моћи да приступи подацима који чине кружну структуру која се назива стаза. Површина диска приказана на Слици 4. организована је као низ концентричних стаза, које су међусобно одвојене размацама како

би се спречила интерференција између суседних стаза приликом операција уписа и читања.

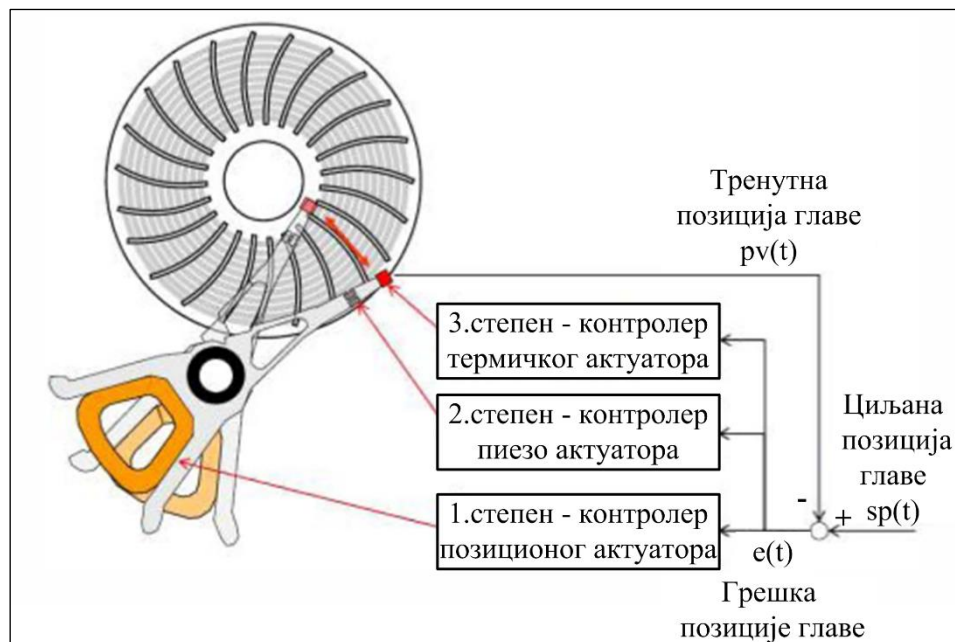
Савремене уписно-читајуће главе користе троструки позициони механизам, који се састоји од позиционог актуатора, пиезо актуатора и термичког актуатора [35]. Позициони актуатор састоји се од проводног намотаја који се налази у магнетном пољу сталног магнета. Пропуштањем струје кроз намотај долази до закретања целокупног актуаторског механизма. Овај актуатор омогућава кретање по целом полупречнику диска али његова брзина и прецизност позиционирања је мања у односу остала два актуатора.



Слика 4. Подела диска на стазе и секторе

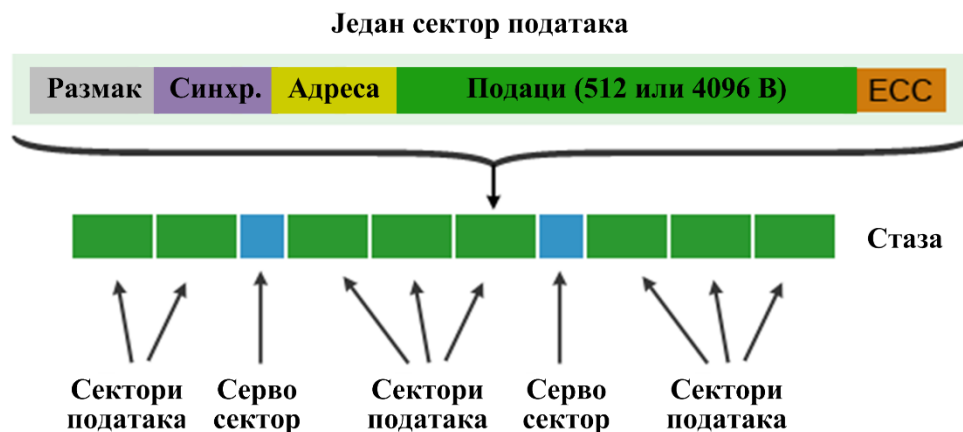
Пиезо електрични актуатор налази се на средини уписно-читајуће главе и састоји се од пиезо електричног елемента чијом активациом долази до закретања горње половине уписно-читајуће главе. Његов померај је знатно мањи у односу на позициони актуатор али је брзина и прецизност знатно већа, и он се користи за прелазак на суседне стазе. На самом врху уписно-читајуће главе налази се термички актуатор који врши микроскопске помераје уписно-читајуће главе одржавајући је на задатој стази. Структура троструког позиционог актуатора приказана је на Слици 5. Сигнал грешке позиције генерише се као разлика циљане и стварне позиције уписно-читајуће главе. Контролери актуатора користе овај сигнал грешке како би га поништили и довели главу у жељену позицију. Сваки од актуатора ће реаговати у зависности од величине сигнала грешке. За велике сигнале грешке реаговаће контролер позиционог актуатора, за средње сигнале контролер пиезо актуатора, а за најмањи сигнал грешке реаговаће контролер термичког актуатора [36].

Услед ротирања плоче долази до струјања ваздуха које подиже главу на висину од неколико стотина нанометара захваљујући којем глава лебди на ваздушном јастуку не остварујући контакт са плочом диска. Магнетни дискови су осетљиви на потресе и вибрације који могу узроковати да уписно-читајућа глава оствари контакт са површином диска и изазове оштећење магнетног материјала. Приликом изненадног нестанка напајања позициони механизам мора померити главу ван површине диска у паркирни положај, како услед губитка аеродинамичког ефекта не би дошло до контакта главе са површином диска.



Слика 5. Структура троструког позиционог актуатора уписно-читајуће главе [36]

Свака стаза је подељена на сегменте који се називају сектори и они представљају најмању јединицу података која се може прочитати и уписати на диск. Број сектора по стази код првих дискова је био константан за све стазе, што је за последицу имало слабије искоришћење периферних стаза на ободу диска. Овај недостатак се превазишао поделом стаза на зоне при чему стазе које припадају спољашњим зонама имају већи број сектора у односу на стазе које припадају унутрашњим зонама. Сваки сектор је одвојен размаком од суседног сектора како би се избегла интерференција приликом уписа и читања података. Стаза поред сектора података садржи и неколико серво сектора који се користе како би глава диска одржала прецизну позицију изнад стазе. Серво сектори представљају секторе који се уписују у поступку производње и садрже шаблоне који помажу у прецизном позиционирању главе на стази. Структура сектора података приказана је на Сlici 6.

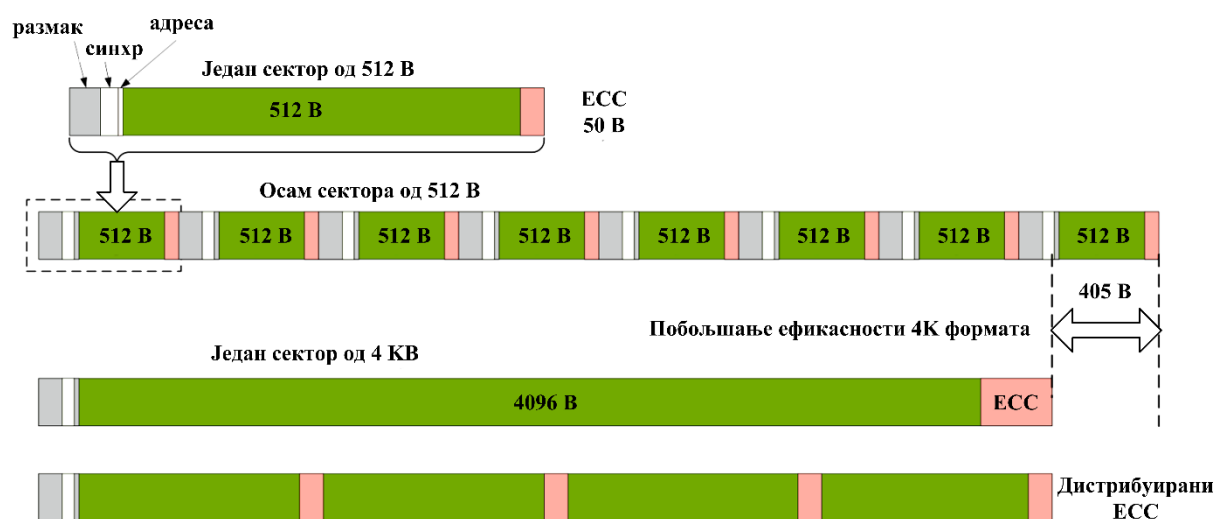


Слика 6. Структура сектора података [37]

Стандардна величина сектора података је 512 бајтова и он поред података садржи контролна поља која су неопходна за лоцирање сектора. Контролно поље величине 15 бајтова, које се налази на почетку сектора, садржи преамбулу која се користи за синхронизацију главе диска и праћено је адресним пољем које идентификује адресу сектора. Након контролног поља складишти се 512 бајтова корисничких података.

Последње ECC (енг. *Error Code Correction*) поље садржи код за проверу и отклањање грешака које омогућава да се детектују и отклоне грешке у подацима приликом операције читања. ECC поље садржи 50 бајтова Рид-Соломоновог (енг. *Reed Solomon*) кода који се генеришу на основу битова података и уписују се заједно са подацима у жељени сектор [38]. Приликом операције читања података генеришу се битови Рид-Соломоновог кода који се пореде са прочитаним ECC пољем на крају сектора. Уколико се кодови поклопе, то значи да прочитани подаци не садрже ни једну битску грешку и могу се проследити кориснику. У случају да се битови кода међусобно разликују, Рид-Соломонов код омогућава лоцирање одређеног броја битских грешки у податку који се након корекције грешака могу проследити кориснику. Ови кодови омогућавају и детекцију веће групе грешака које се не могу кориговати и у том случају се подаци из сектора проглашавају не валидним.

Новији дискови користе секторе података величине 4 KB који мењају осам стандардних сектора величине 512 бајта [39]. Они користе само једно контролно поље за разлику од стандардног формата који захтева осам контролних поља за исту количину података. Структура сектора од 4 KB приказана је на *Слици 7.* и упоређена са секторима стандардне величине. У *Табели 1* међусобно су упоређене величине поља стандардног сектора и сектора величине 4K при чему 4K сектори заузимају 10% мање простора на површини диска, па се њиховим коришћењем може ускладиштити 10% више корисничких података у односу на стандардне секторе.

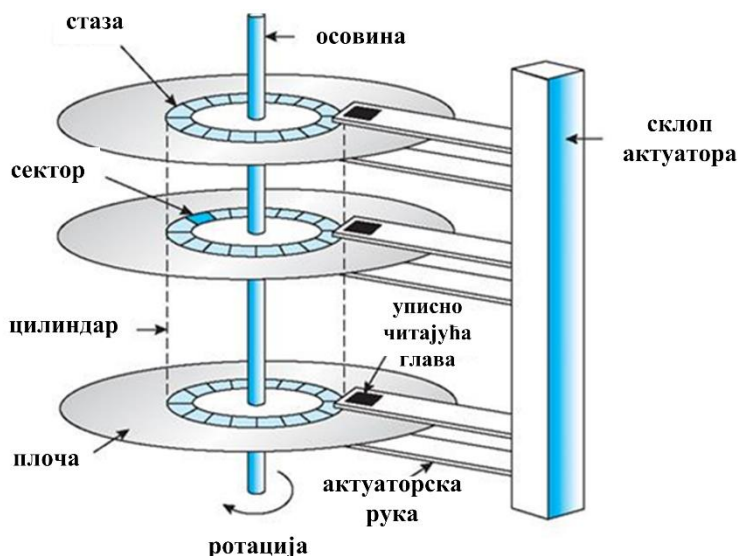


Слика 7. Поређење структура стандардног и 4K сектора [40]

Табела 1. Величине поља стандардног и 4K сектора

| Тип сектора      | Стандардни формат 512 | Унапређени формат 4 К |
|------------------|-----------------------|-----------------------|
| Контролно поље   | 15 В                  | 15 В                  |
| Подаци           | 512 В                 | 4096 В                |
| ECC поље         | 50 В                  | 100 В                 |
| Величина сектора | 577 В                 | 4211 В                |
| Ефикасност       | 88.7%                 | 97.3%                 |

Савремени дискови користе већи број двострано намагнетисаних плоча које су наслагане једна изнад друге на истој погонској осовини око које се заједно ротирају. Свака површина магнетне плоче захтева посебну уписно-читајућу главу које су механички спрегнуте на исти позициони механизам приказан на *Слици 8*. Уписно-читајуће главе позиционираће се над скупом стаза које су подједнако удаљене од осе ротације, и оне заједно формирају виртуелну структуру која се назива цилиндар. Овакво решење значајно поједностављује структуру диска што омогућава достизање високих капацитета једноставним додавањем већег броја плоча. У последњем периоду појављују се дискови испуњени хелијумом који пружа знатно мањи отпор ротације у односу на дискове испуњене ваздухом. Овакво решење омогућава израду тањих плоча при чему најновији дискови имају и до десет магнетних плоча и достижу капацитете до 36 ТВ.



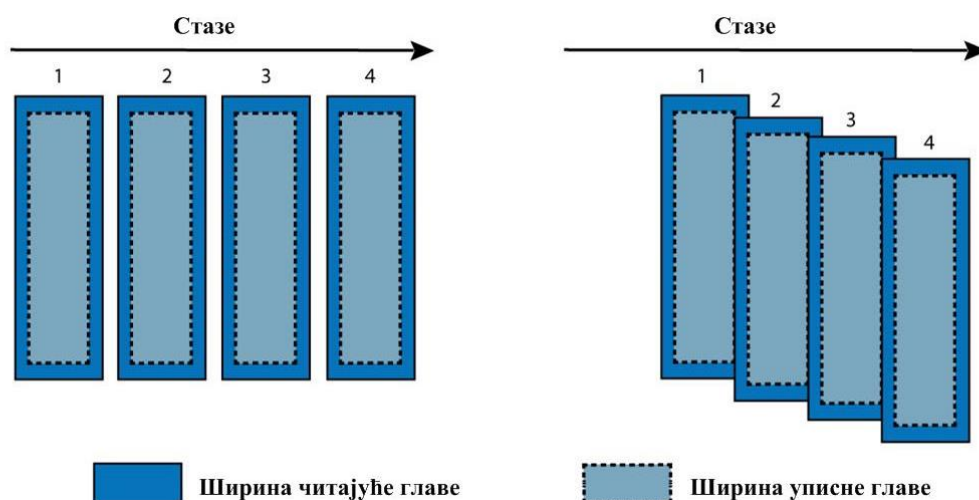
Слика 8. Организација података код дискова са већим бројем магнетних плоча [41]

Магнетни диск представља меморију са секвенцијалним приступом код које време приступа секторима није једнако и зависи од релативне позиције уписно-читајуће главе у односу на жељени сектор. Приступ сектору одвија се кроз низ операција које захтевају позиционирање уписно-читајуће главе на одговарајућу стазу и проналазак одређеног сектора на тој стази. Време позиционирања главе на стазу зависи од брзине кретања актуаторског механизма и најчешће се представља као средње време позиционирања које износи десетак милисекунди. Када се глава позиционира на стазу потребно је да сачека да се плоча диска заротира док жељени сектор не дође испод главе диска, ово време се дефинише као ротационо кашњење и оно представља време потребно да се плоча диска заротира за половину круга. Време приступа подацима код савремених дискова зависи од области њихове намене и приказано је у *Табели 2*.

Табела 2. Времена приступа код савремених магнетних дискова [42]

| Тип диска                | Брзина ротирања          | Средње време позиционирања | Ротационо кашњење | Време приступа |
|--------------------------|--------------------------|----------------------------|-------------------|----------------|
| Екстерни преносни диск   | 5.400 min <sup>-1</sup>  | 12.0 ms                    | 5.6 ms            | 17.6 ms        |
| Интерни диск             | 7.200 min <sup>-1</sup>  | 8.0 ms                     | 4.2 ms            | 12.2 ms        |
| Диск за центре података  | 10.000 min <sup>-1</sup> | 3.5 ms                     | 3.0 ms            | 6.5 ms         |
| Диск високих перформанси | 15.000 min <sup>-1</sup> | 2.0 ms                     | 2.0 ms            | 4.0 ms         |

Конвенционални приступ CMR (енг. *Conventional Magnetic Recording*) организацији стаза на површини магнетног диска подразумева да између две суседне стазе постоји ненамагнетисани простор који формира размак између стаза, као што је приказано на *Слици 9а*. Размак спречава интерференцију између суседних стаза приликом операција читања и уписа. Оваква организација стаза са размацима се користи код већине дискова високих перформанси, јер се свакој стази може директно приступити. У последње време код преносних дискова, користи се технологија SMR (енг. *Shingled Magnetic Recording*) преклапања стаза, која елиминише размаке и повећава густину записа за 11% [43]. Код ове технологије стазе су делимично преклопљене, па упис у једну стазу захтева да се препишу и стазе које се са њом преклапају, као што је приказано на *Слици 9б*. Недостатак технологије са преклапањем стаза изражен је приликом операције уписа, када је најпре потребно прочитати све преклопљене стазе и поново их уписати са измењеним садржајем. У случају операција са интензивним уписом код ове врсте дискова може доћи до смањење перформанси диска односно брзине уписа.



Слика 9. Организација стаза (а) са размацима и (б) стаза са преклапањем

### 2.3 Класификација грешака код магнетних дискова

Приликом читања ускладиштених података са диска могућа је појава битских грешака. Контролер диска користи низ техника за отклањање битских грешака које се јављају приликом читање сектора. Када се открије битска грешка најпре се примењују једноставније технике а ако оне не успеју да отклоне битску грешку користи се техника за отклањање грешака вишег нивоа, која захтева више времена за корекцију. Коришћене технике зависе од самог произвођача, али генерално се састоје од следећих фаза:

1. Фаза – ЕСС детекција грешака: приликом читања сектора чита се и поље ЕСС кода помоћу којег контролер проверава да ли постоје битске грешке. Уколико нема грешака подаци из сектора се прослеђују преко интерфејса и операција читања се сматра успешно завршеном.
2. Фаза – ЕСС корекција грешака: ако су детектоване битске грешке контролер ће покушати да их коригује помоћу ЕСС кода. Процес корекције се обавља веома брзо без кашњења, подаци са коригованим грешкама се прослеђују кориснику и операција читања се сматра успешном након корекције.
3. Фаза – поновни покушај читања: уколико је број битских грешака исувише велики да би био коригован ЕСС кодом, контролер диска ће сачекати да се плоча диска направи пун круг како би поново прочитао секторе. Грешке читања у

претходном кораку могу бити узроковане лошом позицијом главе, вибрацијама, потресом или другим не понављајућим проблемом, па поновљени покушај читања може дати боље резултате. Исправљени подаци се прослеђују преко интерфејса, а грешка се сматра поправљеном након поновног покушаја. Контролер диска ће покушати да прочита сектор неколико пута пре него што одустане. Време потребно за поновни покушај читања сектора износи десетину милисекунди по покушају.

4. Фаза – напредно исправљање грешака: након више неуспешних читања контролер ће искористити прочитане податке из свих претходних покушаја како би коришћењем напредних алгоритама покушао да отклони грешку. Ови алгоритми захтевају много више времена у односу на регуларни ЕСС код, али имају много већу шансу за успех. Исправљени подаци се шаљу преко интерфејса и операција се проглашава успешном после вишеструких читања.
5. Фаза – грешка читања сектора: ако сектор на крају не може бити прочитан онда се проглашава непоправљива грешка читања која подразумева да су подаци изгубљени.

Било какви проблеми током читања, чак и када се операција заврши успешно, биће евидентирани као потенцијални разлог за отказ диска. Контролер диска ће предузети корективне мере ако неки сектор мора прочитати из више покушаја или користити напредне алгоритме за отклањање грешака како би прочитао садржај сектора. Такав сектор се проглашава непоузданим и кориговани садржај се пресликава у неки од резервних сектора са којима неће бити проблема у наредним читањима.

Сваки модел магнетног диска поседује спецификације вероватноће настанка различитих грешака, које се утврђују фабричким тестирањем. Спецификација вероватноће грешке се изражава као број прочитаних битова између две појаве истог типа грешке и приказана је Табелом 3.

Табела 3. Учесталост појаве различитих типова грешки

| Тип грешке                | Учесталост појаве |
|---------------------------|-------------------|
| Аутоматска ЕСС корекција  | 1 Mb              |
| Поновно читање сектора    | 1Gb               |
| Вишеструко читање сектора | 1Tb               |
| Неуспело читање           | 100Tb             |

Осим проблема читања сектора проблеми могу настати приликом лоцирања стазе на којој се сектор налази и такве грешке се називају грешке тражења стазе. Оне настају услед оштећења уписно-читајуће главе или серво сектора, или оштећења позиционог механизма. Уколико након неколико сукцесивних покушаја глава не успе да се постави на стазу сматраће се да је диск отказао.

Општи узроци формирања лошег сектора могу бити физичко оштећење магнетног медијума или појава привремених битских грешака унутар сектора. До физичких оштећења сектора најчешће долази услед контакта главе са површином сектора који може настати због потреса, удараца или изненадног нестанка напајања. Површина магнетног материјала заштићена је слојем од карбона како би се магнетни материјал заштитио од оштећења услед контакта са главом. Физичка оштећења најчешће доводе до трајног оштећења сектора услед којих се не може даље користити. Разлози за појаву

привремених битских грешака најчешће су узроковани уписом и читањем података под неповољним условима који подразумевају нетачну позицију главе на стази или превелики размак главе у односу на површину диска. У оба случаја долази до читања неисправног садржаја сектора који не пролази контролу исправности јер садржи велики број битских грешака које се не могу исправити. Садржај таквог сектора не може бити испоручен кориснику и проглашава се изгубљеним. Применом специјалних дијагностичких програма може се прочитати садржај сектора у изворном облику како би се спровела процедура опоравка података чији позитиван исход није увек изван. Такав сектор се проглашава непоузданим од стране контролера диска и даље се неће користити док се над њим не изврши дијагностика. Контролер диска ће покушати да упише нови садржај у сектор и прочита исти и уколико се и тада појави грешка сектор се проглашава невалидним јер не може поуздано да складишти податке и неће бити коришћен у даљем раду диска. Неки од невалидних сектора се детектују током процеса производње, док се други јављају током радног века диска. Контролер магнетног диска води евиденцију о овим секторима у форми Р-листе (енг. *Production list*) и G-листе (енг. *Grown list*), тако да их не користи када је потребно да сачува податке. Примарна Р-листа се формира током процеса производње и садржи све секторе за које је утврђено да су непоуздани у фази фабричког тестирања диска. Ова листа се уписује у контролер диска и она се не мења током животног века диска. Током радног века диска откривају се сектори који нису поуздани и уписују се у G-листу која се временом повећава и представља добар индикатор тренутног стања магнетног диска.

Магнетни диск представља електромеханички систем који је израђен са врло високом прецизношћу. Како би се дугорочно остварио поуздан рад диска неопходно је да низ параметара остане у прописаним границама. Због електромеханичке конструкције неизбежно ће долазити до хабања механичких компоненти које ће имати негативан утицај на поузданост диска. Ако неки од параметара значајно одступи од граница прописаних од стране произвођача, онда постоји велика вероватноћа да ће доћи до отказа диска. Под отказом диска подразумева се немогућност диска да кориснику испоручи захтеване податке односно да их испоручује брзином која је далеко нижа од минималне брзине гарантоване у произвођачкој спецификацији. У најгорем случају отказ диска доводи до трајног губитка ускладиштених података, који се једино могу повратити у специјализованим лабораторијама након сервисирања механичких делова диска.

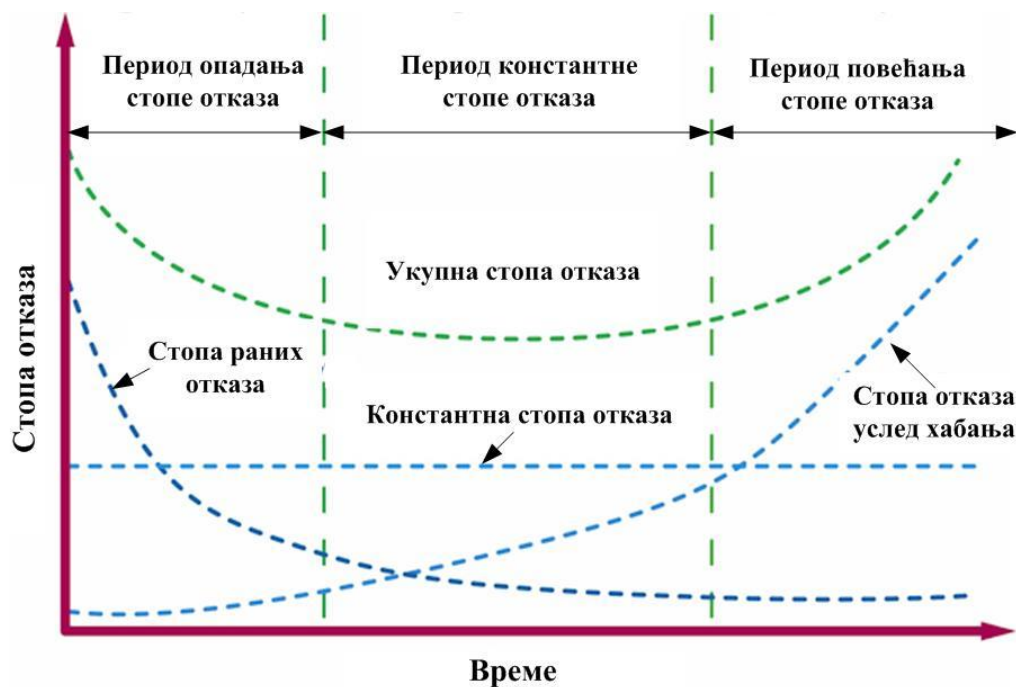
Перформансе магнетног диска које су битне за обичног корисника су капацитет, брзина приступа и цена. Са становишта основне намене магнетног диска за трајно складиштење података, његова најбитнија карактеристика је поузданост. Захваљујући напредној технологији коју користе магнетни дискови поузданост дискова је знатно повећана, међутим услед електромеханичке конструкције неизбежно је да ће диск после неког времена отказати што доводи до губитка ускладиштених података.

Магнетни дискови прате такозвани модел криве кварова која се састоји од три периода, приказаних на *Слици 10*. Први период је период раног отказивања, кога карактерише висока стопа отказа, која настаје и открива се приликом фабричког тестирања. Други период се назива радни период са малом стопом кварова који траје неколико година. Трећи период карактерише постепено повећање стопе кварова услед хабања компоненти које наступа када диск достигне свој корисни век трајања. Сервисни век трајања (енг. *Service life*) је време које је дефинисао произвођач након којег диск постепено губи на поузданости. Радни век диска се типично креће од три до пет година током којег постоји релативно мала шанса да он откаже. Ако се настави са употребом диска након овог периода, вероватноћа отказа ће се повећавати што се он дуже користи. У случају да је поузданост диска кориснику јако битна, потребно је извршити замену

диска по истеку његовог сервисног века. Овом приликом треба нагласити да ако је кориснику битна поузданост, он би требао да креира редовне резервне копије података. Услови околине и начин на који се диск користи могу значајно утицати на дужину његовог сервисног века, уколико се разликују од оних које је прописао произвођач.

Најчешће коришћен параметар квалитета и поузданост магнетних дискова је средње време између отказа MTBF (енг. *Mean Time Between Failures*). Овај параметар представља просечно време, изражено у часовима, које протекне између два насумична квара на диску датог типа. Код данашњих дискова овај параметар се креће у распону од 300.000 до 2.500.000 часова у зависности од намене диска [44]. MTBF параметар не одређује вероватноћу отказа појединачног диска већ се примењује за процену вероватноће отказа веће популације дискова.

На пример ако је MTBF параметар једнак 300.000 часова, то не значи да ће појединачни диск отказати након 34 године континуираног рада (8.766 часова по години рада) јер је његов сервисни век вишеструко краћи. Правилно тумачење MTBF параметра подразумева да ће доћи до једног отказа диска у популацији од  $N$  дискова који су укупно радили 300.000 сати (под условом да дискови нису прекорачили декларисани животни век од стране произвођача). Ако популацију чини 100 дискова чији је животни век три године они ће укупно радити 2.629.800 часова и за тај период ће у просеку отказати осам дискова из ове популације.



Слика 10. Стопе отказа у току животног века диска [45]

Постоји велики број различитих начина услед којих магнетни диск може отказати, и они се могу поделити на следеће типове отказа:

- Механички откази, представљају кварове услед хабања механичких компоненти диска. Откази овог типа везани су за мотор који ротира плоче диска, оштећење лежајева мотора и актуаторског механизма. Овакви откази су последица хабања, рада на високим температурама и присуства вибрација и потреса.
- Откази уписно-читајуће главе, настају услед непрописне висине лебдења главе изнад површине диска, што за последицу има појаву битских грешака приликом уписа и читања. Непрописна висина може бити узрокована вибрацијама,

потресом, контаминацијом ситним честицама и у случају рада на великим висинама где услед ниског притиска ваздуха глава не може постићи правилну висину лебдења, у том случају може доћи и до контакта главе са плочом и њеног оштећења.

- Оштећење магнетног медијума, настаје услед контакта главе са површином диска или услед деформације и оштећења магнетних плоча.
- Отказ контролера диска, настаје услед оштећења интегрисаних кола и других компоненти као и софтвера који контролише рад диска.

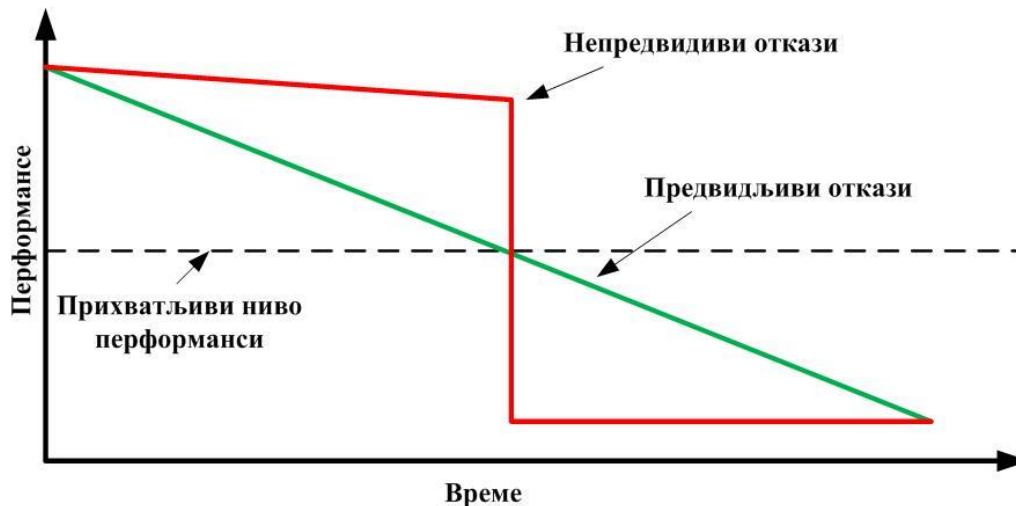
Произвођачи магнетних дискова константно унапређују квалитет процеса производње како би њихови дискови поуздано чували корисничке податке. Такође, произвођачи додају посебне функционалности које за циљ имају смањење шансе за појаву проблема или обавештавају корисника да ће се проблем појавити пре него што дође до губитка података.

### 2.4 Технологија самонадзорне анализе и извештавања SMART

Како би се избегао губитак корисничких података, произвођачи дискова уграђују систем раног упозоравања за наступајуће проблеме у раду диска који се назива технологија самонадзорне анализе и извештавања, познатији као SMART. Контролер надзире рад магнетног диска и у комбинацији са подацима са сензора надгледа различите аспекте перформанси диска на основу којих утврђује да ли се диск понаша нормално или не. С обзиром да је велики број отказа узрокован постепеним губитком перформанси развијен је SMART систем за праћење битних параметара диска. Овај систем прати низ параметара поузданости рада диска који се пореде са вредностима од стране произвођача. Када неки од SMART параметара пређе прописане границе корисник се упозорава да може доћи до отказа диска и да је потребно направити копију података (енг. *backup*) како би се спречио потенцијални губитак корисничких података уколико диск откаже. SMART се појавио у форми стандарда 1995. године од стране водећих произвођача дискова. Састоји се од скупа од 255 параметара који описују различите показатеље рада диска. Већина битних параметара је заједничка за све произвођаче дискова, али постоје и параметри који су јединствени за одређене произвођаче. Сваки параметар дефинише скуп вредности и вредност прага коју параметар не сме прекорачити у нормалном режиму рада. Параметри су приказани у сировом и нормализованом облику. Сирова вредност параметра исказана је у децималном или хексадецималном облику и представља неку вредност, чије значење дефинише произвођач (број значајних догађаја или нека физичка вредност). Нормализована вредност параметара представљена је у опсегу од 1 до 253, при чему 1 представља најнеповољнију вредност параметара, а 253 најповољнију вредност. Почетна нормализована вредност поставља се на 100, али она може варирати од произвођача до произвођача.

Основна претпоставка SMART технологије је да откази диска настају као резултат постепене деградације електромеханичких компоненти [46]. Он је настао из IBM технологије под називом предиктивна анализа кварова PFA (енг. *Predictive Failure Analysis*), која дели отказе у две категорије, предвидљиве и непредвидљиве. Предвидљиви кварови настају постепено током времена и често остављају трагове на основу којих се предстојећи отказ може предвидети. Пример предвидљивог отказа је квар код покретања погонског мотора који се може предвидети праћењем времена потребног да мотор достигне номиналну брзину. Номинално време покретања утврђује се фабричким тестовима на исправним дисковима. Током експлоатације долази до

хабања лежајева или отказа једног од намотаја погонског мотора што за последицу има повећање времена покретања у односу на номиналне вредности. Време покретања ће постајати све дуже и када пређе границе прописане од стране произвођача врло је вероватно да ће наступити отказ диска услед овог проблема. Пример непредвидљивог квара је отказ интегрисаног кола на контролеру диска који настаје изненада најчешће услед спољних утицаја и његов настанак се не може предвидети. Утицај предвидљивих и непредвидљивих отказа на перформансе диска приказан је на *Слици 11*.



Слика 11. Утицај предвидљивих и непредвидљивих перформанси на отказ диска

Основни принцип код предвиђања отказа је да поједини кварови узрокују постепене промене у различитим индикаторима перформанси који се могу пратити како би се открио тренд који указује на предстојећи отказ диска. Произвођачи дискова утврђују номиналне (прихватљиве) вредности код различитих параметара које се утврђују статистичком анализом и који се памте у контролеру диска. Током експлоатације контролер диска прати параметре перформанси и у колико оне одступе из прихватљивог опсега дефинисаног од стране произвођача, уписује се статус упозорења у одговарајући регистар како би упозорио корисника о настанку проблема. SMART технологија само омогућава извештавање о статусу диска, али није у стању да самостално генерише упозорења о проблемима у раду диска. Ова технологија захтева постојање дијагностичког програма инсталираног на корисничком рачунару који периодично чита статусе SMART параметара и у случају да они садрже упозорења алармира корисника о детектованом проблему.

SMART технологија је корисна али није потпуно сигурна, јер може детектовати само неке проблеме настале током рада. Приликом појаве упозорења не би га требало игнорисати већ се обратити техничкој подршци произвођача. Поједини произвођачи дискова сматрају да су ова упозорења довољан доказ да је диск неисправан и покрећу процедуру замене диска у гарантном року, док други произвођачи могу додатно тражити покретање дијагностичког програма како би се утврдио стварни статус диска. У сваком случају циљ произвођача је да дефинисане границе номиналних параметара буду довољно широке како би се спречила појава детекције лажних отказа.

Магнетни дискови поседују ЕСС механизам за опоравак грешака који се користи реактивно, односно процедура опоравка грешака започиње се када корисник затражи податке из жељеног сектора. Услед постепене деградације магнетног медијума могу се појављивати битске грешке чији се број временом повећава. Уколико при захтеву за читање података број битских грешака буде већи од могућности корекције ЕСС кодом, подаци неће моћи бити испоручени кориснику и сматраће се изгубљеним. Како би се

побољшала поузданост складиштених података поједини произвођачи омогућавају проактивну проверу грешака коју самостално покреће контролер диска када нема захтева од корисника. Контролер чита целокупне секторе на диску и уколико наиђе на битске грешке врши њихово исправљање, у постојећем сектору или премешта податке у нови слободан сектор. Захваљујући проактивним техникама које се користе у режиму мировања диска грешке се откривају и коригују на време како би се спречио губитак података, услед акумулирања неисправљених грешака. У следећој Табели 4, приказани су најбитнији SMART параметри који могу бити индикатор непосредног отказа диска.

Табела 4. Битни SMART параметри који утичу на отказ магнетног диска

| Назив SMART параметра                                      | Опис SMART параметра                                           |
|------------------------------------------------------------|----------------------------------------------------------------|
| SMART 1<br>( <i>Read Error Rate</i> )                      | Број хардверски грешака при операцији читања са површине диска |
| SMART 5<br>( <i>Reallocated Sectors Count</i> )            | Број лоцираних и ремапираних лоших сектора                     |
| SMART 7 ( <i>Seek Error Rate</i> )                         | Стопа грешака приликом позиционирања главе на серво сектору    |
| SMART 10<br>( <i>Spin Retry Count</i> )                    | Број покушаја покретања погонског мотора                       |
| SMART 174<br>( <i>Unexpected Power Loss Count</i> )        | Број небезбедних искључења                                     |
| SMART 187<br>( <i>Reported Uncorrectable Errors</i> )      | Број грешака које не могу бити исправљене коришћењем ECC       |
| SMART 188<br>( <i>Command Timeout</i> )                    | Број обустављених операција                                    |
| SMART 194 ( <i>Temperature Celsius</i> )                   | Интерна температура диска                                      |
| SMART 196<br>( <i>Reallocation Event Count</i> )           | Укупан број операција ремапирања сектора                       |
| SMART 197<br>( <i>Current Pending Sector Count</i> )       | Број нестабилних сектора                                       |
| SMART 198<br>( <i>Offline Uncorrectable Sector Count</i> ) | Укупан број непоправљивих грешака                              |
| SMART 206 ( <i>Flying Height</i> )                         | Висина главе у односу на површину диска                        |
| SMART 241 ( <i>Total Host Writes</i> )                     | Укупан број уписаних сектора                                   |
| SMART 242 ( <i>Total Host Reads</i> )                      | Укупан број прочитаних сектора                                 |
| P-list ( <i>Production list</i> )                          | Листа оштећених сектора откривених током фабричких тестова     |
| G-list ( <i>Grown list</i> )                               | Листа оштећених сектора насталих током оперативног рада диска  |

## 2.5 RAID систем независних редундантних дискова

Досадашња разматрања односила су се на побољшања поузданости индивидуалног магнетног диска, међутим и ако би се дискови производили као високо поуздани што би за последицу имало високу цену коштања, и даље постоји могућност да они могу отказати на непредвидив начин. У областима у којима је поузданост података императив, уместо једног диска користи се више дискова у редундантним или делимично редундантним конфигурацијама. Овакве конфигурације називамо RAID систем који представља низ редундантних независних дискова. RAID систем састоји се од већег броја независних физичких дискова, најчешће истих карактеристика, који су повезани са RAID контролером преко којег корисник види све ове дискове као јединствен логички диск. Употребом већег броја дискова омогућава RAID систему да повећа капацитет, перформансе и поузданост у односу на појединачне магнетне дискове. RAID системи се могу имплементирати на седам различитих начина (нивоа), при чему постоји још пуно хибридних решења.

Најједноставнији начин формирања низа дискова је у RAID 0 нивоу, где се подаци деле на стазе које се уписују на појединачне дискове у низу. RAID 0 ниво не поседује имплементирану редундантност и стога у свом називу користи ознаку нула. Отказом било ког диска у низу, подаци на осталим дисковима постају неупотребљиви. Овај ниво користи се у апликацијама код којих је неопходан висок капацитет и велика брзина приступа подацима, а безбедност података није битна. RAID 1 ниво представља потпуну супротност у односу на RAID 0, јер се код овог нивоа на свим дисковима налази идентична копија података. Отказом диска, подаци остају доступни у RAID 1 нивоу све док постоји бар један оперативни диск. Овај RAID ниво постиже изузетну поузданост, при чему се не повећава ни капацитет ни брзина приступа подацима на диску.

Остали RAID нивои користе делимичну редундантност која се остварује са минимално три диска и по карактеристикама се налазе између нивоа RAID 0 и RAID 1. Делимично редундантни нивои памте стазе или блокове података на независне дискове при чему користе један или два додатна диска за складиштење кода за детекцију грешака. У случају отказа једног или два диска у низу ови RAID нивои ће бити у стању да реконструишу податке који су се налазили на отказалим дисковима.

Употреба RAID система економски је оправдана у случају већих организација и центара података, док употреба код персоналних рачунара није економски оправдана због неопходности коришћења већег броја дискова. Отказом једног диска у редундантном низу не долази до губитка ускладиштених података. Када се отказали диск замени новим празним диском, RAID контролер започиње поступак реконструкције изгубљених података на основу података складиштених на осталим дисковима. Када се заврши реконструкција података RAID систем прелази у регуларни радни режим. Претпоставка је да током фазе реконструкције неће доћи до отказа још неког диска у низу, па се у таквим случајевима може користити RAID 6 ниво који је отпоран на истовремени отказ два диска у низу.

Хибридне RAID конфигурације укључују комбиновање два различита RAID нивоа на различитим нивоима организације низа дискова којима се постижу карактеристике система које су погодне за поједину област примене. Хибридне конфигурације захтевају четири и више физичких дискова па често нису исплативе за употребу у персоналним рачунарима. Карактеристике различитих RAID конфигурација приказане су у Табели 5.

Табела 5. Карактеристике различитих RAID конфигурација

| RAID ниво              | RAID 0 | RAID 1 | RAID 5 | RAID 6 | RAID 10 | RAID 50 | RAID 60 |
|------------------------|--------|--------|--------|--------|---------|---------|---------|
| Минимални број дискова | 2      | 2      | 3      | 4      | 4       | 6       | 8       |
| Цена                   | Ниска  | Висока | Ниска  | Ниска  | Висока  | Средња  | Висока  |
| Искоришћење капацитета | 100%   | 50%    | 67-94% | 50-80% | 50%     | 67-94%  | 50-80%  |
| Заштита података       | Нема   | Висока | Средња | Висока | Средња  | Ниска   | Средња  |

## 2.6 Предвиђање отказа магнетних дискова

SMART технологија може указати на потенцијалне отказе диска, само у случају када вредност праћеног атрибута опадне испод или пређе вредност прага дефинисаног од стране произвођача. Вредност прага се одређује на основу искуства, праћењем рада серије дискова у фабричким тестовима. FMMEA (енг. *Failure Modes, Mechanisms, and Effects Analysis*) је систематска методологија заснована на разумевању односа између физичких карактеристика одређеног система и услова под којим тај систем ради, како би се утврдила подложност система квару под одређеним условима употребе [47]. Она изучава интеракцију компоненти система са корисничким оптерећењем услед којих долази до квара. FMMEA методологија идентификује начине и механизме кварова и анализу њихових ефекта на отказ целокупног система. У овој методологији механизми кварова се класификују према утицају на отказ система и вероватноћи појаве квара, како би се одредили узроци отказу. Коришћењем ове методологије идентификовани су следећи доминантни механизми кварова: хабање магнетног медијума, оштећења заштитних слојева на контакту главе са површином магнетног диска и резонанција актуатора уписно-читајуће главе. Ова методологија захтева детаљно познавање физичких карактеристика магнетног диска и доступне податке добијене из фабричких деградационих тестова спроведених са високим степеном оптерећења, па се она углавном користи од стране произвођача и везана је за одређен модел или фамилију дискова. Независна истраживања се ослањају на податке који су прикупљени током регуларног рада дискова и она захтевају коришћење другачијих методологија.

## 2.7 Машинско учење

Машинско учење користи приступ заснован на прикупљеним подацима на основу којих се може предвидети отказ магнетног диска. Приступ заснован на подацима укључује праћење и анализу функционалних параметара система без увида у физичке процесе који се одвијају унутар система, већ на основу података он креира модел система учећи се на основу претходних инстанци података. Задатак машинског учења је да креира модел система без потребе за експлицитним програмирањем. Алгоритми машинског учења користе скуп података за тренирање из којег се креира модел, чије понашање се побољшава са квалитетом и квантитетом прикупљених података и претходних искустава. Тако формиран модел у стању је да на основу нових улазних података који су из опсега података за тренирање генерише очекивану вредност излаза. Машинско учење, за разлику од вештачке интелигенције, у већини случајева неће

генерисати рационалну вредност излаза уколико се на улаз модела доведу подаци за које модел није трениран.

Традиционално програмирање, приказано на *Слици 12а*, користи улазне податке и програм чију логику је развио и написао човек, који се покреће на рачунару да би се добио одговарајући излаз. Артур Ли Самјуел је 1959. године започео развој области машинског учења, које је пружио рачунарима могућност да уче без потребе да буду експлицитно програмирани од стране човека [48]. Машинско учење приказано на *Слици 12б*, користи скуп познатих улазних и излазних података, над којима рачунар током фазе обуке извршава неки алгоритам машинског учења, где се као резултат извршавања креира сопствени програм, односно логику, чија тачност се може проценити током фазе тестирања. Артур Ли Самјуел је користио машинско учење како би научио рачунар да игра друштвену игру „даме“ (енг. *checkers*).



Слика 12. Поређење (а) традиционалног програмирања и(б) машинског учења

Процес машинског учења започиње припремом квалитетног скупа података, и наставља се израдом модела учења користећи различите алгоритме машинског учења. Избор алгоритама зависи од врсте података и какав задатак је потребно аутоматизовати. За рачунар се каже да учи из искуства с обзиром на неку класу задатака, ако се његово обављање у задатку побољша с искуством. Рачунарски програм, према дефиницији Тома Мичела, учи из искуства  $E$  (скуп података), решавајући задатак  $T$  са неком мером перформанси  $P$  ако се његов рад на решавању задатка  $T$ , мерено перформансом  $P$ , побољшава са искуством  $E$  [49].

У процесу машинског учења користи се појединачан узорак података који називамо *инстанца (опажање)*. Инстанца садржи категоризоване податке, *атрибуте*, који ближе описују неке њене особине и представљају појединачно мерљиво својство података. Инстанце могу поседовати и *циљну променљиву (ознаку)* која дефинише исход (излаз, одговор, резултат) коју модел треба да предвиди. *Модел* је специфична репрезентација одређене врсте знања научена из података применом неког алгоритма машинског учења. *Обука* је поступак у којем се за дати скуп улазних променљивих (атрибута) и очекиваних резултата (циљна променљива), креира модел који ће бити способан пресликати нове податке у једну од категорија. *Предвиђање* представља употребу обученог модела да за дати скуп улазних података генерише предвиђени исход. Током обуке модела могућа су погрешна предвиђања исхода која се мере *функцијом грешке* и она представља меру одступања између очекиваног исхода и исхода којег је предвидео модел. Циљ фазе обуке је да модел минимизује функцију грешке над скупом података како би генерисао што тачније исходе.

Алгоритми машинског учења класификовани су према у четири главне категорије, приказане на *Слици 13*, у зависности од тога који подаци су на располагању систему учења:



Слика 13. Типови машинског учења

Надгледано (енг. *supervised*) учење обучава алгоритам на основу инстанци које укључују улазне податке и придружене циљне одговоре како би се генерисао предиктивни модел приказан на *Слици 14*. Предиктивни модел након процеса обуке је у стању да предвиди тачан одговор када му се представе нови улазни подаци. Овај приступ је сличан људском учењу под надзором учитеља који пружа добре примере ученику који их памти и из тих конкретних примера извлачи општа правила. Надгледано учење користи улазне променљиве  $x$  које алгоритам за учење пресликава у функцију мапирања  $f(x)$  коју тежи да изједначи са познатом излазном променљивом  $Y$ . Циљ надгледаног учења је приближити функцију мапирања  $f(x)$  довољно добро циљном одговору  $Y$ , све док алгоритам не постигне прихватљив ниво перформанси којим се оцењује квалитет мапирања. Овако обучен модел је у стању да за нове улазне податке  $x$  предвиди рационалну вредност излазне променљиве  $Y$ .



Слика 14. Процес изградње предиктивног модела код надгледаног машинског учења

Излазна променљива може садржати дискретне вредности попут класа или ознака, или континуалне нумеричке вредности и у зависности од њеног типа надгледано учење се дели на две категорије:

1. Класификација: Излазна променљива има дефинисане дискретне вредности (ознаке). На пример, у случају куповине дефинисане су ознаке 0 и 1; 1 значи да ће купац купити производ, а 0 значи да купац неће купити. Циљ класификације је предвидети дискретне вредности које припадају одређеној класи и проценити на основу тачности. У случају бинарне класификације модел предвиђа 0 или 1, а у случају вишекласне класификације, модел предвиђа између више дискретних класа као што је препознавање руком писаних децималних цифара.

2. Регресија: Излазна променљива има континуалну вредност попут "цене" или "тежине". Циљ регресије је да модел предвиди излазну вредност која је што ближа стварној вредности излазне променљиве, како би грешка регресионог модела била што мања. Код једноструке регресије излазна променљива зависи само од једне улазне променљиве, а у случају вишеструке регресије од већег броја улазних променљивих. У случају да је излазна вредност линеарно зависна од улазних променљивих, у питању је линеарна регресија, а у супротном ради се о нелинеарној регресији.

Најзаступљенији алгоритми за класификацију код надгледаног учења су:

1. Стабла одлучивања (енг. *Decision Trees*) - овај алгоритам гради хијерархијску структуру одлука, где сваки чвор представља услов који дели податке на основу одређених атрибута. Листови стабла су коначне одлуке. Предност стабала је њихова интерпретабилност, док је главни недостатак склоност претераном уклапању модела, што се може ублажити техникама као што су обрезивање стабла (енг. *pruning*) или коришћењем методе „случајне шуме“ (енг. *Random Forest*).
2. Наивни Бајесов класификатор (енг. *Naïve Bayes*) - овај модел се заснива на Бајесовој теорему и претпоставци да су сви атрибути међусобно независни једни од других (што у пракси ретко важи, али алгоритам и даље често даје добре резултате).
3. Алгоритам потпорних вектора SVM (енг. *Support Vector Machines*) - SVM је класификациони модел који тражи оптималну хиперраван која најбоље раздваја податке две класе. У случајевима када подаци нису линеарно одвојиви, користе се методе кернела (енг. *kernel*) како би се подаци трансформисали у вишедимензионални простор у којем је линеарно раздвајање могуће.
4. Неуронске мреже ANN (енг. *Artificial Neural Networks*) - ови модели су инспирисани начином рада људског мозга и састоје се од слојева вештачких неурона. Дубоке неуронске мреже (енг. *Deep Learning*) користе вишеслојну архитектуру да би училе сложене обрасце из података.

Ненадгледано (енг. *unsupervised*) учење обучава алгоритам на основу скупа примера који укључују само улазне податке без унапред познатог одговора, препуштајући алгоритму да сам одређује зависност између улазних података. Овај тип учења користи се у процесу проналажења структуре, карактеристика и груписања улазних података. Постоје два облика ненадгледаног учења кластеризација тј. груписање и удруживање односно асоцијативност. Кластер значи скуп инстанци који се деле у групе са истим или сличним вредностима атрибута, а који су различити од инстанци у другој групи. Најчешће коришћени алгоритми за кластеризацију код ненадгледаног учења су: *k-means* [50], *DBSCAN* [51], *Hierarchical Clustering* [52]. Други облик ненадгледаног учења је удруживање и има задатак да открије које инстанце из скупа података се појављују заједно. Алгоритми груписања који се користе код ненадгледаног учења су: *Apriori* [53], *FP-Growth* [54], *Eclat* [55].

Полу-надгледано (енг. *semi-supervised*) учење обучава алгоритам са скупом улазних података код којих недостаје неки (често многи) од циљних одговора. Употреба циљних одговора увелико повећава тачност учења и такође је ефикаснија од надгледаног учења.

Подржано (енг. *reinforcement*) учење (учење условљавањем) обучава алгоритам на основу скупа примера који укључују само улазне податке као код ненадгледаног учења, при чему алгоритам добија повратну информацију од стране окружења (позитивну или

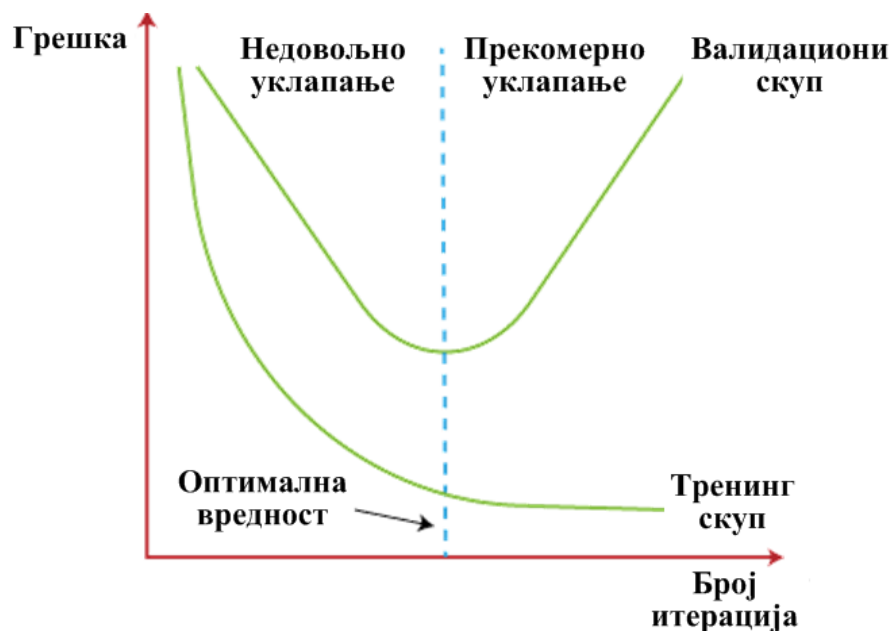
негативну) према решењу које алгоритам предлаже. Овај тип учења заснован је на учењу путем покушаја и грешке при чему наредна акција се одлучује учењем понашања која су заснована на тренутном стању и која ће у будућности минимизирати грешке. Пример подржаног учења је у случају када рачунар сам учи да игра видео игру која омогућава алгоритму да зна исход акција које предузима, а учење се догађа док покушава да избегне оно што открије да је неповољно на резултат игре.

### 2.8 Фазе у процесу машинског учења

Задатак машинског учења је изградња рачунарских модела са високим могућностима предвиђања и генерализације [49]. Циљ надгледаног учења је да се обучи рачунарски модел да предвиди излазе непознате циљне функције која је представљена коначним скупом података за обуку, који се састоји из улазних атрибута и одговарајућих жељених излазних вредности. Скуп атрибута који се користе за обуку модела називамо скуп за тренирање. Алгоритми машинског учења у стању су да открију зависности између атрибута у скупу за тренирање, што омогућава откривање новог знања о некој појави. Ово је нарочито корисно ако је зависност између атрибута толико комплексна да се до новог знања не може доћи аналитичким путем. Обучени модел би требао да предвиди исправне излазне вредности на основу улазних атрибута и скупа за обуку. Алгоритми машинског учења покушавају да пронађу зависност између атрибута максимизирањем параметара перформанси предиктивног модела као што су тачност, прецизност, итд. Поред тога он би требао да буде у стању да добро генерализује излазне вредности када се модел сусретне непознатим скупом улазних вредности. Остваривањем максималних перформанси у фази обуке модела најчешће води ка лошој генерализацији и ова појава се назива претерана обука или претерано уклапање.

Проблем претераног уклапања настаје када модел у процесу обуке тежи да максимизује тражени параметар перформанси при чему долази до значајног повећања комплексности модела. У том случају модел почне да „памти“ податке скупа за обуку (еквивалентно учењу напамет) уместо да генерализује знање из скупа података (еквивалентно учењу са разумевањем). Овакав модел ће остварити јако добре резултате предикције над подацима са којима је обучаван, међутим када се модел сусретне са новим непознатим подацима из скупа за тестирање перформансе предикције ће бити значајно лошије. Као резултат може доћи до генерисања модела који приближно или тачно одговара одређеном скупу података за обуку а не успева да се уклопи у нове податке или поуздано предвиди будуће запажање [56]. Суштина проблема прекомерног уклапања је да модел током обуке погрешно препознаје шум у подацима као релевантно знање и уграђује га у структуру модела, уместо да научи опште законитости [57].

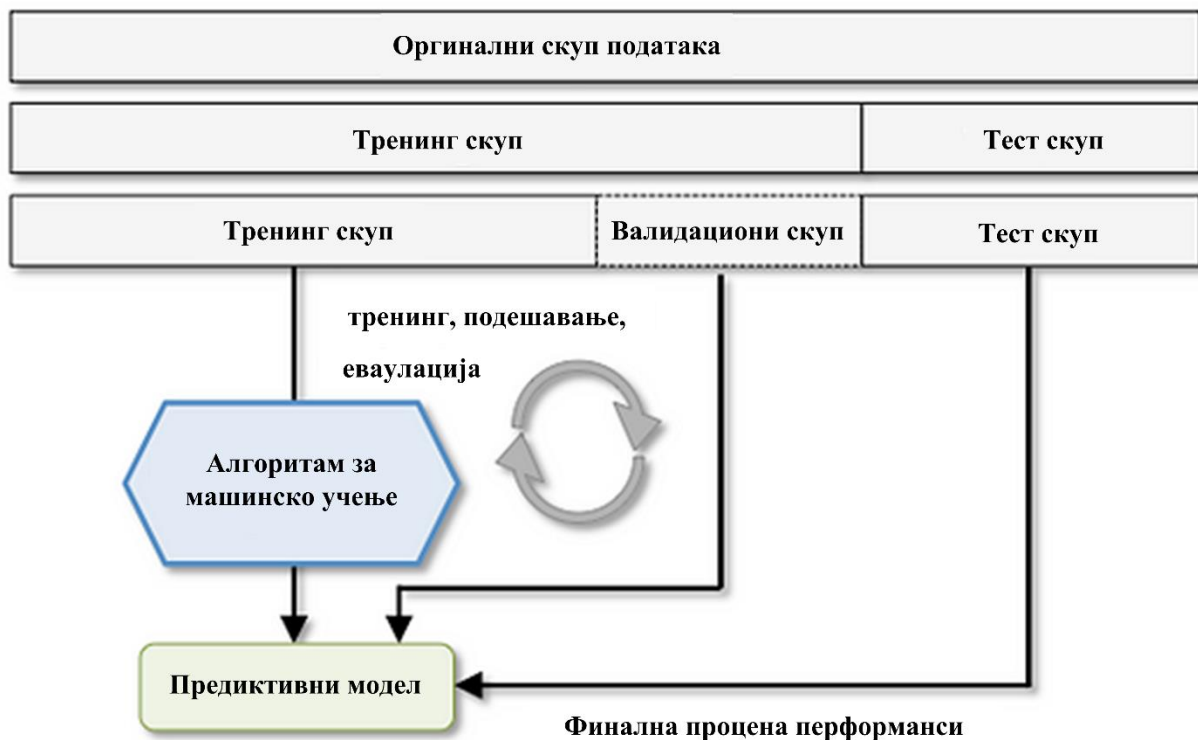
Током процеса обуке модел се обучава са инстанцама за обуку постижући све боље предикционе перформансе што се огледа у смањењу грешке. Циљ процеса тренирања је да се пронађе оптимална граница која не доводи до недовољног ни до прекомерног уклапања као што је приказано на *Слици 15*.



Слика 15. Одабир оптималне вредности уклапања

Како би се избегла појава претераног уклапања модела параметри перформанси предиктивног модела се процењују помоћу података са којима се модел није обучавао. У ту сврху користи се независан скуп података који нису коришћени у поступку тренирања модела и представљају потпуно нова опажања које се називају подаци за тестирање. Коришћењем независних података за процену параметара перформанси модела омогућена је објективна процена валидности генерисаног предиктивног модела. Због тога се скуп података код алгоритама машинског учења дели на две независне целине, од којих се једна користи у фази тренирања модела, а друга у фази тестирања параметара перформанси након обуке модела. Поступак обуке модела најчешће захтева знатно више података за тренирање у односу на број података потребан за тестирање перформанси обученог модела. У пракси најчешће се користи приступ 80:20, што значи да се од укупног броја, 80% опажања користи за скуп за обуку, док се преосталих 20% користи у фази тестирања обученог модела. Такође, распрострањено је коришћење и других односа поделе података, као што су 70:30, 60:40 или 50:50.

С обзиром да су захтеви за добро предвиђање на скупу за обуку и добра генерализација модела на скупу за тестирање међусобно супротстављени, потребно је наћи компромисно решење. Обучени модели поседују подесиве хиперпараметре који могу значајно утицати на перформансе предикције. Како би се одабрала најоптималнија вредност хиперпараметра користи се скуп за валидацију који је независан од скупа за тренирање и скупа за тестирање. Скуп за валидацију користи се како би се на основу његових инстанци одабрала вредност хиперпараметра која максимизује перформансе обученог модела. Поступак обуке се зауставља када се постигне задовољавајући ниво предиктивних перформанси на подскупу за валидацију или када ниво предиктивних перформанси престане да се побољшава. Овај приступ се назива техника раног заустављања (енг. *early stopping*) која омогућава да се модел обучи тако да достигне оптималну границу одлучивања. Фазе у процесу машинског учења које користи скуп за валидацију за подешавање хиперпараметра модела приказане су на *Слици 16*.

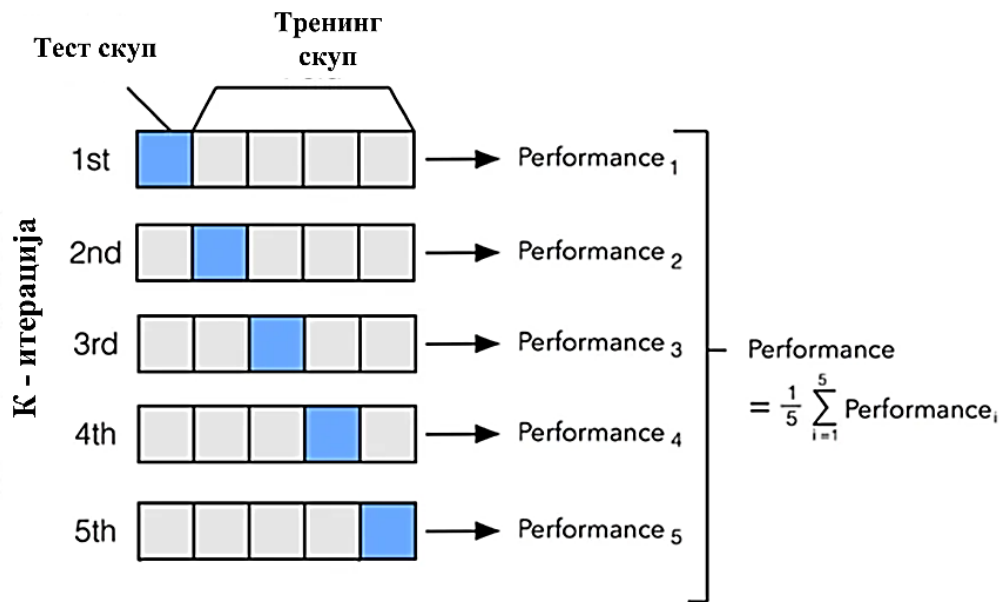


Слика 16. Фазе у процесу машинског учења [58]

Технике унакрсне валидације користе се како би се постигла добра генерализација модела на целокупном скупу података. Уобичајене методе које се користе за унакрсну валидацију су: унакрсна валидација са задржавањем,  $k$ -струка унакрсна валидација и унакрсна валидација са једним опажањем [59].

Код унакрсне валидације са задржавањем, скуп за обуку и скуп за тестирање представљају два независна скупа података између којих нема преклапања. Када се подаци поделе, скупови остају фиксни. Овај тип валидације не користи све податке за тестирање перформанси модела и процена перформанси зависиће од начина на који је скуп података подељен. Скуп података се најчешће дели коришћењем једноставног насумичног узорковања, где се узорци одабирају са једнаком вероватноћом, коришћењем униформне расподеле.

Применом  $k$ -струке унакрсне валидације [60] побољшавају се перформансе обученог модела, јер користи све податке у поступку унакрсне валидације. Ова метода дели скуп података на  $k$  једнаких делова, при чему се  $k-1$  подскупова користи у фази обуке модела а преостали подскуп за тестирање модела. Поступак се понавља  $k$  пута док се сваки од  $k$  подскупова не искористи за тестирање модела, као што је представљено на Слици 17. Перформансе модела се израчунавају на основу средње вредности резултата тестирања перформанси  $k$  итерација. Типичне вредности за параметар  $k$  су  $k=5$  или  $k=10$ . Недостатак ове методе је што се поступак обуке понавља  $k$  пута што захтева знатно више времена за обуку у односу на унакрсну валидацију са задржавањем, па је препоручена да се користи код мањих скупова података.



Слика 17. Метода k-струке унакрсне валидације

Унакрсна валидација са једним опажањем је специјални случај  $k$ -струке валидације код које је параметар  $k$  једнак броју података у скупу, што значи да се само један податак користи у скупу за валидацију а преосталих  $n-1$  за обуку модела. Поступак се понавља  $n$  пута при чему се сваки пут појединачни податак користи за унакрсну валидацију. Овај метод има високу тачност, али је погодан само за мале скупове података јер се поступак обуке понавља  $n$  пута.

## 2.9 Процена перформанси предвиђања

Процена перформанси предвиђања у машинском учењу користи се како би се одредила мера ефикасности и поузданости обученог модела. Како би ова процена била што објективнија, обучени модел врши предвиђања над независним скупом података, који није коришћен у фази обуке. На тај начин се добијају поуздани показатељи генерализације обученог модела у непознатом окружењу. Процена перформанси код надгледаног учења врши се поређењем вредности излаза коју је генерисао обучен модел за одређени скуп улазних података са очекиваном вредношћу излаза која је саставни део скупа за тестирање. У случају да модел генерише нумеричку излазну вредност користе се регресивне метрике, као што су средња апсолутна грешка MAE (енг. *Mean Absolute Error*), средња квадратна грешка MSE (енг. *Mean Squared Error*), корен средње квадратне грешке RMSE (енг. *Root Mean Squared Error*) и коефицијент детерминације  $R^2$  (енг. *R-squared*) [61].

Класификационе метрике пореде превиђене резултате класификације са стварном вредношћу класе из скупа података. Перформансе класификационог модела приказују се помоћу матрице конфузије, у којој је приказан однос између предвиђене класе и стварне класе. У случају бинарне класификације матрица конфузије садржи два реда за приказ резултата предвиђене класе (позитивне и негативне ознаке) и две колоне које садрже резултате из стварне класе (позитивне и негативне ознаке), приказано на *Слици 18*. Све класе које је модел тачно предвидео и које се поклапају са очекиваном класом називају се стварне позитивне и стварне негативне вредности. У случају да модел погрешно одреди класу која се разликује од очекиване класе називамо лажним позитивним вредностима и лажним негативним вредностима. Лажно позитивне вредности су све вредности чија је стварна класа негативна а погрешно су класификоване од стране

модела као позитивне. Насупрот томе, лажно негативне вредности су погрешно класификоване од стране модела као негативне вредности. Идеално обучени модел исправно класификује класе као стварно позитивне и стварно негативне, без лажно позитивних и лажно негативних вредности. На основу ових показатеља у матрици конфузије могу се одредити следеће метрике: тачност, прецизност, осетљивост, стопа погрешних детекција и стопа исправних детекција, позитивни однос вероватноће и негативни однос вероватноће.

|                  |    | Стварна класа              |                            |
|------------------|----|----------------------------|----------------------------|
|                  |    | да                         | не                         |
| Предложена класа | да | Стварно позитивни<br>$t_p$ | Лажно позитивни<br>$f_p$   |
|                  | не | Лажно негативни<br>$f_n$   | Стварно негативни<br>$t_n$ |

Слика 18. Матрица конфузије

Тачност  $T$  (енг. *Assurasy*) представља однос исправних класификација и укупног броја класификација и израчунава се према формули (1).

$$T = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \quad (1)$$

Метрика тачности је непоуздана ако постоји неравномерна расподела инстанци између две класе. На пример, ако у скупу података постоји само 1% отказалих дискова и ако класификатор сваки диск класификује као исправан, игноришући неисправне дискове, тачност класификације износи 99%.

Прецизност  $P$  (енг. *Precision*) или  $PPV$  (енг. *Positive Predictive Value*) представља однос између броја стварно позитивних класификација и укупног броја позитивних класификација који укључује и лажно позитивне класификације (2).

$$P = \frac{T_p}{T_p + F_p} \quad (2)$$

Прецизност представља меру релевантности резултата предвиђања, она тежи вредности 1 код добро обучених модела.

Осетљивост (енг. *Sensitivity*) или опозив (енг. *Recall*) представља однос броја стварно позитивних класификација и збира стварно позитивних и лажно негативних класификација дефинисана према формули (3).

$$TPR = \frac{T_p}{T_p + F_n} \quad (3)$$

Осетљивост се дефинише и као стопа стварно позитивних детекција  $TPR$  (енг. *True Positive Rate*) и она представља способност класификатора да исправно идентификује позитивне инстанце. Метрике прецизности и осетљивости су међусобно зависне, повећањем једне метрике смањује се друга. Повећањем прецизности модела смањује се број лажно позитивних класификација, што за последицу има да се смањује и број стварно позитивних класификација, односно долази до смањења осетљивости.

Метрика која комбинује прецизност и осетљивост назива се F мера (енг. *F-score*). Она користи параметар  $\beta$  који контролише колико се више значаја придаје осетљивости у односу на прецизност према (4).

$$F = \frac{(1 + \beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R} \quad (4)$$

У случају да је подједнак значај прецизности и осетљивости дефинишемо балансирану F1 меру (енг. *F1-score*) према формули (5).

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} = \frac{2 \cdot T_p}{2 \cdot T_p + F_p + F_n} \quad (5)$$

F0.5 мера вреднује прецизност двоструко више у односу на осетљивост према формули (6) и ова мера се најчешће користи за мерење перформанси алгоритама за предикцију отказа магнетних дискова.

$$F0.5 = \frac{(1 + 0.5^2) \cdot P \cdot R}{0.5^2 \cdot P + R} = \frac{1.25 \cdot P \cdot R}{0.25 \cdot P + R} \quad (6)$$

Стопа лажно позитивних детекција FPR (енг. *False Positive Rate*) представља број погрешно класификованих позитивних класа које су стварно биле негативне, односно овај параметар показује стопу лажних детекција према (7).

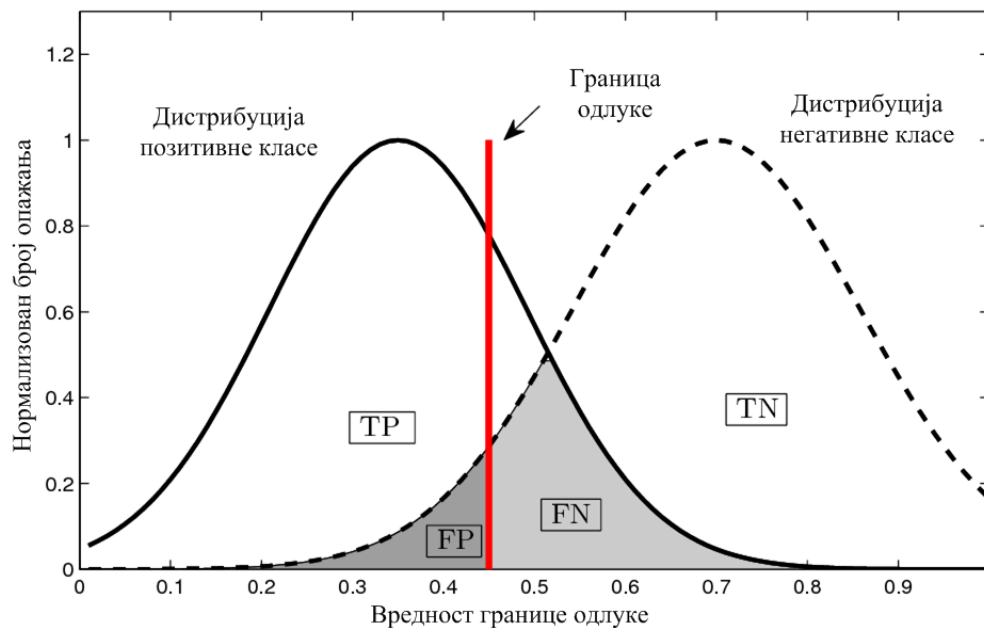
$$FPR = \frac{F_p}{T_n + F_p} \quad (7)$$

Метрика позитивног односа вероватноће LR+ (енг. *Positive Likelihood Ratio*) показује однос између осетљивости и стопе лажно позитивних детекција према (8), а негативни однос вероватноће LR- (енг. *Negative Likelihood Ratio*) је однос између стопе лажно негативних детекција и прецизности према (9).

$$LR+ = \frac{TPR}{FPR} \quad (8)$$

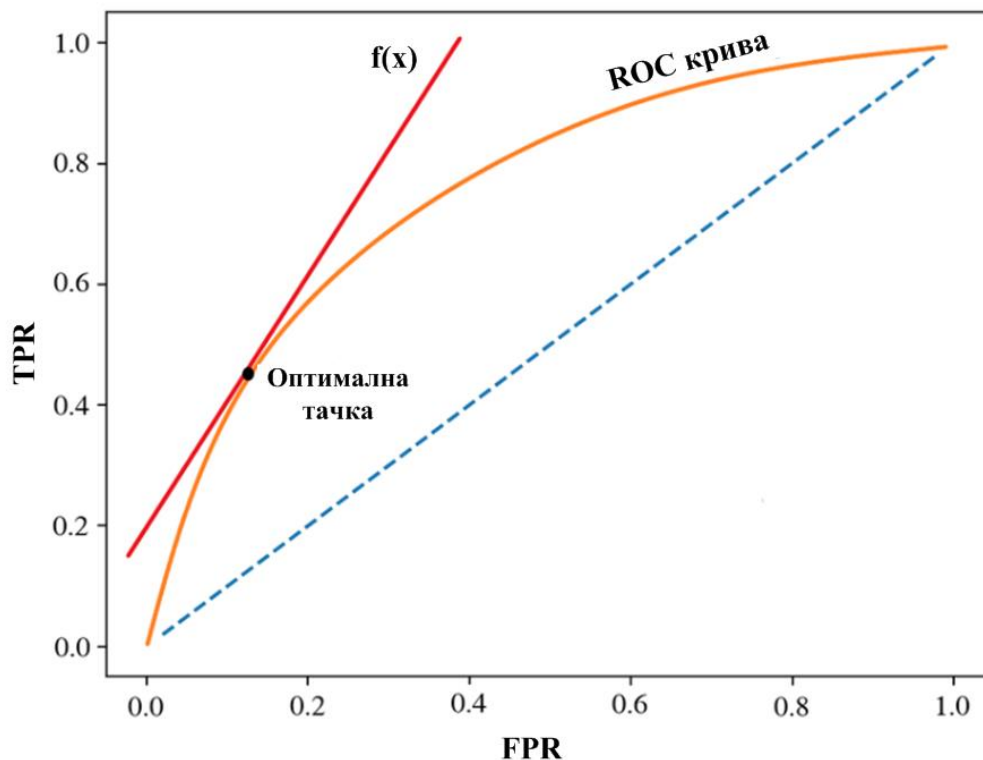
$$LR- = \frac{FNR}{TNR} \quad (9)$$

Типична вредност позитивног односа вероватноће је значајно већа од 1, док је код негативног односа вероватноће значајно мања од 1, што их разликује од осталих метрика које имају вредности између 0 и 1. Предиктивне перформансе модела зависе од подешавања његових параметара. На Слици 19, приказане су дистрибуције вредности позитивне и негативне класе, које се међусобно преклапају услед чега долази до погрешних класификација позитивне класе (FN) и погрешних класификација негативне класе (FP). Променом границе одлуке смањује се број погрешних класификација једне класе док се повећава број погрешних класификација супротне класе. Оптимална вредност границе одлуке тежи да максимизује број тачних позитивних класификација (TP), при чему минимизује број погрешних класификација негативне класе (FP). Како би се пронашла оптимална вредност границе одлуке потребно је испитати предиктивне перформансе модела у опсегу у којем он класификује тачне одговоре са стопом TPR у опсегу од 0.0 до 1.0. Том приликом модел ће генерисати и лажно тачне одговоре који се представљају помоћу FPR параметра, чија вредност се креће у опсегу од 0.0 до 1.0.



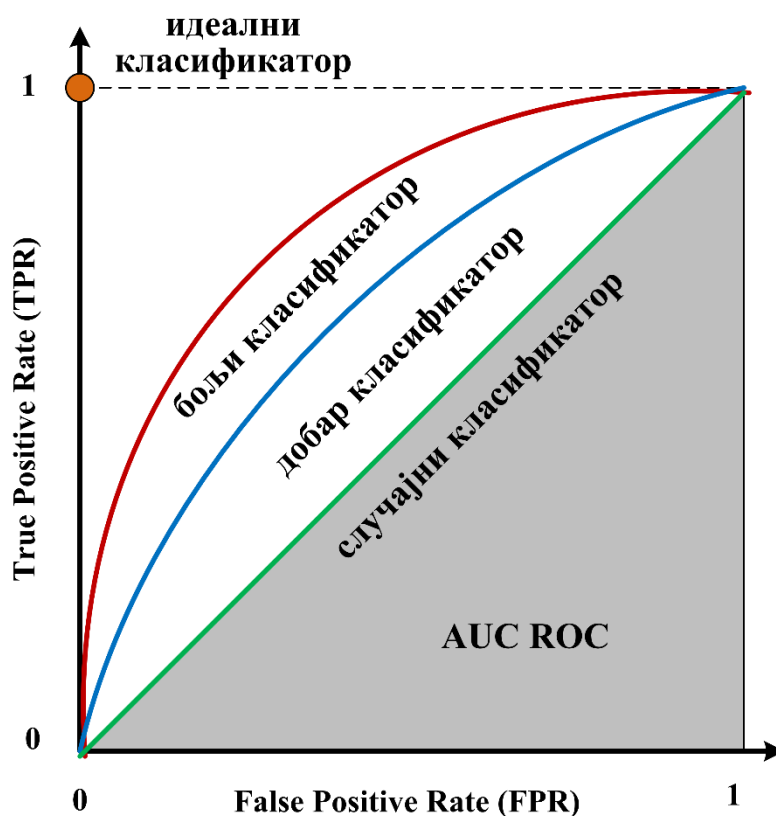
Слика 19. Утицај границе одлуке на погрешне класификације [62]

Подешавањем прага класификације модела за сваку нову вредност прага мењаће се матрица конфузије и сви показатељи перформанси који се из њих израчунавају. Параметри TPR и FPR су међусобно независни, и њихов однос за различите вредности прага дефинише ROC (енг. *Receiver Operating Characteristic*) криву, приказану на Слици 20. На  $x$  оси ROC криве приказује се вредност стопе лажно позитивних детекција FPR, а на  $y$  оси приказује се стопа стварно позитивних детекција TPR. Променом прага класификације вредности TPR и FPR метрика ће се мењати у интервалу 0 до 1 и добиће се одговарајућа крива, за коју можемо одабрати најоптималнију вредност прага класификације која се налази на “колелу” ROC криве. Што је ROC крива ближа горњем левом углу графика то су боље предиктивне перформансе модела.



Слика 20. Одређивање оптималне вредности прага класификације помоћу ROC криве

Предиктивне перформансе одабраног класификатора могу се изразити израчунавањем површине која се налази испод ROC криве. Вредност ове метрике назива се AUC-ROC (енг. *Area Under the Receiver Operating Characteristic*) и она представља збирну меру перформансе неког модела за све могуће вредност прага класификације. Вредности овог параметра се крећу између 0 и 1. У случају насумичног класификатора вредност AUC-ROC параметра износи 0.5. Што је вредност параметра AUC-ROC већа то су предиктивне перформансе боље и код идеалног класификатора ова вредност је једнака јединици, као што је приказано на *Слици 21*. Предност ове метрике је што приказује генерализоване перформансе модела независно од одабраног прага класификације. У случају да постоје јако велике разлике у значају лажно позитивних у односу на лажно негативне исходе предикције модела потребно је минимизирати једну од ове две врсте грешака у класификацији. Примера ради код предикције отказа магнетних дискова циљ је да се смањи стопа лажно позитивних детекција отказа диска како би се смањиле непотребне рекламације исправних дискова што ствара додатне трошкове произвођачима. Из тог разлога AUC-ROC метрика није погодна за овакав тип оптимизације класификационог модела.



Слика 21. Поређење модела класификатора на основу површине AUC-ROC криве

## 2.10 Преглед модела за предикцију отказа магнетног диска

С обзиром на проблематику предвиђања отказа магнетних дискова која се ослања на скупове података, који садрже класификоване односно означене инстанце података са класама отказао или исправан, углавном се користе надгледани или полу-надгледани модели учења [63]. На *Слици 22* су приказане различите категорије модела машинског учења који се користе за предикцију отказа.



Слика 22. Класификација алгоритама машинског учења за предвиђање отказа [63]

Модели вероватноће за предикцију отказа магнетних дискова базирани су на дистрибуцији вероватноће атрибута. У случају наивног Бајесовог класификатора за сваку од класа из скупа за тренирање одређују се параметри нормалне дистрибуције сваког атрибута за исправне и отказале дискове [15]. На основу скупа за тестирање, класификује се свака инстанца за претходно одређене дистрибуције вероватноће. Класификатор за неку нову инстанцу одређује вероватноћу вредности атрибута за класе исправан/отказао и на основу њиховог производа одређује укупну вероватноћу на основу које се класификује да ли инстанца припада класи исправних или отказалих дискова. Овај алгоритам је надограђен тако што користи вишеструке инстанце у предикцији са наивним Бајесовим класификатором [14]. У истраживањима за предвиђање отказа коришћен је Wilcoxon rank-sum тест статистичких хипотеза [16], [17]. Новија истраживања заснована су на употреби скривених Марковљевих модела, који представљају статистичке моделе који се користе за опис догађаја који су зависни од интерног стања система [18], [64], [65], [66], [67]. TSP алгоритам користи параметарску дистрибуцију Махаланобисовог растојања како би предвидео отказ диска [20].

Модели засновани на правилима користе табелу правила на основу којих се врши поређење вредности атрибута, како би се класификовало да ли ће доћи до отказа магнетног диска. Табела правила на почетку обуке обухвата најосновнија правила, базирана на вредностима које прелазе неки праг, која се генеришу из првих инстанци скупа за обуку. Ова правила се примењују на нове инстанце из скупа за обуку и у зависности од исхода врши се прилагођавање параметара правила и усклађивање са претходним правилима пре него што се додају у постојећу табелу правила. Ови модели су једноставни за имплементацију код магнетних дискова јер обухватају низ једноставних провера услова дефинисаних у табели правила. Алгоритам правила максималне вероватноће MLRules (енг. *Maximum Likelihood Rules*) користи се за класификацију отказа проценом вероватноће [68], [69]. Експертни систем заснован на правилима користи C4.5 алгоритам за генерисање правила [70].

Модели засновани на инстанцама користе меру удаљености између инстанце и скупа за обуку како би предвидели отказ диска. Као мера удаљености може се користити Еуклидско растојање за међусобно независне атрибуте који су скалирани на сличан опсег вредности. У већини случајева атрибути су приказани у различитим скалама и међусобно могу бити зависни па је тада погодно користити Махаланобисово растојање које мери

удаљеност између инстанце и дистрибуције скупа за обуку, узимајући у обзир коваријансу између атрибута [47], [19], [71]. Када растојање између инстанце и скупа за обуку пређе одређену вредност прага, таква инстанца се проглашава отказом. Вредност прага се одређује у поступку надгледаног учења користећи инстанце отказалих дискова. У случају ненадгледаног учења неправилно понашање дискова може се одредити алгоритмом  $k$ -најближих суседа. Овај алгоритам мери удаљеност сваке инстанце до својих  $k$ -најближих суседа и за инстанце које су значајно удаљене од осталих инстанци се проглашавају за аномалије односно отказ [72], [73].

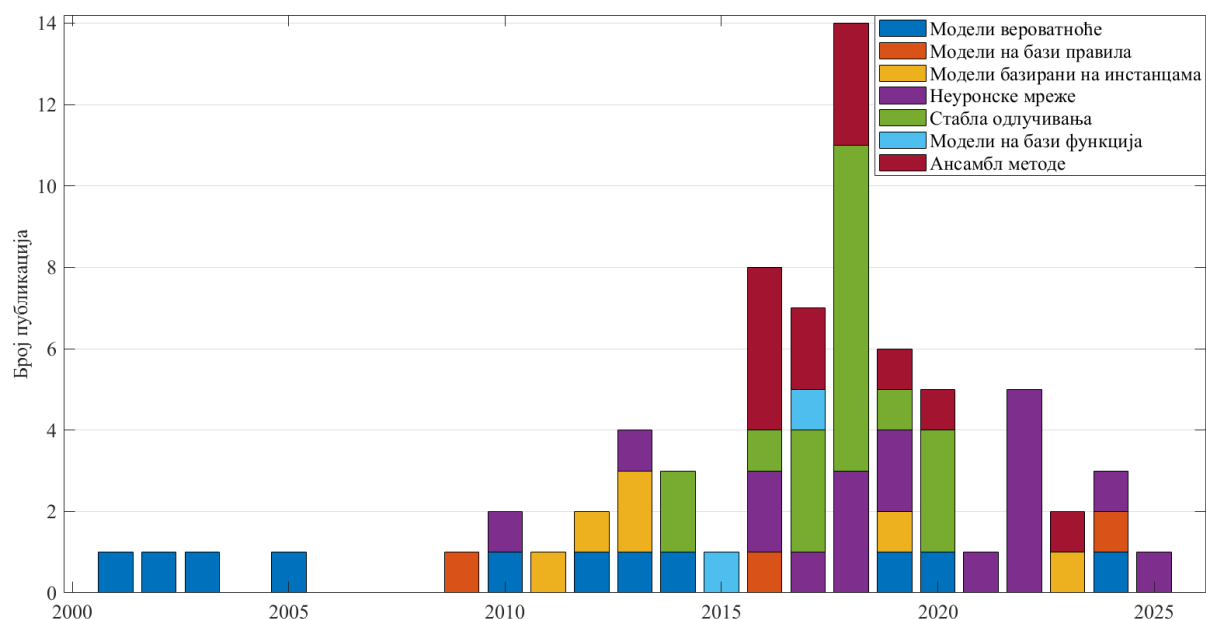
Модели засновани на апроксимацији функција најчешће користе логистичку регресију помоћу које се вредност атрибута скалира у сигмоидну (енг. *sigmoid*) функцију како би се одредила вероватноћа отказа [74]. Применом линеарне регресије могуће је одредити време за које ће вредност одређеног атрибута доћи до критичне вредности прага која сигнализира отказ [75].

Стабла одлучивања припадају надгледаном учењу, најчешће се користе за класификацију и регресију као алгоритам машинског учења. Стабло одлучивања састоји се од чворова где сваки чвор представља тест неког од SMART атрибута. Чвор се даље дели на гране а свака грана стабла одговара некој вредности SMART атрибута посматраног чвора, односно представља исход теста [76], [77], [78]. Насумичне шуме побољшавају недостатке предикције приликом коришћења појединачних стабала тако што комбинују мноштво стабала одлучивања и врше агрегацију резултата класификације тих стабала [79], [80], [81].

Често коришћена метода за предвиђање отказа магнетних дискова су вештачке неуронске мреже. С обзиром да SMART параметри представљају временске серије они су погодни за примену рекурентних неуронских мрежа RNN (енг. *Recurrent Neural Network*) [82], [83], [84], [85], [86]. Истраживачи су често користили мреже дуге краткорочне меморије LSTM (енг. *Long Short Term Memory*) у предикцији отказа магнетних дискова [87], [88], [89].

Веома чест случај је да се комбинацијом различитих алгоритама машинског учења могу добити значајно бољи резултати предвиђања у односу на појединачно коришћене алгоритме. Методе уклапања користе двостепени приступ у којем на нижем степену користе једну врсту алгоритма, а резултате предикције се прослеђују вишем степену у којем други тип алгоритма доноси коначну одлуку. Пример оваквог алгоритма је комбинација XGBoost (енг. *eXtreme Gradient Boosting*) стабала одлучивања са мрежама дуге краткорочне меморије LSTM [90]. Модели засновани на Бајесовим мрежама [91], [92], [93] користе графички модел вероватноће који приказује скуп случајних променљивих и њихове међусобне зависности.

На Слици 23. приказан је временски тренд истраживања предикције отказа магнетних дискова класификован на основу коришћеног модела. Прва истраживања била су заснована на примени једноставних модела вероватноће за предикцију отказа, који су били засновани на појединачним SMART атрибутима. Каснија истраживања ослањала су се на примену модела машинског учења, који су остваривали значајно боље резултате предикције отказа у односу на моделе вероватноће. Најпре су уа предикцију отказа дискова примењивани добро познати алгоритми, попут SVM класификатора и стабала одлучивања. Истраживачи су развијају сложеније моделе за предикцију отказа који су остваривали знатно боље предикционе перформансе у односу на традиционалне алгоритме машинског учења.



Слика 23. Тренд коришћења разних типова алгоритама за предикцију отказа

Претходних неколико година истраживачи су најчешће примењивали алгоритме засноване на рекурентним неуронским мрежама и LSTM мрежама дуге краткорочне меморије. На основу приказаног тренда на Слици 23. долази се до закључка да је тема предикције отказа дискова и даље веома актуелна.

### 3 Прикупљање и обрада скупова података

У овом поглављу описан је поступак анализе и обраде података који су доступни у облику отворених скупова података. Обрада података подразумева примену статистичких метода које се користе за уклањање абнормалних вредности и попуњавање недостајућих вредности података. У наставку поглавља, представљене су технике које се користе за балансирање скупа података.

#### 3.1 Прикупљање података

Најважнији предуслов за успешну примену алгоритама машинског учења је поседовање довољно великог скупа података при чему сами подаци морају бити задовољавајућег квалитета. Прикупљени подаци могу бити необрађене чињенице, нумеричке вредности, текст, звук или слике. Најчешће су представљени у изворном облику и не могу се директно обрађивати од стране алгоритама машинског учења. Скуп података представља структурирану колекцију инстанци које се односе на одређену појаву која је предмет истраживања. Они настају као резултат прикупљања података помоћу мерења, анкета, студија или настају спонтано као резултат извршавања неког пословног или индустријског процеса. Скупови података су најчешће представљени у табеларној форми где сваки ред у табели одговара једној инстанци, док колоне одговарају одређеним атрибутима којим се описује посматрана појава. Такав начин приказивања погодан је и за човека као и за рачунарску обраду података.

У следећој Табели 6. приказан је пример дела скупа CMRR података за магнетне дискове. У табели је приказано 11 различитих опажања које називамо инстанце, које су прикупљене за различите дискове у различитим временским условима. Свако опажање садржи низ атрибута са одговарајућим нумеричким и категоријским описом инстанце. Први атрибут *Serial* представља серијски број диска који је категорисан као нумеричка вредност. Затим следи атрибут *Frame* који представља редни број записа а поред њега се налази атрибут *Hours* који представља број радних сати диска. Следећа три атрибута садрже нумеричке вредности којима се описују температура *Temp1*, број читања *Reads* и број уписа *Writes* на магнетни диск. У последњој колони налази се атрибут бинарне класе *Class* којим је описано да ли је магнетни диск отказао (1) или је исправан (0).

Табела 6. Пример дела скупа CMRR података за магнетне дискове

| Serial | Frame | Hours   | Temp1 | Reads    | Writes   | Class |
|--------|-------|---------|-------|----------|----------|-------|
| 100001 | 20    | 0.133   | 10    | 120832   | 120842   | 1     |
| 100001 | 22    | 0.166   | 7     | 120832   | 120852   | 1     |
| 100001 | 23    | 0.183   | 7     | 120832   | 120841   | 1     |
| 100001 | 25    | 0.216   | 9     | 131072   | 131078   | 1     |
| 100002 | 46    | 77.833  | 29    | 16661145 | 16478748 | 1     |
| 100002 | 47    | 79.833  | 29    | 17905440 | 17512856 | 1     |
| 100002 | 48    | 81.833  | 30    | 17952464 | 17570840 | 1     |
| 100365 | 26    | 720.267 | 50    | 8321122  | 3677769  | 0     |
| 100365 | 27    | 722.267 | 48    | 32142780 | 200118   | 0     |
| 100365 | 28    | 724.267 | 49    | 55630904 | 635185   | 0     |
| 100365 | 29    | 726.267 | 49    | 19732248 | 3714737  | 0     |

Скупови података морају бити прикупљени систематично, при чему је неопходно да се уз саме податке обезбеде и мета подаци који ближе описују значење и начин прикупљања података. У академској заједници све више се промовише принцип отворене науке који се односи и на скупове података који се прикупљају у сврху отворених научних истраживања. Како би ови подаци могли бити дељени са научном заједницом они морају испунити следеће FAIR принципе:

- **F** (енг. *Findable*) - Могућност лаког проналажења – подаци и мета подаци су организовани тако да их истраживачи могу лако пронаћи и користити. Они захтевају коришћење трајних дигиталних идентификатора објекта DOI (енг. *Digital Object Identifier*);
- **A** (енг. *Accessible*) - Доступност – скупови података су смештени на платформама са којих се могу преузети коришћењем стандардизованих комуникационих протокола;
- **I** (енг. *Interoperable*) - Интероперабилност – подаци су представљени у широко коришћеном отвореном формату за представљање података као што је .csv (енг. *comma separated values*) формат са вредностима раздвојених зарезом;
- **R** (енг. *Reusable*) - Вишекратна употреба – скуп података садржи добро дефинисане мета податке који олакшавају разумевање значења података. Овај корак дефинише и ауторска права над скупом података везаних са коришћењем, дистрибуцијом, комерцијализацијом и изменом података.

Скупови података представљају колекцију опажања неке појаве која је описана низом параметара односно атрибута. Опажања могу припадати једној или вишеструким инстанцама неке појаве, а могу бити и прикупљени у дужем временском периоду у виду временских серија. Подаци се врло често представљају у нумеричком облику како би се убрзало учење модела. Ако су подаци у облику текста потребно је претварање у нумерички облик у форми низа карактера. Слика се може претворити у матрицу димензија  $n \times n$ , а вредност сваког елемента матрице представља нумеричку вредност интензитета боје за одређени пиксел слике.

У овој дисертацији предмет истраживања је предикција отказа магнетних дискова на основу анализе скупа података о раду дискова који су отказали у нормалним условима рада. Прикупљање података о отказалим дисковима врши се током његовог животног века све до тренутка његовог отказа. С обзиром да су магнетни дискови високо поуздани уређаји са годишњом стопом отказа 1% до 3% годишње [7], неопходно је прикупљање велике количине података из велике популације дискова. Ови подаци се могу прикупити из центара за складиштење података, прикупљањем података од корисника на терену или спровођењем деградационих тестова.

Подаци о раду дискова из центара за складиштење података су најприкладнији за истраживања јер укључују податке из велике популације дискова који раде у истим условима експлоатације. Један од најпознатијих скупова података је из *Backblaze* центра података и он садржи податке о раду стотинама хиљада дискова од 2013. године до данас [94]. Међутим, проблем са овим скуповима података је јако велика неуравнотеженост класа где постоји јако мали број инстанци са отказима, насупрот милионима инстанци без отказа диска што је јако неповољно за већину алгоритама машинског учења који преферирају балансиране скупове података.

Подаци прикупљени од корисника на терену представљају податке дискова који су отказали током гарантног рока и враћени су произвођачу. Отказали дискови су радили у хетерогеним условима код различитих корисника. Најкоришћенији скуп података овог

типа је из CMRR Центра за истраживање магнетног снимања при *University of California, San Diego* (UCSD) [14], који садржи скуп података о дисковима отказалим код корисника и исправним дисковима који су прошли фабричке тестове. С обзиром да овакви скупови података садрже податке о великом броју отказалих дискова, они имају добру уравнотеженост класа у скупу података.

Подаци прикупљени у деградационим тестовима симулирају оптерећење које је значајно веће од номиналног како би изазвали квар диска у краћем временском периоду. Ови подаци су добијени у различитим деградационим тестовима спроведеним са мањим бројем дискова и нису толико употребљиви, јер су дискови радили у много неповољнијим условима у односу на стандардне услове експлоатације. Најпознатији деградациони скуп података прикупљен је од компаније *Quantum Inc.*, који садржи податке од два модела дискова који су тестирани у трајању од два до три месеца [95]. Овај скуп података има велику неуравнотеженост класа као и у случају *Backblaze* скупа података. У Табели 7. приказане су карактеристике различитих скупова података који се користе за предикцију отказа магнетних дискова.

Табела 7. Поређење различитих скупова података магнетних дискова [63]

| Скуп података               | Backblaze [94]                                 | CMRR [14]                                                                | Quantum [95]                                                              |
|-----------------------------|------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|
| Метод прикупљања података   | Прикупљање у центру података                   | Исправни дискови из фабричких тестова, неисправни дискови из рекламације | Деградациони тестови                                                      |
| Број дискова                | > 300. 000 различитог типа, 4372 отказала      | 369 дискова истог типа, 191 отказао                                      | 1936 дискова типа А са 9 отказалих<br>1808 дискова типа В са 27 отказалих |
| Временски период прикупљања | 2013 – 2024                                    | 25 дана                                                                  | 2 – 3 месеца                                                              |
| Интервал између инстанци    | 24 часа                                        | 2 часа                                                                   | 30 минута                                                                 |
| Број SMART параметара       | 62 параметра у нормализованом и сировом облику | 64 параметра у сировом облику                                            | 15 кумулативних и инкременталних параметара                               |

Магнетни дискови су сложени електромеханички уређаји који се састоје од различитих подсистема чији квар може довести до отказа магнетног диска. Након завршетка производње магнетног диска неопходно је спровести низ фабричких тестова како би се утврдило да ли диск испуњава очекиване параметре перформанси, ови тестови обухватају низ сценарија у којем се врше насумични или секвенцијални уписи и читања података на диск, током којих магнетни диск мора остварити прописане нивое перформанси. У случају да се у току производње магнетног диска појавио дефект он може утицати на перформансе магнетног диска. Циљ фабричких тестова је благовремено детектовање дефеката у раду диска како би се избегло да се магнетни дискови са дефектом испоруче купцима. Током фабричких тестова прикупљају се подаци у виду SMART параметара који се могу користити за предвиђање отказа магнетних дискова.

Скупови података фабричких тестирања дискова нису доступни широј академској заједници. Редак изузетак је скуп отворених података који је обезбедио CMRR при Универзитету Калифорније у Сан Дијегу, који садржи податке од 369 дискова неименованог модела. CMRR скуп података не садржи информације о произвођачу и типу магнетног диска.

Скуп података садржи 178 исправних дискова који су прошли фабричке тестове у контролисаним условима при чему сваки диск садржи 300 узорака SMART параметара који су прикупљени у двочасовним интервалима. Узорак представља скуп вредности свих параметара појединачног диска у одређеном интервалу. Атрибути диска укључују серијски број диска, број радних сати и 60 SMART параметара. У скупу се налази и 191 диск који су враћени произвођачу од стране корисника јер су отказали током рада. С обзиром да диск може запамтити само 300 последњих узорака код отказалих дискова доступно је 600 часова рада пре наступања отказа. Укупан број инстанци у скупу података износи 68411.

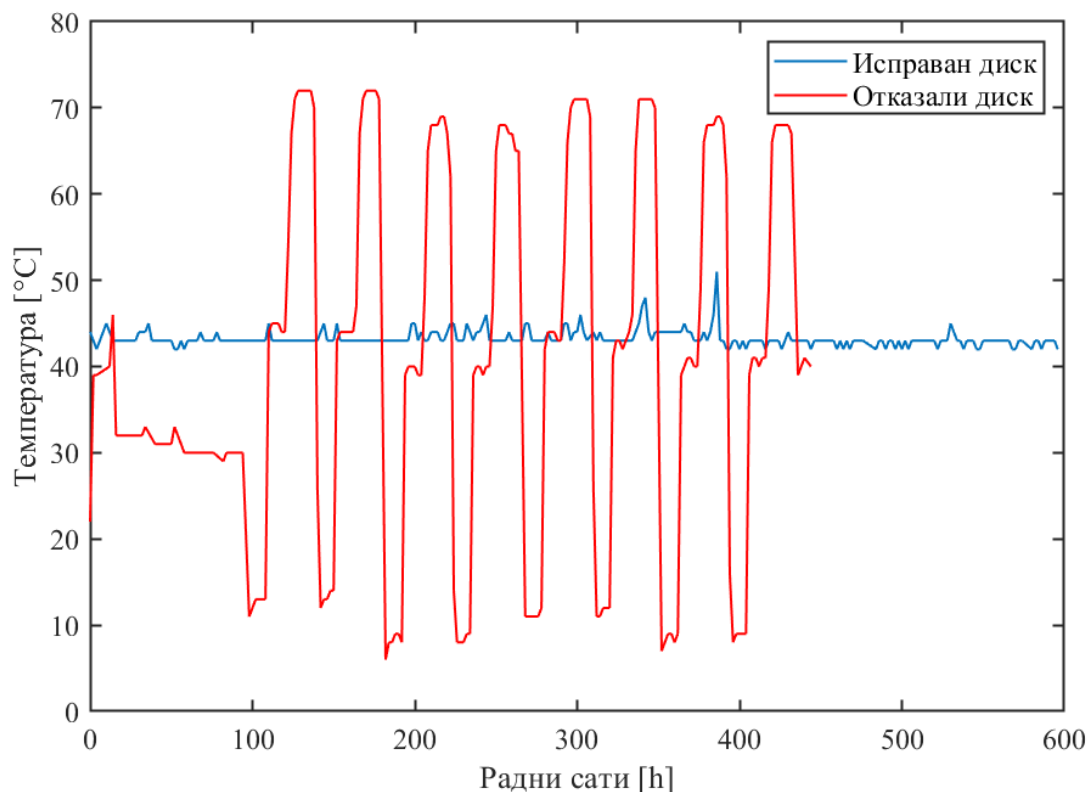
Магнетни дискови су поуздани уређаји са годишњом стопом кварова која се креће од 1% до 3% тако да алгоритми за предвиђање кварова морају бити отпорни на генерисање лажних аларма (ситуација када алгоритам предвиђа квар а диск ради исправно). Прихватљива стопа лажних аларма може бити максимално 0,1%. С обзиром да скуп података садржи релативно мали број од 178 исправних дискова, детекција једног лажног отказа резултоваће стопом од 0.56% лажних отказа. Скуп података доступан је у виду датотеке у *arff*. формату, која се састоји од два одељка, одељак заглавља који садржи називе атрибута и одељак података у којем су наведене инстанце података чије су вредности раздвојене зарезима.

Листа атрибута CMRR скупа укључује:

- *Serial* - серијски број диска који се креће у опсегу од 100001 до 100369;
- *Frame* - редни број узорка у опсегу од 0 до 299;
- *Hours* - број радних сати диска;
- *Hours before failure* - време до отказа;
- *Temp* - шест атрибута за мерење температуре диска;
- *FlyHeight* - шеснаест атрибута за мерење висине уписно-читајуће главе;
- *Servo* - десет атрибута за серво позиционирање;
- *Read Error* - двадесет атрибута за грешке приликом читања;
- *Write Error* - један атрибут за грешке приликом уписа;
- *CSS* - број циклуса укључења;
- *PList* - фабричку листу лоших сектора;
- *Glist* - три корисничке листе лоших сектора;
- *Reads* - укупан број прочитаних података са диска;
- *Write* - укупан број уписаних података на диск;
- *Class* - ознаку класе да ли је диск отказао или не.

На *Слици 24* приказана је вредност параметра температуре *Temp4*, за исправни диск са серијским бројем 100302 и отказали диск са серијским бројем 100003 из које се може запазити неисправан рад диска пре његовог отказа. Услед неправилног рада долазило је до великих скокова радне температуре која је достигала 70°C, па је убрзо након тога

дошло до његовог отказа. Насупрот томе исправни диск није показивао значајне промене радне температуре током посматраног периода.



Слика 24. Приказ радне температуре Tempr4 исправног и отказалог диска

## 3.2 Обрада скупа података

Обрада података је задатак трансформације података из датог облика у много кориснији облик, тј. чинећи га смисленијим и информативнијим. Процес обраде података се може аутоматизовати применом алгоритама машинског учења, математичког моделирања и статистичких тестова.

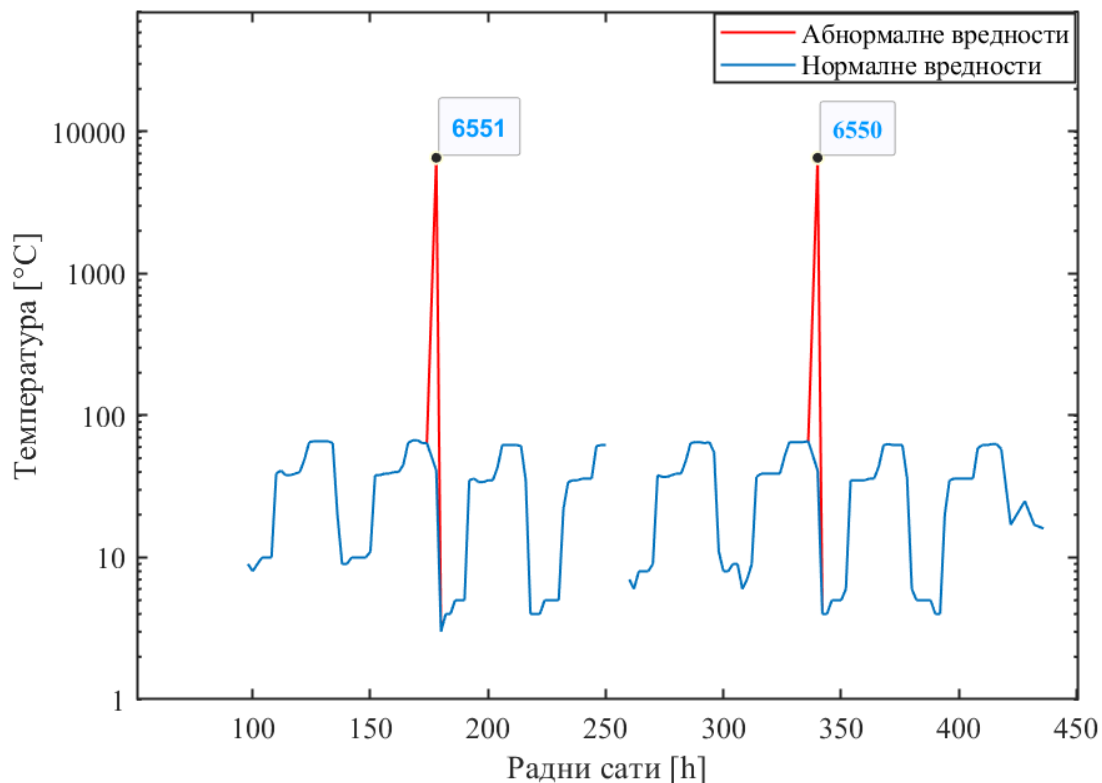
У процесу прикупљања података, одређени подаци могу имати недостајуће вредности за одређене атрибуте, који морају бити испуњени одговарајућим вредностима како би се применили алгоритми машинског учења. Недостајуће вредности за нумеричких података могу се заменити средњом вредношћу атрибута, док недостајуће вредности за категоријске атрибуте могу се заменити најчешће заступљеном вредношћу атрибута.

Прикупљени подаци осим недостајућих вредности могу садржати и неконзистентне односно абнормалне вредности. Абнормалне вредности (енг. *outliers*) представљају вансеријске вредности атрибута које се значајно разликују од осталих вредности. Појава оваквих вредности може бити изазвана грешкама у мерењу или извођењу експеримената, прикупљању и складиштењу података.

Абнормалне вредности се значајно разликују од осталих вредности у скупу података, и оне најчешће настају као резултат грешака приликом прикупљања података. С обзиром да због својих високих вредности могу негативно утицати на резултате предикције ове вредности је потребно кориговати пре примене алгоритама машинског учења. Услед великог броја узорака у скупу података неопходно је користити алгоритме за детекцију и корекцију абнормалних вредности. Неки од најчешће коришћених

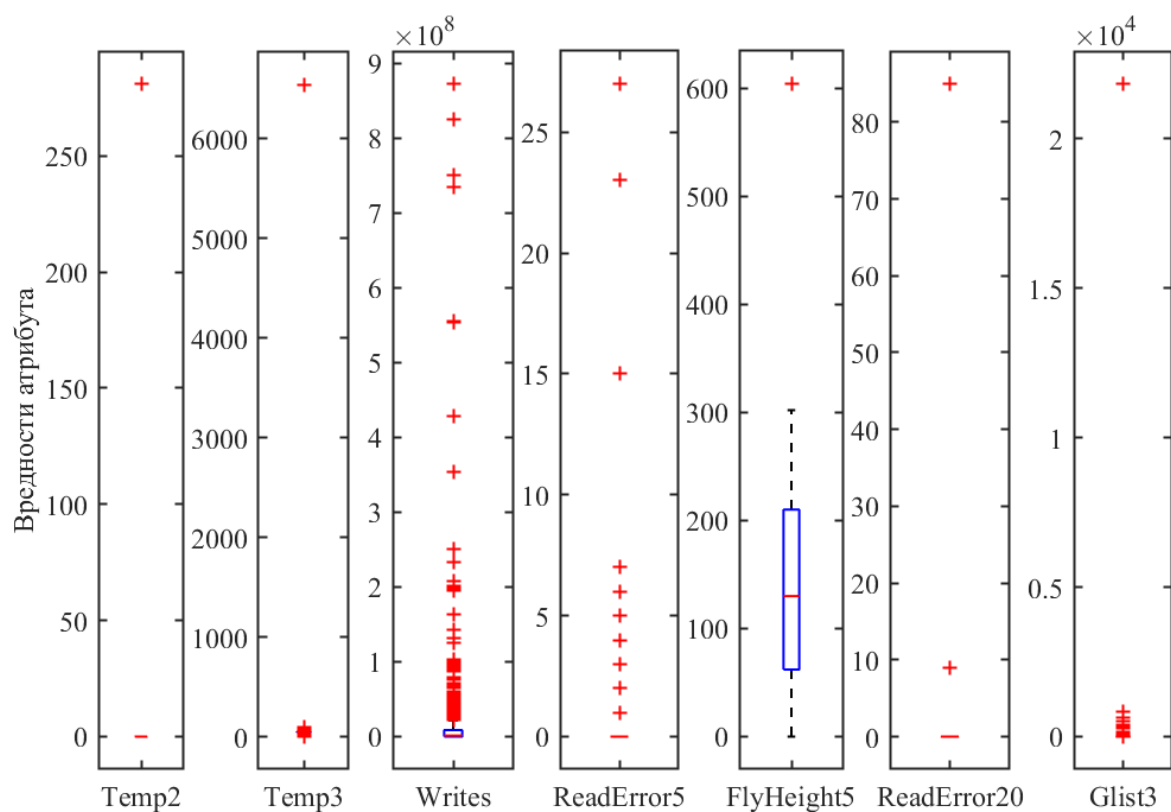
алгоритама су статистички тест абнормалних вредности и алгоритам заснован на међуквartilном растојању. Када су подаци нормално распоређени све вредности које су удаљене одређен број стандардних одступања од средње вредности дистрибуције сматрају се абнормалним вредностима. У том случају се користи статистички тест абнормалних вредности. Најчешће коришћено растојање у пракси је  $\pm 3$  стандардна одступања, која ће препознати 0.3 % података као абнормалне вредности. У случају да је удео абнормалних вредности у подацима мањи или већи од наведеног потребно је користити другачије вредности границе одлуке. Када се вредност препозна као абнормална, њена вредност се замењује средњом вредношћу дистрибуције. Ако подаци не прате нормалну расподелу или нису симетрично дистрибуирани примењује се метода међуквartilног растојања.

Скуп података CMRR у изворном облику садржи одређени број инстанци код којих су присутне абнормалне вредности атрибута. Ове вредности је потребно отклонити пре примене алгоритама машинског учења, како би њихов утицај на резултате предикције био минималан. На *Слици 25* црвеном бојом приказане су изворне вредности атрибута *Temp3*, за магнетни диск са серијским бројем 100004, чије вредности су последица грешака насталих у процесу прикупљања података, које значајно одступају од нормално очекиваних вредности температуре.



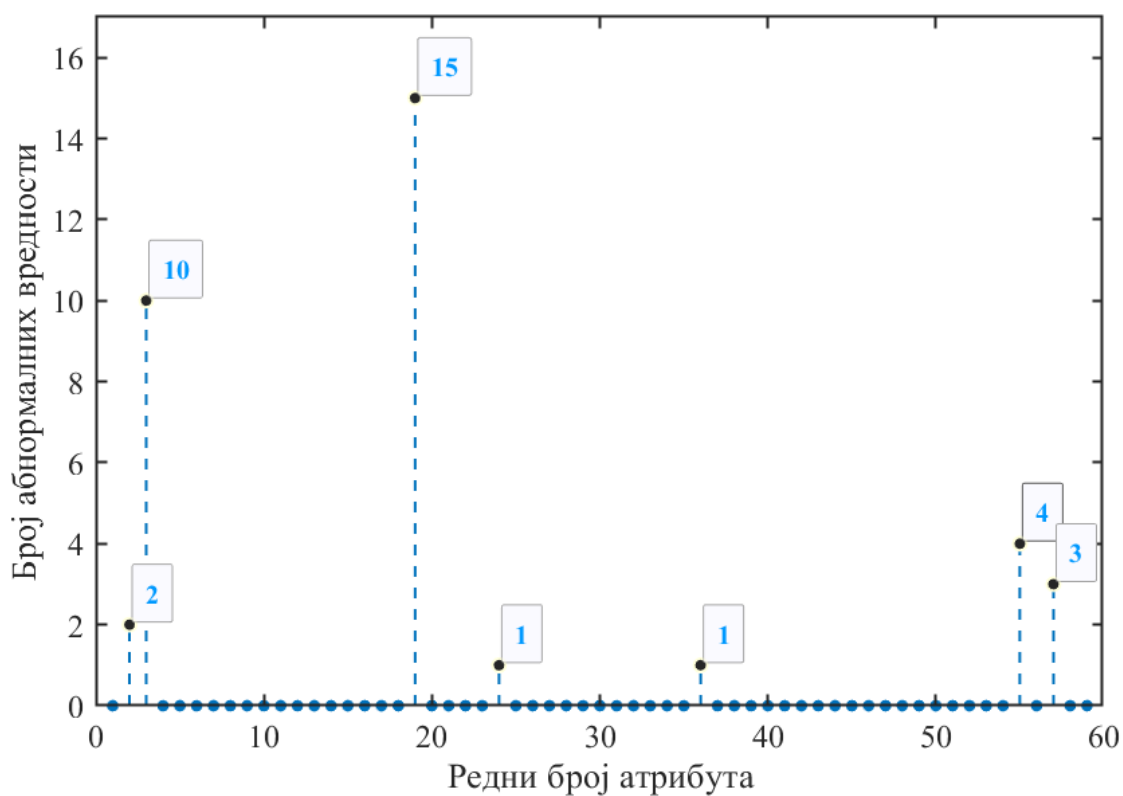
Слика 25. Абнормалне вредности температуре за параметар *Temp3*

Абнормалне вредности је неопходно идентификовати а затим уклонити, при чему се њихова вредност мења са статистички релевантним садржајем, у овом случају замењене су са средњом вредношћу нормалних инстанци за тај атрибут. Овакве вредности се идентификују коришћењем међуквartilног растојања које приказује дистрибуцију вредности неког атрибута. Абнормалне вредности препознају се по значајној удаљености у односу на нормалне вредности. Међуквartilна растојања су одређена за све вредности атрибута, а на *Слици 26* приказана су за седам атрибута код којих су идентификоване абнормалне вредности. Корекцијом ових вредности добија се тренд који прати нормални ток вредности атрибута као што је означено плавом бојом на *Слици 25*.



Слика 26. Идентификовање абнормалних вредности атрибута

За претходно идентификоване атрибуте које садрже абнормалне вредности извршена је корекција и укупан број коригованих вредности по атрибутима приказан је на Слици 27.

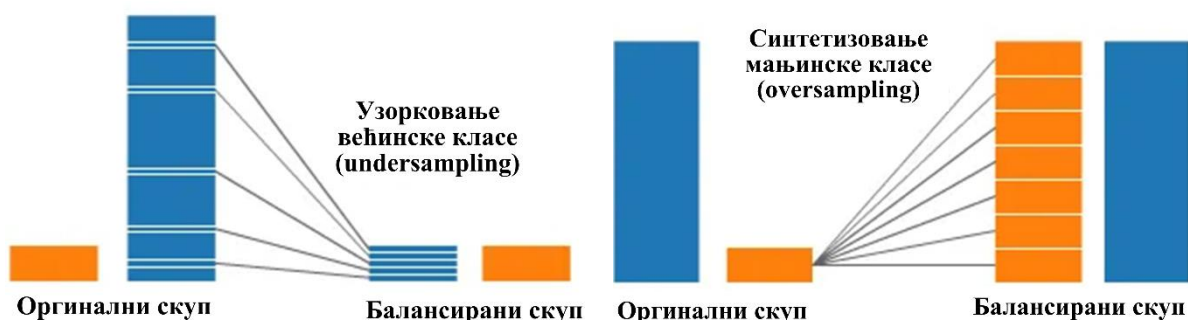


Слика 27. Број коригованих абнормалних вредности за све атрибуте скупа података

### 3.3 Технике за балансирање скупа података

У великом броју случајева не постоји равномерна заступљеност инстанци обе класе што је предуслов како би алгоритми машинског учења могли исправно да се обуче. Стога је неопходно да се приликом обуке модела користе скупови података који су уравнотежени, што подразумева да постоји једнак број инстанци за сваку од класа. У случају магнетних дискова присутна је велика неуравнотеженост мањинске класе која описује отказе диска, јер су они јако ретки догађаји.

Уравнотеженост скупова података може се побољшати техником редукованог узорковања већинске класе (енг. *downsampling*) или синтетичког повећавања броја узорака мањинске класе (енг. *upsampling*), као што је приказано на Слици 28. Код технике редукованог узорковања већинске класе, узима се само део већинских узорака тако да њихов број буде приближно једнак броју узорака мањинске класе. Ова техника се користи код магнетних дискова, када је број отказалих дискова довољно велики за процес обуке модела. Узорковање већинске класе се у том случају најчешће спроводи методом случајног узорковања. У случају да је број инстанци мањинске класе исувише мали како би се спровела обука модела неопходно је извршити вештачко генерисање нових инстанци мањинске класе које се изводи синтетичким генерисањем помоћу SMOTE алгоритма [96]. Овај алгоритам користи постојеће инстанце мањинске класе, тако да се између њих генеришу нове синтетичке инстанце мањинске класе.



Слика 28. Технике за балансирање скупа податка

У случају коришћеног CMRR скупа података присутна је добра уравнотеженост по броју инстанци исправних и отказалих дискова што је приказано у Табели 8, тако да није било потребе за коришћење техника за балансирање скупа података.

Табела 8. Заступљеност класа у CMRR скупу података

| Класа            | Број инстанци | Заступљеност класе |
|------------------|---------------|--------------------|
| Исправни дискови | 178           | 48.2%              |
| Отказали дискови | 191           | 52.8%              |
| Укупно дискова   | 369           | 100%               |

## 4 Селекција значајних атрибута

У овом поглављу представљен је поступак селекције значајних атрибута применом алгорита рекурзивне елиминације. Као резултат примене наведеног алгорита, добија се уређена листа атрибута у којој први атрибут у листи највише утиче на предикцију отказа. Селекцијом атрибута смањује се број атрибута који се користи за обуку модела чиме се смањује сложеност модела а побољшавају предиктивне перформансе модела.

### 4.1 Алгоритми за селекцију атрибута

Алгоритми машинског учења имају за циљ да пронађу зависности између вредности атрибута и очекиваних исхода предвиђања. У већини случајева постоји комплексна релација између већег броја атрибута и исхода, при чему је тешко одредити утицај појединачних атрибута на исход. Неки од атрибута ће имати значајнији утицај на исход од других, док неки ирелевантни атрибути неће уопште доприносити перформансама предвиђања, поред тога, неки атрибути могу и негативно утицати на перформансе предвиђања код појединих алгорита. Селекција атрибута представља процес утврђивања међусобне корелације између два атрибута и уклањање зависног атрибута јер он са собом не носи корисно знање већ само повећава комплексност модела. Такође, у овом кораку се уклањају и атрибути који имају константну вредност на целом скупу података. Алгоритми машинског учења као што су SVM (енг. *Support Vector Machine*) и неуронске мреже осетљиви су на ирелевантне атрибуте који у значајној мери утичу на њихове предиктивне перформансе. Поред њих, модели засновани на линеарној и логистичкој регресији су осетљиви на атрибуте који су међусобно јако корелисани, па је потребно смањити међусобну зависност атрибута како би се побољшале перформансе ових модела.

С тога је један од циљева селекције атрибута да се из скупа података уклоне сви они атрибути који не доприносе предиктивним перформансама модела. Смањењем броја атрибута превазилази се проблем прекомерног уклапања (прилагођавања) обученог модела. Ако би се модел обучио са свим расположивим атрибутима, без претходне селекције значајних атрибута, постоји могућност да се појави проблем прекомерног уклапања током фазе тестирања. Модели са оптималним скупом атрибута ће бити прецизнији а такође ће се смањити и количина прикупљених података и повећати брзина обучавања предиктивног модела. Због тога је потребно смањити број атрибута колико год је то могуће, без угрожавања перформанси предвиђања [97]. Уклоњени атрибути немају утицај на перформансе класификације и представљају тзв. шум у подацима [98].

Методологије избора атрибута могу се поделити у три опште класе:

- имплицитне методе (енг. *implicit methods*),
- методе филтрирања (енг. *filter methods*),
- методе омотача (енг. *wrapper methods*).

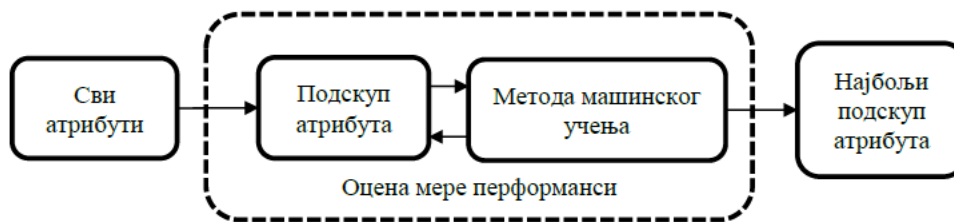
Имплицитне методе поседују природно уграђену селекцију атрибута у процесу моделовања, и не захтевају коришћење додатних екстерних модела за селекцију атрибута. Ове методе обезбеђују директну везу између селекције атрибута и циљане функције. Употреба ових метода је оправдана у случају да предикциони алгоритам није

превише захтеван за обучавање па се помоћу њих може пронаћи оптималан подскуп атрибута. Неке од често коришћених имплицитних метода за селекцију атрибута су:

1. Стабла одлучивања (енг. *decision trees*) врше процену значаја сваког атрибута у поступку гранања где се у тачки раздвајања одваја најутицајнији атрибут од осталих атрибута. Поступак се понавља како би се нашли следећи значајни атрибути. У случају да се атрибут не појави у некој од подела сматра се да је функционално независан од модела предвиђања и може се искључити из скупа података. Ако се користи алгоритам случајне шуме селекција значајних атрибута се може извршити употребом оцене важности атрибута (енг. *feature importance score*) [99].
2. Линеарни модели са регуларизацијом користе L1 и L2 регуларизацију како би умањили или елиминисали утицај мање релевантних атрибута. Lasso алгоритам је заснован на L1 регуларизацији и он додаје вредност коефицијентима како би елиминисао утицај мање значајних атрибута. Ridge алгоритам користи L2 регуларизацију да смањи ефекат небитних атрибута али их не уклања у потпуности. Elastic Net користи комбинацију L1 и L2 регуларизације како би остварио оптималну селекцију атрибута [100].
3. Неуронске мреже (енг. *neural networks*) користе уграђени механизам који у фази обуке искључује мање значајне атрибуте. Коришћењем механизма пажње (енг. *attention mechanism*) у трансформаторским моделима омогућено је фокусирање на најважније атрибуте.

Методе филтрирања проверавају сваки атрибут посебно, како би се утврдило да ли он има утицај на исход предвиђања пре него што се укључи у модел. Ове методе су једноставне и брзе али су склоне избору великог броја атрибута који најчешће неће побољшати перформансе предвиђања.

Методе омотача користе итеративни поступак претраживања где се испитује утицај подскупа атрибута како би се извршила процена перформанси обученог модела. Ове методе генеришу подскуп атрибута који има боље перформансе од оригиналног скупа, као што је приказано на Слици 29. Алгоритми секвенцијалног одабира атрибута постепено изграђују оптимални скуп атрибута, елиминисањем мање битних атрибута (енг. *backward elimination*) или додавањем значајних атрибута (енг. *forward selection*), на основу мере перформанси коришћене методе машинског учења. Најчешће коришћена метода омотача је са рекурзивним елиминисањем атрибута RFE (енг. *Recursive Feature Elimination*). Ова метода у првом кораку обучава модел користећи комплетан скуп атрибута, који се рангирају по мери и важности. У следећем кораку се уклања најмање значајни атрибут и врши се поновно рангирање атрибута. Овај поступак се наставља све док се не добије минимални број атрибута који пружају задовољавајуће перформансе предикције. Поступак рекурзивне елиминације атрибута је погодан за коришћење ако се већина атрибута у скупу података сматра релевантним, па ће ова метода брзо елиминисати неколико небитних атрибута. Насупрот овој методи, метода додавања атрибута се користи када је у скупу података мали број релевантних атрибута, док је већина атрибута ирелевантна. Метода додавања атрибута брзо проналази скуп најзначајнијих атрибута, који највише доприносе перформансама модела машинског учења.

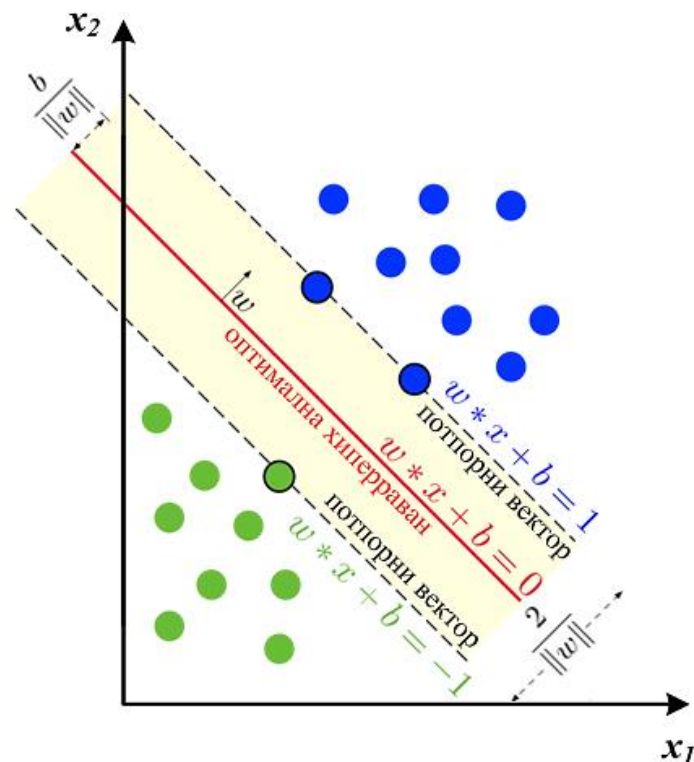


Слика 29. Поступак селекције атрибута помоћу методе омотача [98]

## 4.2 Алгоритам рекурзивне елиминације атрибута

У овој дисертацији извршена је селекција атрибута коришћењем методе рекурзивне елиминације атрибута (RFE) уз подршку класификатора заснованог на подржавајућим векторима (SVM). Овај алгоритам рангира атрибуте на основу њихове релативне важности, коришћењем SVM класификатора са линеарним кернелом (енг. *kernel*), тако што у свакој итерацији уклања атрибут са најмањим утицајем за предикцију отказа диска.

SVM представља алгоритам надгледаног учења који се најчешће користи код проблема класификације. Основна верзија алгоритма користи се за класификовање података који су линеарно раздвојиви у две класе, што подразумева да између две класе нема преклапања и да се оне могу међусобно раздвојити правом линијом. У случају вишедимензионалних података, класе је могуће раздвојити помоћу хиперравни (енг. *hyperplane*). Димензија хиперравни зависи од броја атрибута присутних у скупу података, што значи да, ако постоје два атрибута (као што је приказано на Слици 30), онда је хиперраван права линија. У случају да постоје три атрибута онда ће хиперраван бити дводимензионална раван. Генерално, ако постоји  $n$  атрибута димензија хиперравни ће бити  $n-1$ .



Слика 30. Одређивање оптималне хиперравни на основу SVM алгоритма

Циљ SVM алгоритма је да одреди најбољу хиперраван (границу одлуке) која одваја  $n$ -димензионални простор у класе, при чему је растојање између хиперравни и инстанци две класе максимално.

Једначина хиперравни може се представити обликом (10):

$$wx + b = 0 \quad (10)$$

Када се за неки скуп података одреди оптимална хиперраван, свака нова инстанца података се класификује у једну од две класе, у зависности од тога са које стране хиперравни се налази посматрана инстанца, према (11):

$$y = \begin{cases} +1 & wx + b \geq 0 \\ -1 & wx + b < 0 \end{cases} \quad (11)$$

SVM алгоритам бира инстанце обе класе које су међусобно најближе и кроз њих конструише потпорне векторе који су паралелни. На основу потпорних вектора алгоритам је и добио изворни назив алгоритам подржавајућих вектора.

Нормално растојање подржавајућих вектора се може израчунати према (12):

$$\frac{|wx + b|}{\|w\|} = \frac{1}{\|w\|} \quad (12)$$

па се укупно нормално растојање између потпорних вектора може изразити са (13):

$$\frac{2}{\|w\|} \quad (13)$$

Број потпорних вектора који се могу конструисати је бесконачан и потребно је пронаћи један пар потпорних вектора који омогућава најбоље раздвајање између класа. Оптимални пар потпорних вектора добија се налажењем максимума функције растојања између потпорних вектора  $2/\|w\|$ , под условом да између потпорних вектора нема инстанци података, дефинисано је формулом (14):

$$y_i(wx_i + b) - 1 \geq 0, \quad i = 1 \dots m \quad (14)$$

где је:

$x_i$  – улазни вектор података,

$y_i \in \{+1, -1\}$  – класа инстанце,

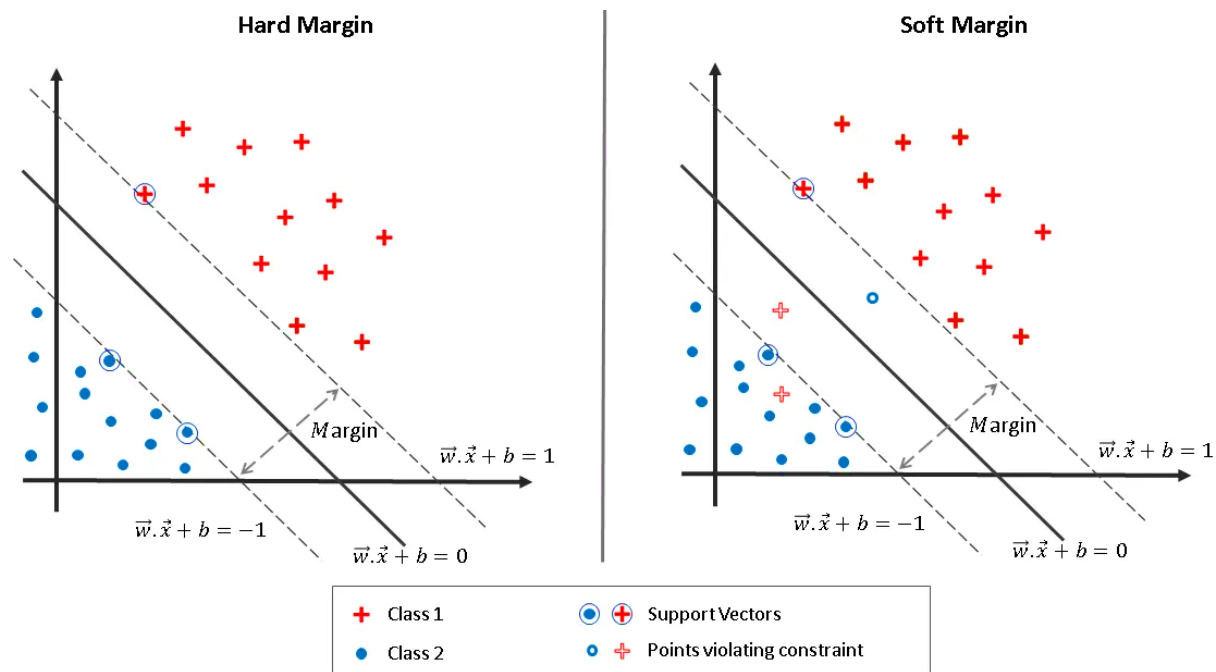
$w$  – вектор тежина који одређује правац хиперравни,

$b$  – померај хиперравни (енг. *bias*)

$m$  – представља број инстанци у скупу за обуку.

Проблем налажења максималне вредности између потпорних вектора  $2/\|w\|$  еквивалентан је проблему минимизације  $\frac{1}{2} \|w\|^2$ . Налажење оптималног решења за  $w$  добија се применом Лагранжових множитеља (енг. *Lagrange multipliers*) [101].

Тако одређен скуп потпорних вектора дефинише оптималну хиперраван која се налази на једнаком растојању између њих. Овакав начин одређивања оптималне хиперравни назива се SVM са тврдом маргином (енг. *hard-margin SVM*), помоћу које се потпуно раздвајају инстанце података које припадају различитим класама као што је приказано на Слици 31. Он захтева савршену класификацију, што значи да је применљив само ако су подаци потпуно линеарно раздвојиви. Недостатак оваквог начина одређивања маргине је велика осетљивост алгоритма на присуство абнормалних вредности (енг. *outliers*). Разлог томе што једна погрешно означена инстанца може значајно да утиче на положај хиперравни, што може довести до погрешних класификација на новим подацима.



Слика 31. SVM алгоритам са а) „тврдом” маргином и б) „меким” маргином [102]

У случају да класе инстанци нису јасно одвојиве, користи се SVM са меким маргином (енг. *soft-margin SVM*), што значи да алгоритам толерише одређени број погрешних класификација, смањујући утицај абнормалних вредности (енг. *outliers*). Алгоритам са меким маргином дозвољава да одређене инстанце података буду унутар маргине или на погрешној страни хиперравни, тако што умањује утицај таквих инстанци на избор оптималне хиперравни помоћу ненегативног релаксационог параметра  $\xi_i$ . У том случају оптимална хиперраван налази се минимизацијом према (15) уз услов (16):

$$\min_{w, \xi} \frac{\|w\|^2}{2} + C \sum_{i=1}^m \xi_i \quad (15)$$

$$y_i(wx_i + b) \geq 1 - \xi_i, \xi_i > 0, \quad i = 1 \dots m \quad (16)$$

где је:

$x_i$  – улазни вектор података,

$y_i \in \{+1, -1\}$  – класа инстанце,

$w$  – вектор тежина који одређује правац хиперравни,

$b$  – померај (енг. *bias*) хиперравни,

$m$  – представља број инстанци у скупу за обуку,

$\xi_i$  – релаксациони параметар,

$C$  – параметар регуларизације (енг. *regularization*).

Минимизовањем члана  $\frac{1}{2} \|w\|^2$  у (15), алгоритам настоји да пронађе највећу могућу маргину између класа. Члан  $C \sum_{i=1}^m \xi_i$  у (15), „кажњава“ погрешне класификације и омогућава толеранцију на грешке. Толеранција алгоритма на погрешне класификације контролише се помоћу параметра регуларизације  $C$ . Што је вредност параметра  $C$  већа алгоритам се понаша сличније алгоритму са тврдом маргином, тј. алгоритам ће покушати да минимизује грешке, чак и ако то значи да ће одабрати мању маргину. Ако је  $C$  мало алгоритам ће дозволити више грешака, али ће настојати да има ширу маргину.

1. Ако је  $\xi_i = 0$  инстанца  $x_i$  је исправно класификована и налази се ван маргине.
2. Ако је  $0 < \xi_i < 1$  инстанца  $x_i$  је унутар маргине али је и даље тачно класификована.
3. Ако је  $\xi_i \geq 1$ , инстанца  $x_i$  је погрешно класификована (тј. налази се на погрешној страни хиперравни).

Претходно разматрани алгоритми погодни су само код линеарно раздвојивих скупова података, што није чест случај код већине скупова података. Коришћењем нелинеарних функција кернела, могуће је пресликати нелинеарно раздвојиве класе података у простор веће димензије у којем они постају линеарно раздвојиви што омогућава примену регуларног SVM алгоритма. Најчешће коришћене нелинеарне функције кернела су: полиномски кернели (енг. *polynomial kernel*), Гаусов кернел (енг. *Gaussian kernel*) и кернел сигмоидне функције (енг. *sigmoid kernel*).

Процес селекције атрибута може се посматрати као проблем бинарне класификације чији је улаз вектор са  $n$  атрибута у  $n$ -димензионалном простору  $F$ . Скуп за обуку састоји се од низа инстанци  $\{x_1, x_2, \dots, x_k, \dots, x_m\}$  са познатим ознакама класе  $\{y_1, y_2, \dots, y_k, \dots, y_m\}$ , где је  $y_k \in \{-1, +1\}$ , како би се креирала функција одлуке  $D(x)$ . Функције одлуке које представљају тежинску суму инстанци обуке и коефицијента пристрасности, називају се линеарно одвојиве функције и приказане су у облику (17):

$$D(x) = w \cdot x + b \quad (17)$$

где је,

$w$  - тежински вектор,

$b$  - је вредност коефицијента пристрасности.

Нове инстанце  $x_i$  ( $i=1, 2, \dots, m$ ) се класификују у зависности од знака функције одлуке према (18):

$$\begin{aligned} D(x_i) > 0 &\Rightarrow x_i \in \text{класа} + 1 \\ D(x_i) < 0 &\Rightarrow x_i \in \text{класа} - 1 \\ D(x_i) = 0 &, \text{граница одлуке} \end{aligned} \quad (18)$$

Скуп података се може сматрати да је линеарно одвојив ако је могуће пронаћи линеарно одвојиву функцију која је у стању да класификује скуп података без грешака. Према наведеном облику функције одлуке, може се уочити да ће атрибути помножени са највећим апсолутним вредностима коефицијента вектора тежине  $w$ , имати највећи утицај на одлуку о класификацији. Сходно томе, највећи коефицијенти вектора тежине указују на атрибуте који највише доприносе класификацији. За класификацију линеарно одвојивих података најчешће се користи линеарни SVM модел, који одређује оптималну хиперраван за раздвајање класа.

Алгоритам рекурзивне елиминације атрибута RFE (енг. *Recursive Feature Elimination*) примењује критеријум рангирања над тежинама вектора добијених обуком SVM модела. Овај алгоритам полази од претпоставке да атрибут са најмањом придруженом вредношћу вектора тежине  $w$ , има најмањи утицај на функцију одлуке, па се исти може уклонити без значајнијег смањења нивоа предиктивних перформанси. У следећем кораку, алгоритам не наставља са уклањањем следећег атрибута са најмањом придруженом тежином, већ започиње поновну обуку SVM модела са редукованим скупом атрибута како би одредио нову функцију одлуке. Следеће итерације ће уклањати по један атрибут са најмањом придруженом тежином и поново обучавати SVM модел све док у скупу атрибута не остане само један најутицајнији атрибут. На тај начин у

свакој итерацији биће извршено рангирање при чему ће атрибути који су уклоњени у првим итерацијама бити ранжирани као мање утицајни у односу на оне уклоњене у каснијим итерацијама алгоритма [103]. Алгоритамска шема SVM RFE методе приказана је на Слици 32.



Слика 32. Алгоритам рекурзивне селекције атрибута

CMRR скуп података садржи 59 атрибута, као што је наведено у 3.1, прикупљених током рада магнетних дискова и ознаке класе које означавају да ли је диск отказао или не. Анализом CMRR скупа података коришћеног у дисертацији утврђено је да су вредности одређених атрибута константе за све инстанце података, тако да су они уклоњени јер немају никакав утицај на отказ магнетног диска. У Табели 9 приказани су уклоњени атрибути из скупа података који имају константне вредности.

Табела 9. Листа уклоњених атрибута са константним вредностима

| Редни број | Назив атрибута | Константна вредност |
|------------|----------------|---------------------|
| 1.         | Temp2          | 0                   |
| 2.         | Servo4         | 0                   |
| 3.         | ReadError13    | 0                   |
| 4.         | ReadError14    | 0                   |
| 5.         | ReadError15    | 0                   |
| 6.         | ReadError16    | 0                   |
| 7.         | FlyHeight14    | 1                   |
| 8.         | FlyHeight15    | 1                   |
| 9.         | FlyHeight16    | 1                   |
| 10.        | Temp5          | 1                   |
| 11.        | Temp6          | 57005               |
| 12.        | ReadError20    | 0                   |
| 13.        | WriteError     | 57005               |

Уклањањем атрибута са константном вредношћу, од почетних 59 атрибута скупа података, преостало је 46 који имају различит утицај на отказ диска.

Корелација представља меру међусобне повезаности два атрибута и она се може проширити на више атрибута помоћу корелационе матрице. Вредност коефицијента корелације дефинисана је опсегом од -1 до 1, при чему је један од најчешће коришћених метода за израчунавање корелације Пирсонов коефицијент (енг. *Pearson coefficient*) линеарне корелације [104] чија је математичка дефиниција дата у формули (19):

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (19)$$

где је:

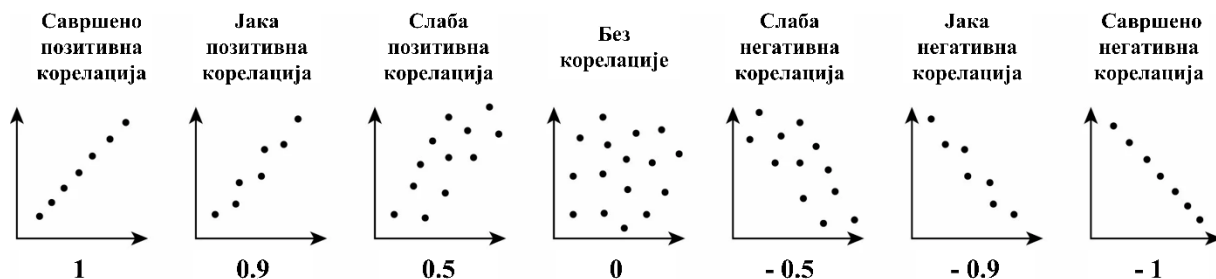
$n$  - број инстанци,

$x_i$  и  $y_i$  су вредности атрибута инстанце  $i$ ,

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  је средња вредност за атрибут  $x$ ,

$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  је средња вредност за атрибут  $y$ .

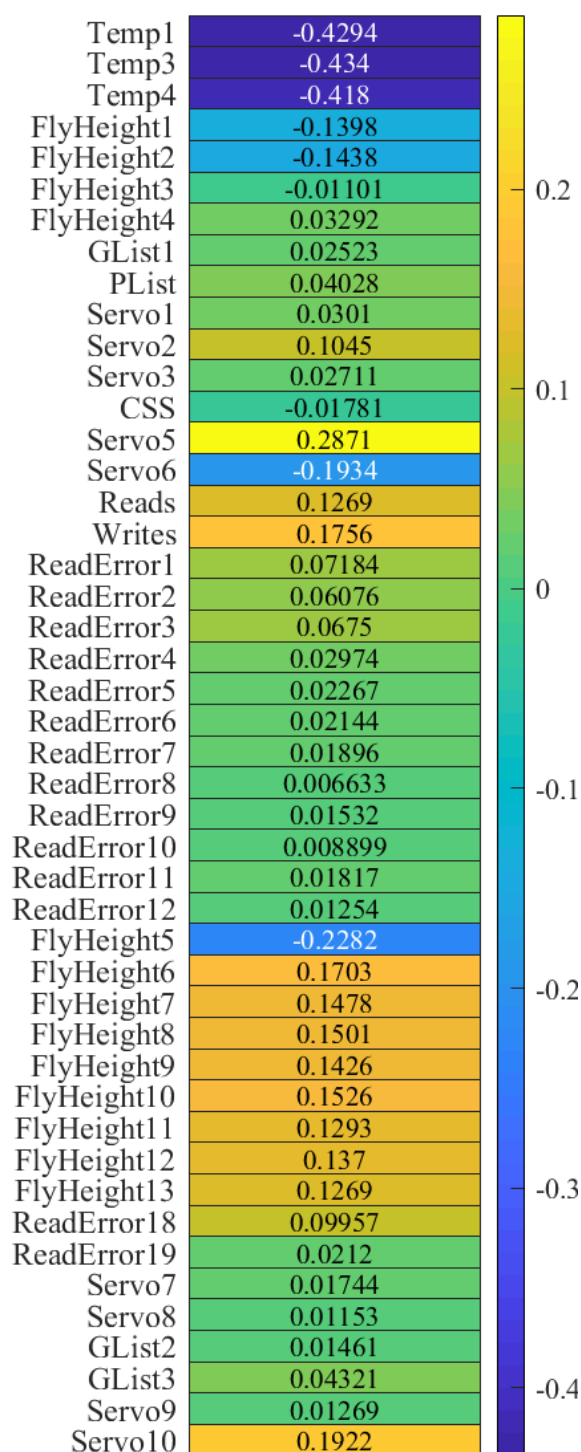
Зависност коефицијента корелације за различите односе између две променљиве приказана је на *Слици 33*.



Слика 33. Коефицијент корелације за различите односе између две променљиве

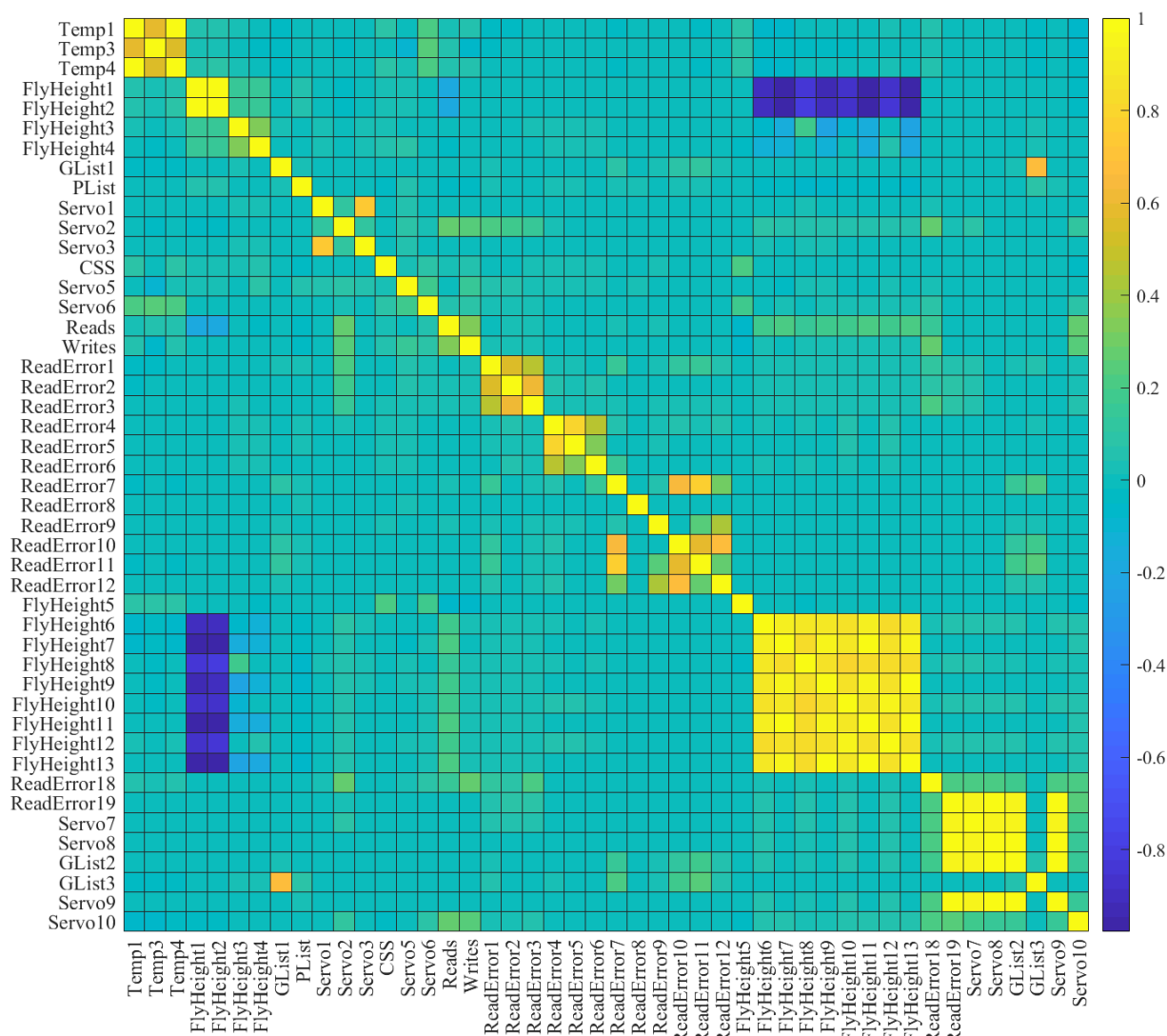
Апсолутна вредност коефицијента корелације показује јачину линеарне повезаности између две променљиве, што је она ближа јединици то је њихова линеарна повезаност већа. Ако је вредност ближа нули то указује на слабу или непостојећу линеарну повезаност између променљивих. Позитивна вредност корелације значи да се повећањем вредности једне променљиве повећава и вредност друге променљиве. У супротном код негативне корелације, повећањем вредности једне променљиве долази до смањења вредности друге променљиве.

На *Слици 34*, приказане су вредности корелације које су одређене између појединачних атрибута CMRR скупа података и индикатора отказа магнетног диска. Вредности корелације одређене су коришћењем уграђене функције *corr* програмског пакета *Matlab*. Атрибути са великом апсолутном вредношћу корелације показују јачу линеарну повезаност са појавом отказа у односу на оне код којих је вредност корелације блиска нули. Негативне вредности коефицијента корелације указују да до отказа долази смањењем вредности атрибута. Насупрот томе, код позитивних вредности коефицијента корелације до отказа долази повећањем вредности атрибута.



Слика 34. Корелација атрибута и отказа магнетног диска

Вредности међусобне корелације између 46 атрибута CMRR скупа приказане су помоћу топлотне мапе (енг. *heatmap*) на Слици 35. Топлотна мапа је генерисана коришћењем уграђене функције *heatmap* програмског пакета *Matlab*. Може се закључити да је присутна јака позитивна корелација (приказана на слици топлим бојама) између атрибута температуре. Висока вредност корелације постоји између атрибута који мере висину уписно-читајуће главе у односу на површину диска, где доминира јака позитивна корелација између ових атрибута. Изузетак су атрибути *FlyHeight1* и *FlyHeight2* који су у јакој негативној корелацији (приказана на слици хладним бојама) са осталим атрибутима који мере висину уписно-читајуће главе.



Слика 35. Међусобна корелација атрибута CMRR скупа података

На топлотној мапи се запажа и јака позитивна корелације између атрибута серво позиционог механизма. Сви атрибути који имају висок степен међусобне корелације неће значајно допринети предиктивним перформансама модела, па ће у поступку селекције атрибута бити замењени једним од најутицајних атрибута из те групе.

Атрибути CMRR скупа података имају различит утицај на предикцију отказа магнетног диска па их је потребно рангирати коришћењем алгоритма за рекурзивну елиминацију атрибута заснованог на линеарном SVM класификатору. На тај начин може се одабрати оптималан подскуп атрибута CMRR скупа података који има највећи утицај на предикцију отказа, чиме се смањује рачунска комплексност модела, јер се елиминишу атрибути који имају мали утицај на предикцију отказа диска.

За обуку предиктивног модела коришћен је линеарни SVM алгоритам са меком границом из разлога што инстанце података нису потпуно линеарно раздвојиве. Линеарни SVM класификатор обучен је коришћењем 10.000 насумично одабраних инстанци података, јер би обука над целокупним CMRR скупом података захтевала 35 пута више рачунарског времена. Подскуп инстанци за обуку прослеђује се *svm-rfe* функцији имплементираној у програмском пакету R. Ова функција итеративно рангира атрибуте на основу вектора тежине добијених обуком SVM класификатора. У сваком кораку атрибут са најмањом придруженом тежином вектора уклања се из листе атрибута.

Процес се понавља све док у листи не остане само један атрибут са највећим утицајем на предикцију отказа. Као резултат позива *svm-rfe* функције добијена је листа ранжираних атрибута приказана у Табели 10. Најзначајнији атрибути који доприносе отказу диска су *GList1* и *GList3*, који представљају листу лоших сектора насталих током рада диска. Остали високо ранжирани атрибути везани су за серво позициони механизам и атрибуте које мере висину уписно-читајуће главе.

Табела 10. Листа ранжираних атрибута SVM-RFE [120]

| Ранг атрибута | Назив атрибута | Ранг атрибута | Назив атрибута |
|---------------|----------------|---------------|----------------|
| 1.            | GList1         | 24.           | FlyHeight5     |
| 2.            | GList3         | 25.           | CSS            |
| 3.            | Servo10        | 26.           | PList          |
| 4.            | ReadError18    | 27.           | Servo2         |
| 5.            | Servo9         | 28.           | Writes         |
| 6.            | FlyHeight1     | 29.           | ReadError3     |
| 7.            | FlyHeight2     | 30.           | ReadError2     |
| 8.            | Temp4          | 31.           | ReadError5     |
| 9.            | FlyHeight11    | 32.           | ReadError9     |
| 10.           | Servo5         | 33.           | FlyHeight7     |
| 11.           | FlyHeight10    | 34.           | FlyHeight4     |
| 12.           | ReadError19    | 35.           | Servo7         |
| 13.           | Servo6         | 36.           | Temp3          |
| 14.           | FlyHeight9     | 37.           | Servo3         |
| 15.           | FlyHeight6     | 38.           | ReadError1     |
| 16.           | ReadError4     | 39.           | ReadError7     |
| 17.           | Servo1         | 40.           | ReadError6     |
| 18.           | FlyHeight12    | 41.           | ReadError11    |
| 19.           | FlyHeight8     | 42.           | GList2         |
| 20.           | Temp1          | 43.           | ReadError12    |
| 21.           | FlyHeight13    | 44.           | Servo8         |
| 22.           | Reads          | 45.           | ReadError10    |
| 23.           | FlyHeight3     | 46.           | ReadError8     |

## 5 Трансформација непараметарски дистрибуираних података и мерење удаљености

Ово поглавље описује поступак трансформације непараметарски дистрибуираних скупова података у нормално дистрибуиран облик применом *Box-Cox* трансформације. Након трансформације скупа података у нормализован облик могућа је примена Махаланобисове удаљености (енг. *Mahalanobis distance*) за мерење растојања у мултиваријабилном простору између инстанци података.

### 5.1 Дистрибуција података

Већина модела машинског учења преферира параметарски дистрибуиране скупове података како би се тачно предвидела опажања код нових инстанци података. Сваки од SMART параметара представља скуп нумеричких вредности које су прикупљене током рада диска. Учестаност појаве појединих вредности параметара може се представити помоћу хистограма који приказује дистрибуцију, односно расподелу вредности једне променљиве. Хистограм класификује вредности података у групе опсега вредности (енг. *bins*) и приказује број вредности променљиве који припада свакој групи опсега. На основу изгледа хистограма одређена појава може се представити неком од математичких функција густине расподеле вероватноће. У зависности од облика функције густине расподеле вероватноће постоје различити типови статистичких дистрибуција, као што су: униформна (енг. *uniform*), нормална (енг. *normal*), Поасонова (енг. *Poisson*), Бернулијева (енг. *Bernoulli*), итд. Наведене расподеле описане су математичким функцијама где облик расподеле зависи од параметара функције па се оне често називају параметарске расподеле.

Униформна расподела дефинисана је функцијом густине вероватноће која на интервалу од  $a$  до  $b$  има константну густину вероватноће која износи  $1/(b-a)$ , што значи да ће се опажања на овом интервалу појављивати са једнаком вероватноћом. Генерално, већина појава не прати униформну расподелу већ имају тенденцију да се групишу око неке централне вредности са мањим или већим одступањима. Овакав облик расподеле познат је као нормална или Гаусова (енг. *Gaussian*) расподела, приказана на Слици 36, која се описује помоћу два параметра, средње вредности  $\mu$  и стандардног одступања  $\sigma$ . Густина расподеле вероватноће дефинисана је функцијом (20).

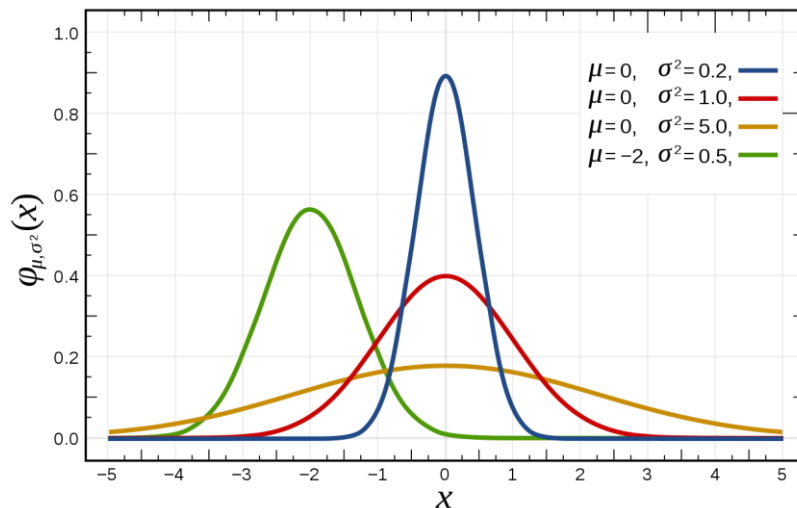
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (20)$$

где је:

$\mu$  – параметар средње вредности дистрибуције;

$\sigma$  – параметар стандардног одступања;

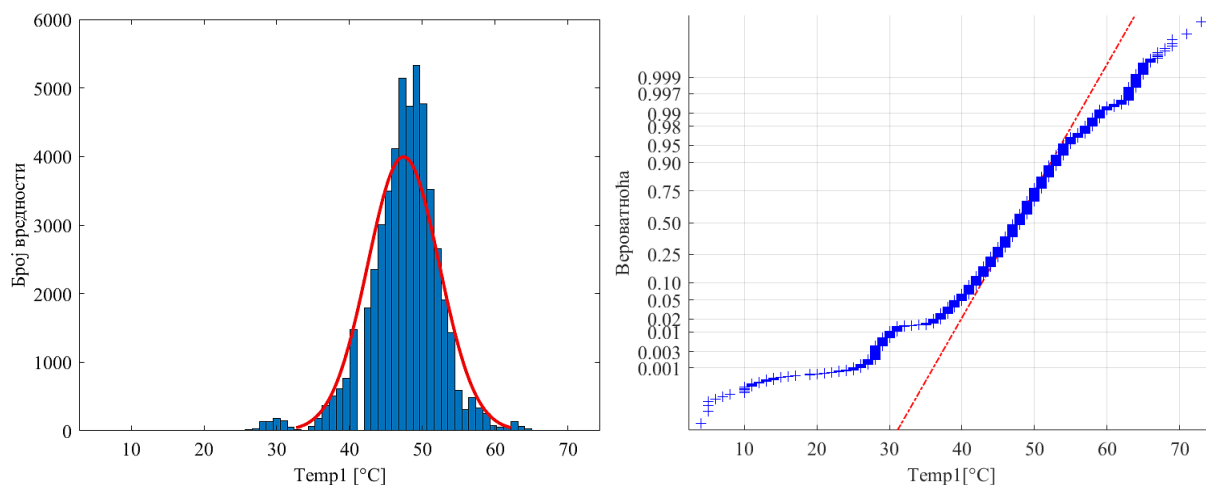
$x$  – вредност случајне променљиве.



Слика 36. Различити облици нормалне дистрибуције

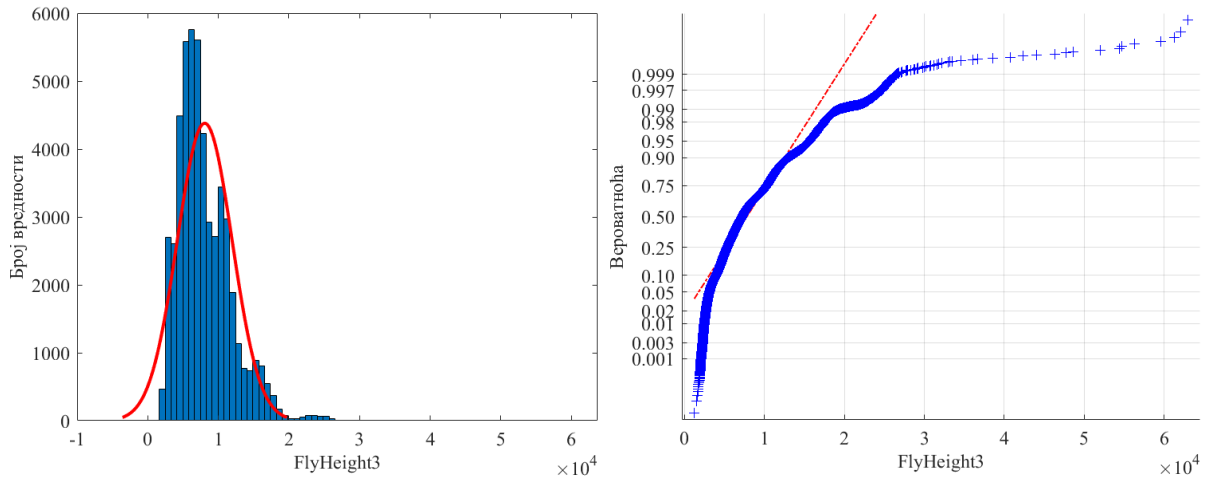
Стандардна нормална расподела дефинисана је са средњом вредношћу  $\mu=0$  и стандардним одступањем  $\sigma=1$ . Код стандардне нормалне расподеле 68.2% опажања појавиће се унутар интервала  $\pm 1\sigma$  око средње вредности, 95.4% опажања појавиће се унутар интервала  $\pm 2\sigma$ , а 99.7% опажања појавиће се унутар интервала  $\pm 3\sigma$ . Опажања која се појаве изван интервала од  $\pm 3\sigma$  обухватају свега 0.3% свих опажања. Променом средње вредности расподеле крива функције се помера у лево или десно при чему не долази до промене облика криве. Променом параметра стандардног одступања долази до промене облика криве, што је параметар  $\sigma$  већи од 1 крива је шири у односу на криву стандардне нормалне расподеле. Без обзира на параметре нормалне расподеле  $\mu$ ,  $\sigma$  вероватноћа опажања по интервалима  $\pm 1\sigma$ ,  $\pm 2\sigma$  и  $\pm 3\sigma$  око средње вредности се не мења.

Већина атрибута у CMRR скупу података не прати нормалну расподелу података. Изузетак је дистрибуција вредности температуре *Temp1* где је на Слици 37а приказана расподела вредности температуре за исправне дискове у односу на нормалну дистрибуцију. Расподела вредности атрибута генерисана је коришћењем уграђене функције *hist*, док је нормална расподела за вредности атрибута генерисана функцијом *normfit* програмског пакета *Matlab*. На Слици 37б приказан је дијаграм вероватноће генерисан *Matlab* функцијом *normplot*, који упоређује дистрибуцију вредности температуре *Temp1* (приказане плавим маркерима) са нормалном дистрибуцијом (приказаном црвеном линијом).



Слика 37. Дистрибуција вредности температуре *Temp1* а) за исправне дискове и б) поређење са нормалном дистрибуцијом

Такође на *Слици 38а* приказана је дистрибуција вредности атрибута *FlyHeight3* висине уписно-читајуће главе за исправне дискове. На дијаграму вероватноће приказаном на *Слици 38б* запажа се да дистрибуција висине уписно-читајуће главе *FlyHeight3* у значајнијој мери одступа од нормалне расподеле. Вредности атрибута отказалих дискова значајно одступају од вредности атрибута исправних дискова па се из тог разлога не користе за креирање дистрибуције вредности атрибута.



Слика 38. Дистрибуција вредности висине уписно читајуће главе *FlyHeight3* а) за исправне дискове и б) поређење са нормалном дистрибуцијом

## 5.2 Трансформација података

Дистрибуције вредности SMART параметара се називају непараметарске расподеле јер у већини случајева не прате познате дистрибуције као што су нормална, експоненцијална, униформна, итд. Из тог разлога потребно је извршити трансформацију у нормално дистрибуиран облик пре примене метода машинског учења. Неке од најчешће коришћених трансформација су *Box-Cox* трансформација [105] и *Yeo-Johnson* трансформација [106].

*Box-Cox* трансформација трансформише позитивне непараметарски дистрибуиране податке, њиховим степеновањем за експонент степена  $\lambda$ , како би трансформисани подаци били нормално распоређени. Овом методом се проналази оптимална вредност експонента степена  $\lambda$ , која минимизира стандардну девијацију трансформисаних података, при чему трансформисани подаци имају највећу вероватноћу да буду нормално распоређени. Трансформација сваког атрибута, који мора да има позитивне вредности, изводи се према једначини (21).

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases} \quad (21)$$

У случају да су присутне негативне вредности података, њима се додаје константна вредност  $\lambda_2$  како би се сви подаци претворили у позитивне вредности пре примене *Box-Cox* трансформације за одређивање експонента степена  $\lambda_1$  према (22)

$$y(\lambda) = \begin{cases} \frac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \lambda_1 \neq 0 & y > -\lambda_2 \\ \ln(y + \lambda_2) & \lambda_1 = 0 & y > -\lambda_2 \end{cases} \quad (22)$$

## 5. Трансформација непараметарски дистрибуираних података и мерење удаљености

где је:

$\lambda_1$  – параметар (експонент) који одређује врсту трансформације;

$\lambda_2$  – померај који омогућава трансформацију негативних вредности у позитивне;

$y$  – посматране вредности података;

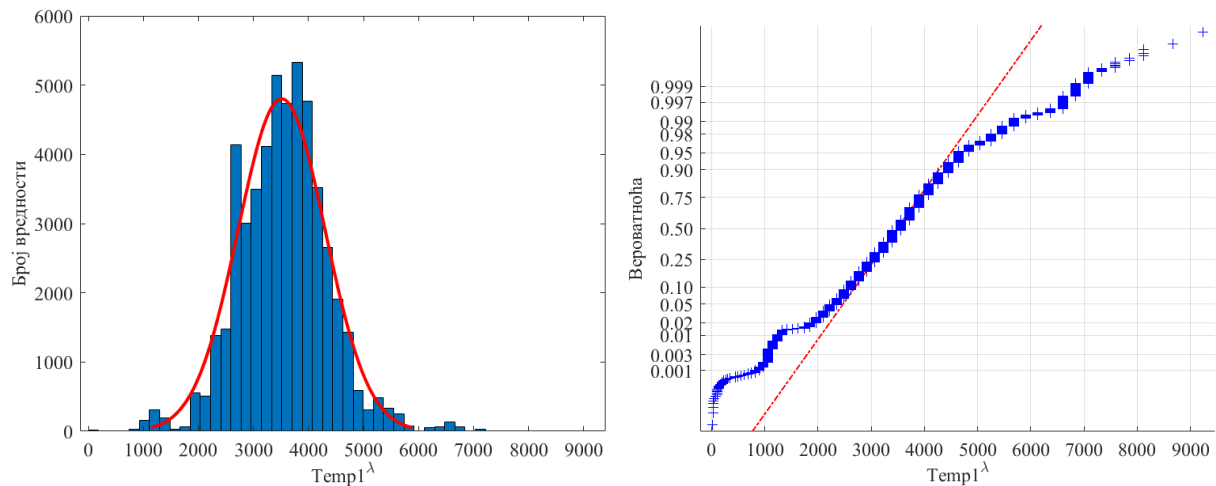
услов  $y > -\lambda_2$  осигурава да аргумент логаритамске функције буде позитиван.

Трансформација CMRR скупа података изводи се тако што се за скуп свих инстанци исправних дискова одређује оптимална вредност параметра  $\lambda$  за сваки атрибут, тако да се вредности тог атрибута што боље приближе нормалној расподели. С обзиром на то да вредности атрибута могу бити једнаке нули, пре примене *Box-Cox* трансформације неопходно је увећати све вредности атрибута за један, како би се обезбедило да су све вредности позитивне. Коришћењем *Matlab* функције *boxcox* одређене су вредности експонента  $\lambda$  приказане у Табели 11. којима се врши трансформација целокупног скупа података укључујући и отказале дискове.

Табела 11. Вредности  $\lambda$  за *Box-Cox* трансформацију атрибута CMRR скупа

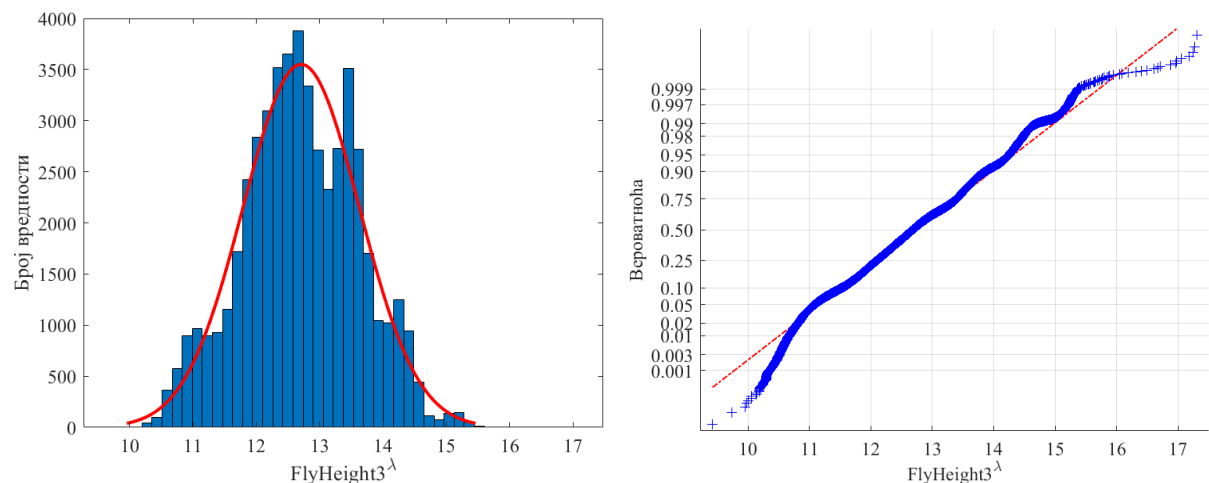
| Р. бр. | Атрибут    | $\lambda$ | Р. бр. | Атрибут     | $\lambda$ |
|--------|------------|-----------|--------|-------------|-----------|
| 1.     | Temp1      | 2.317     | 24.    | ReadError7  | 0         |
| 2.     | Temp3      | 2.370     | 25.    | ReadError8  | 0         |
| 3.     | Temp4      | 2.252     | 26.    | ReadError9  | 0         |
| 4.     | FlyHeight1 | 0.008     | 27.    | ReadError10 | 0         |
| 5.     | FlyHeight2 | 2.608     | 28.    | ReadError11 | 0         |
| 6.     | FlyHeight3 | 0.076     | 29.    | ReadError12 | 0         |
| 7.     | FlyHeight4 | 0.161     | 30.    | FlyHeight5  | 0.709     |
| 8.     | GList1     | 1.172     | 31.    | FlyHeight6  | 0.259     |
| 9.     | PList      | 1.159     | 32.    | FlyHeight7  | 0.218     |
| 10.    | Servo1     | 1.940     | 33.    | FlyHeight8  | 0.071     |
| 11.    | Servo2     | 1.281     | 34.    | FlyHeight9  | 1.108     |
| 12.    | Servo3     | 0.968     | 35.    | FlyHeight10 | 0.142     |
| 13.    | CSS        | 1.155     | 36.    | FlyHeight11 | 0.377     |
| 14.    | Servo5     | 1.230     | 37.    | FlyHeight12 | 0.119     |
| 15.    | Servo6     | 0.016     | 38.    | FlyHeight13 | 0.829     |
| 16.    | Reads      | 1.167     | 39.    | ReadError18 | 1.346     |
| 17.    | Writes     | 0.062     | 40.    | ReadError19 | 1.559     |
| 18.    | ReadError1 | 0.737     | 41.    | Servo7      | 3.108     |
| 19.    | ReadError2 | 0.968     | 42.    | Servo8      | 0         |
| 20.    | ReadError3 | 0.731     | 43.    | GList2      | 0         |
| 21.    | ReadError4 | 0         | 44.    | GList3      | 1.264     |
| 22.    | ReadError5 | 0         | 45.    | Servo9      | 1.047     |
| 23.    | ReadError6 | 0         | 46.    | Servo10     | 0.316     |

Применом вредности експонента  $\lambda=2.317$  одређеног за атрибут температуре *Temp1* врши се трансформација вредности атрибута степеном функцијом, чија је дистрибуција приказана на *Слици 39а*. Поређењем дистрибуције трансформисаних вредности атрибута температуре *Temp1* са нормалном расподелом приказаном на *Слици 39б* уочава се да трансформисани подаци боље прате нормалну расподелу у односу на нормалну расподелу оригиналних вредности приказану на *Слици 37б*.



Слика 39. Дистрибуција трансформисаних вредности температуре *Temp1* а) за исправне дискове и б) поређење са нормалном дистрибуцијом

Трансформацијом вредности атрибута висине уписно-читајуће главе *FlyHeight3*, које не прате нормалну расподелу, примењује се степена трансформација са вредношћу експонента  $\lambda=0.076$ . Добијена дистрибуција трансформисаних вредности атрибута знатно је ближа нормалној расподели као што је приказано на *Слици 40*.



Слика 40. Дистрибуција трансформисане вредности висине уписно читајуће главе *FlyHeight3* а) за исправне дискове и б) поређење са нормалном дистрибуцијом

### 5.3 Мерење растојања у вишедимензионалном простору

Инстанце скупа података представљају се као скуп тачака у вишедимензионалном простору где је позиција тачке дефинисана вредностима њених атрибута по осама вишедимензионалног простора. У случају бинарне класификације ове тачке су означене одговарајућим позитивним и негативним ознакама класе. Тачке доминантне класе имају тенденцију да се групишу око тачке која је дефинисана као средња вредност, односно центар масе. Тачке мањинске класе обично су расуте и налазе се на већим растојањима

од центра масе доминантне класе. Како би се тачке мањинске класе исправно класификовале неопходно је извршити мерење њиховог растојања у вишедимензионалном простору у односу на центар масе. Растојање се може мерити различитим метрикама као што су Еуклидска (енг. *Euclidean*), Менхетен (енг. *Manhattan*), Чебишевљева (енг. *Chebyshev*), Минковски (енг. *Minkowski*), косинусна (енг. *Cosine*) или Махаланобисова удаљеност (енг. *Mahalanobis distance*) [107].

Растојање у вишедимензионалном простору, при чему све димензије доприносе подједнако, мери се коришћењем Еуклидског растојања (енг. *Euclidean*). Ова метрика, познато као  $L_2$  норма, представља квадратни корен од суме квадрата растојања између почетне и крајње тачке за сваку димензију понаособ. Еуклидско растојање представља најкраће могуће растојање између две тачке у простору, према (23):

$$d_{L_2}(x_i, x_j) = \left( \sum_{k=1}^m (x_{ik} - x_{jk})^2 \right)^{1/2} \quad (23)$$

Поред Еуклидског растојања користи се и правоугаона (Менхетен) удаљеност позната као метрика  $L_1$ , која представља збир апсолутних разлика координата појединачно по свакој димензије (24). Ова метрика назив је добила по острву Менхетен, чија улична мрежа има правоугаони распоред, па је кретање ограничено на правце паралелне осама координатног система. Области примене ове метрике су алгоритми за проналажење путања, кластеризацију података и детекцију абнормалних вредности.

$$d_{L_1}(x_i, x_j) = \sum_{k=1}^m |x_{ik} - x_{jk}| \quad (24)$$

Чебишевљева удаљеност, позната као максимална метрика или  $L_\infty$  метрика, представља највећу апсолутну разлику у било којој координати између две тачке (25). Ова метрика се често примењује за оптимизацију путање робота, као и у алгоритмима за обраду слике.

$$d_{L_\infty}(x_i, x_j) = \max_k |x_{ik} - x_{jk}| \quad (25)$$

Минковски удаљеност, позната као  $L_p$  метрика, представља општу форму која обухвата више различитих удаљености. Ова метрика користи параметар  $p$  који одређује врсту удаљености (26). Када је вредност параметра  $p=1$ , добија се Менхетен удаљеност, у случају када је  $p=2$  тада је Еуклидско растојање, а када је  $p=\infty$  удаљеност се дефинише као Чебишевљева удаљеност.

$$d_{L_p}(x_i, x_j) = \left( \sum_{k=1}^m (x_{ik} - x_{jk})^p \right)^{1/p} \quad (26)$$

Косинусна удаљеност мери разлику између вектора који дефинишу две тачке у вишедимензионалном простору и израчунава се према (27):

$$d_{\cosine}(x_i, x_j) = 1 - \cos \theta \quad (27)$$

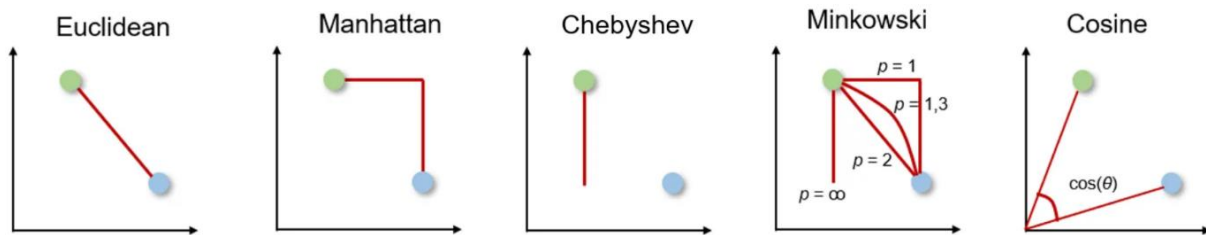
где је:

$\theta$  – угао између вектора  $x_i$  и  $x_j$ .

Вредност косинусне удаљености се креће у границама од 0 до 2. Уколико је вредност косинусне удаљености блиска нули, вектори су усмерени у истом правцу и тачке се сматрају блиским. Тачке се могу сматрати међусобно супротним када је вредност

косинусног растојања 2, док вредност 1 означава да су вектори ортогонални, односно да између њих нема сличности. Овај тип растојања најчешће се користи за анализу текстуалних података и за утврђивање семантичких релација између речи код алгоритама за обраду природних језика.

Наведене метрике су засноване на претпоставци да су вредности по димензијама међусобно независне и имају једнаку важност, као што је приказано на *Слици 41*. Међутим, оне имају одређене недостатке када се примењују на мултиваријантне податке код којих постоји различито скалирање вредности и међусобна корелација димензија. Менхетен метрика ублажава проблем скалирања јер додељује исту вредност свим димензијама, али не решава проблем корелације између њих.



Слика 41. Метрике за мерење растојања у вишедимензионалном простору [108]

Махаланобисово растојање представља меру удаљености у вишедимензионалном простору изражену у стандардним одступањима у односу на тачку која представља центар масе дистрибуције података. За разлику од стандардног Еуклидског растојања, ово растојање је мера без јединице, која узима у обзир како скалирање параметара, тако њихову међусобну корелацију.

Примена метрике засноване на Махаланобисовом растојању корисна је када су испуњени следећи услови [109]:

- присутна је корелација између атрибута,
- атрибути имају различите опсеге вредности, односно различите варијансе,
- вредности атрибута приближно прате нормалну расподелу.

Махаланобисово растојање  $MD_i$  за сваку тачку података  $i$ , израчунава се коришћењем нормализованих вредности тачака података  $z_i = [z_{i1}, z_{i2}, \dots, z_{in}]$  и матрице коваријансе  $C$  према формули (28):

$$MD_i = D_i^2 = \frac{1}{n} z_i^T C^{-1} z_i \quad (28)$$

где је:

$D_i^2$  – квадрат Махаланобисовог растојања за дату тачку  $i$ ;

$z_i$  – нормализоване вредности тачке података  $i$ ;

$z_i^T$  – транспоновани вектор  $z_i$ ;

$C^{-1}$  – инверзна матрица коваријансе.

Нормализоване вредности (енг. *z-score*) атрибута  $z_{ij}$  трансформишу оригиналне вредности атрибута  $x_{ij}$  за све тачке података чиме се елиминише ефекат скалирања између атрибута. Атрибути у нормализованом облику описани су помоћу средње вредности  $\bar{X}$  и стандардног одступања  $S_j$  дистрибуције података према (29):

$$z_{ij} = \frac{x_{ij} - \bar{X}_j}{S_j} \quad (29)$$

Средња вредност односно положај центра масе дистрибуције података  $\bar{X}$  одређен је скупом координата  $\{\bar{X}_1, \dots, \bar{X}_j, \dots, \bar{X}_n\}$  који се добија као аритметичка средина тачака података за сваки атрибут  $j = 1 \dots n$ , према (30):

$$\bar{X}_j = \frac{1}{m} \sum_{i=1}^m x_{ij} \quad (30)$$

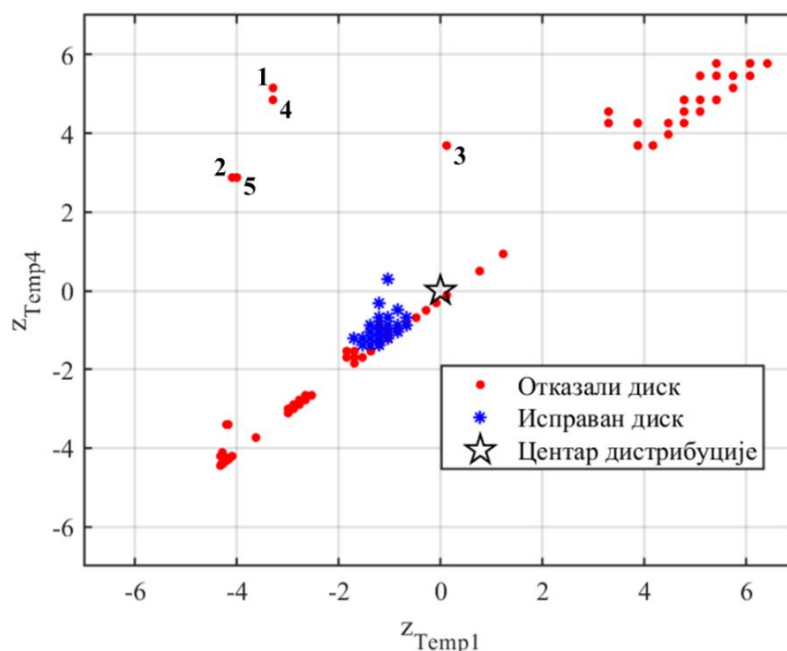
Расејање тачака података око центра масе израчунава се стандардним одступањем  $S_j$  за сваки атрибут  $j$ , према (31):

$$S_j = \sqrt{\frac{\sum_{i=1}^m (x_{ij} - \bar{X}_j)^2}{m - 1}} \quad (31)$$

Матрица коваријансе (енг. *covariance matrix*) представља меру међусобне линеарне зависности променљивих. Свака ћелија у матрици приказује коваријансу између две променљиве, при чему су елементи на главној дијагонали варијансе појединачних променљивих. Коваријанса две променљиве представља просечну вредност суме производа одступања појединачних променљивих од својих средњих вредности (32):

$$C = \frac{1}{m - 1} \sum_{i=1}^m z_i^T z_i \quad (32)$$

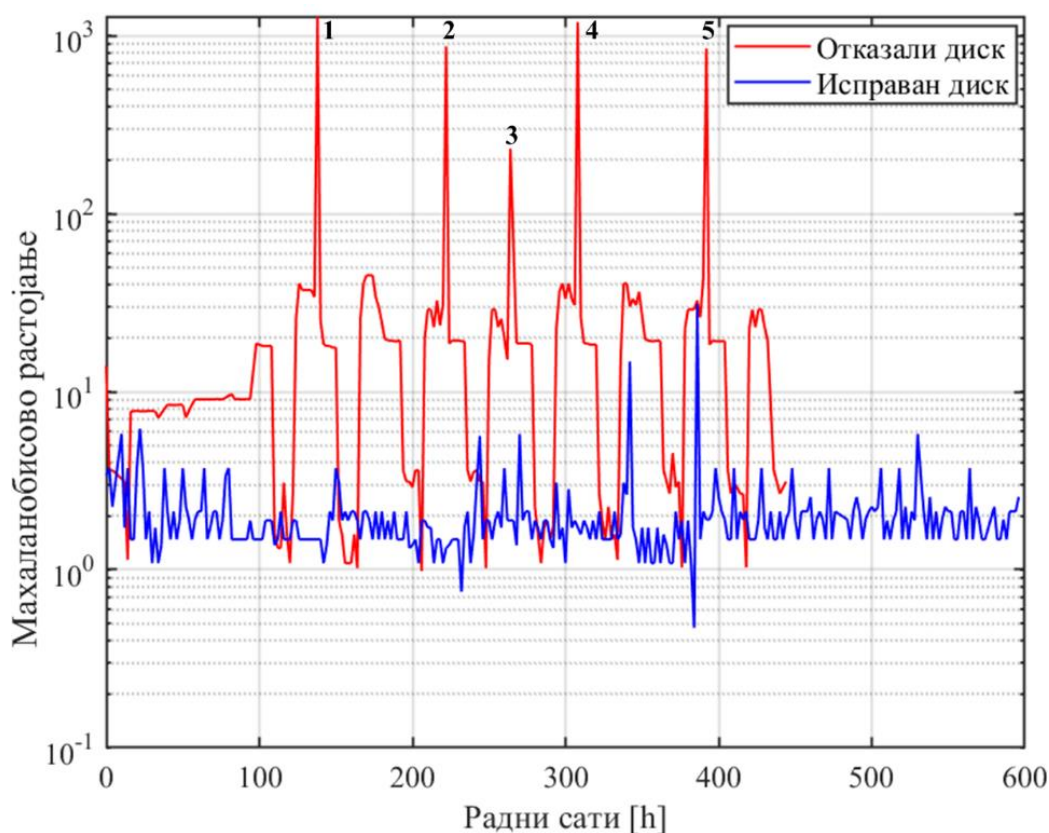
Најчешће се примењује код сложених података јер омогућава анализу односа између свих променљивих, будући да сваки елемент матрице представља коваријансу одговарајућих парова променљивих. Матрица коваријансе је у пракси најчешће скоро сингуларна услед високе корелације између атрибута или када број узорака није довољно велики у односу на број променљивих, што онемогућава проналажење њене инверзне матрице. Овај проблем се решава налажењем *Moore-Penrose* (М-Р) псеудо инверзне матрице, која омогућава добијање најбољег решења у случају сингуларности [110]. Дистрибуција нормализованих вредности за температуре *Temp1* и *Temp4* код исправних и отказалих дискова приказана је на Слици 42, у облику графикона генерисаног помоћу *Matlab* уграђене функције *plot*.



Слика 42. Дистрибуција нормализованих вредности температуре *Temp1* и *Temp4*

На *Слици 42*, се може уочити да вредности температура код отказалих дискова значајније одступају у односу на исправне дискове. Вредности температура исправних дискова углавном су распоређене у непосредној близини центра дистрибуције података, који је дефинисан као средња вредност атрибута за све исправне дискове. С обзиром да су атрибути температуре високо корелирани традиционалне метрике за мерење удаљености попут Еуклидског растојања, нису применљиве у овом случају. Уместо њих користи се Махаланобисово растојање, које узима у обзир корелацију између атрибута.

Временска серија Махаланобисовог растојања добија се када се за сваку инстанцу података одреди Махаланобисово растојање у односу на центар дистрибуције. На *Слици 43*, приказане су временске серије Махаланобисовог растојања за исправни диск са серијским бројем 100302 и отказали диск са серијским бројем 100003, из које се може запазити осцилације у вредностима диска са серијским бројем 100003 у периоду који је претходио његовом отказу.



Слика 43. Временска серија Махаланобисовог растојања за исправан и отказали диск

У нормалном раду магнетног диска може се очекивати висока корелација између измерених температура, јер се повећање радне температуре диска равномерно одражава на све атрибуте за мерење температуре. Насупрот томе, уколико између атрибута температуре постоји значајна разлика, то може указивати на вероватни отказ магнетног диска. Инстанце података отказалог диска између којих постоји значајно одступање између вредности атрибута температура означене су на *Слици 42* са редним бројевима од 1 до 5. На *Слици 43* се код отказалог диска могу уочити максимуми вредности Махаланобисовог растојања који су последица неубичајене разлике између две вредности температуре и одговарају претходно означеним инстанцама на *Слици 42*. Ови максимуми могу бити јасни индикатори предстојећег отказа диска.

## 6 Детекција аномалија

У овом поглављу представљен је појам аномалија и приказана њихова класификација. Аномалије, као ретки догађаји, захтевају посебне алгоритме за њихову детекцију. Најчешће коришћени типови алгоритама за детекцију аномалија су приказани у наставку поглавља.

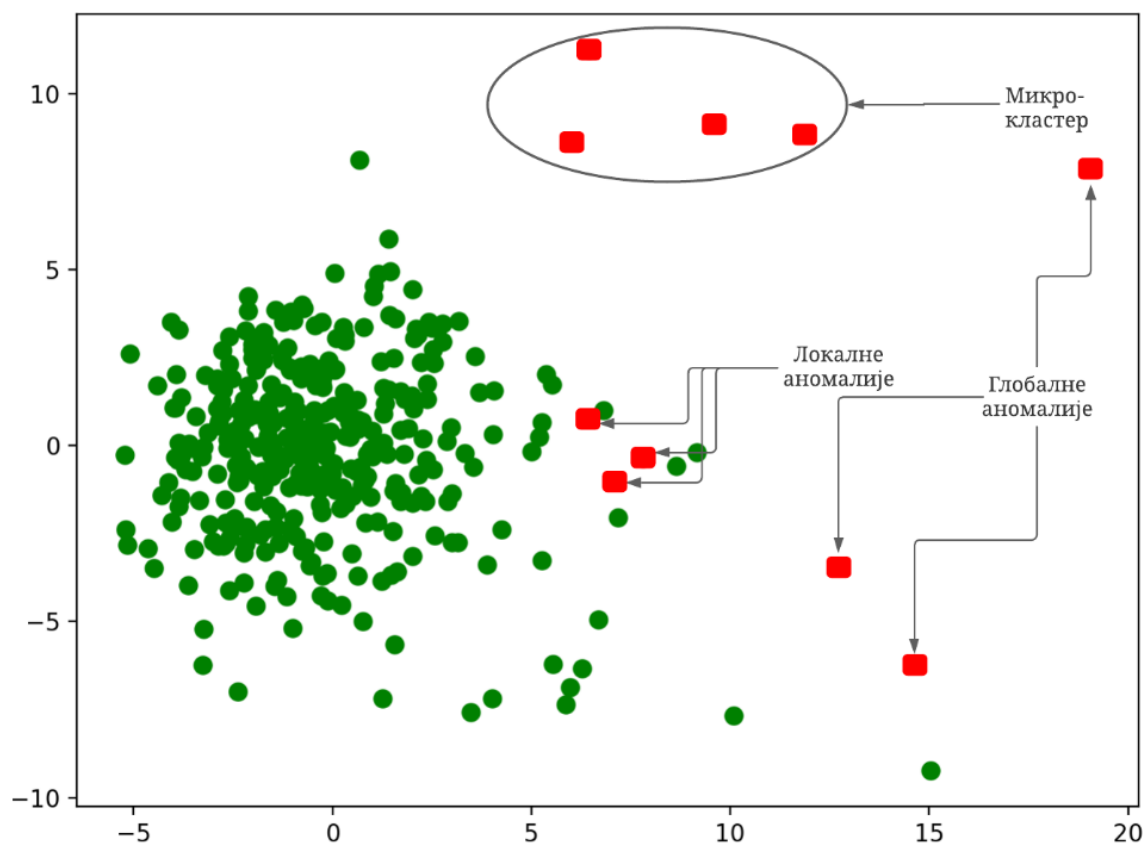
### 6.1 Класификација аномалија

Аномалија представља запажање које значајно одступа од нормалног, тј. уобичајеног запажања карактеристичног за неку појаву. Нормално понашање система подразумева да се основни процес генерисања података значајно не мења током времена, што резултира у томе да генерисани подаци не одступају значајно једни од других. Аномалије се препознају на основу вредности атрибута које значајно одступају од уобичајених нормалних вредности. Аномалије су најчешће генерисане другачијим механизмом у односу на уобичајене податке, а могу бити последица деловања шума, грешака у прикупљању података или присуства нове, непознате категорије података.

Шум у подацима настаје услед спољних утицаја чије се случајне вредности додају вредностима атрибута тренутног опажања. Утицај шума може бити толико велики да опажање значајно одступа од нормалних опажања и може се сматрати аномалијом. Вредност променљиве шума је случајна и не може се предвидети, али њене вредности прате неку статистичку расподелу, најчешће нормалну. Утицај шума на појаву аномалија може се препознати по карактеристичном расејавању опажања у свим правцима у односу на групу нормалних запажања. Аномалије које настају услед утицаја шума у подацима често се називају локалним аномалијама. Инстанце које се, услед утицаја шума препознају као аномалије потребно је уклонити из скупа података јер оне нису релевантне.

Аномалије могу настати и због грешака приликом прикупљања и обраде вредности атрибута који се користе у скупу података. Овакве вредности не задовољавају нормална ограничења одређеног атрибута (нпр. вредност температуре је приказана као 2345°C уместо 23.45°C), или је вредност атрибута приказана помоћу другог типа података (нпр. уместо нумеричке вредности атрибута приказан је низ карактера). Иако су овакве грешке ретке у поређењу са аномалијама насталим услед утицаја шума, њихова одступања су значајно већа у односу на вредности атрибута нормалних опажања. Аномалије настале услед грешака при прикупљању података су углавном значајно удаљене од нормалних инстанци и веома расејане, па се називају глобалне аномалије. Инстанце које су препознате као аномалије услед грешака у подацима најчешће се уклањају из скупа података јер могу негативно утицати на алгоритме машинског учења у процесу обуке модела.

Трећа врста аномалија односи се на нове категорије података. Оне значајно одступају од осталих опажања, али су међусобно груписане, што указује на то да су генерисане другачијим механизмом у односу на нормална опажања и називају се микрокластерима. Овај тип аномалија поседује веродостојне вредности атрибута које се могу користити у даљој анализи. На *Слици 44*. приказане су три врсте аномалија.



Слика 44. Различити типови аномалија [111]

У зависности од броја атрибута који се користе како би се нека инстанца прогласила аномалијом разликујемо тзв. моноваријабилне (енг. *univariate*) познате као тачкасте аномалије, биваријабилне (енг. *bivariate*) односно аномалије контекста и мултиваријабилне (енг. *multivariate*) тзв. колективне аномалије.

Тачкасте аномалије препознају се када вредност једнога атрибута значајно одступа од вредности тог атрибута код нормалних инстанци. Пример овакве аномалије је ако се као вредност атрибута телесне тежине појави негативан број или нека неуобичајено велика вредност.

Аномалије контекста се детектују када вредност неког атрибута није у очекиваном опсегу када се посматра у контексту вредности другог (условљеног) атрибута. Оне се не могу детектовати као тачкасте аномалије јер њихове вредности не одступају у значајној мери од вредности атрибута нормалних инстанци. На пример, негативне вредности температуре ваздуха саме по себи нису неуобичајене, али ако се такве вредности појаве у подацима за летње месеце, може се сматрати да се такве инстанце контекстне аномалије и потребно их је уклонити из скупа података.

Колективне аномалије препознају се када вредности већег броја атрибута заједно одступају у односу на нормалне вредности. Ове аномалије се не препознају на исти начин као тачкасте и као контекстне аномалије. На пример, резултати лабораторијских биохемијских анализа не морају значајно одступати од референтних вредности, али заједничка одступања неколико параметара могу указати на одређено здравствено стање особе.

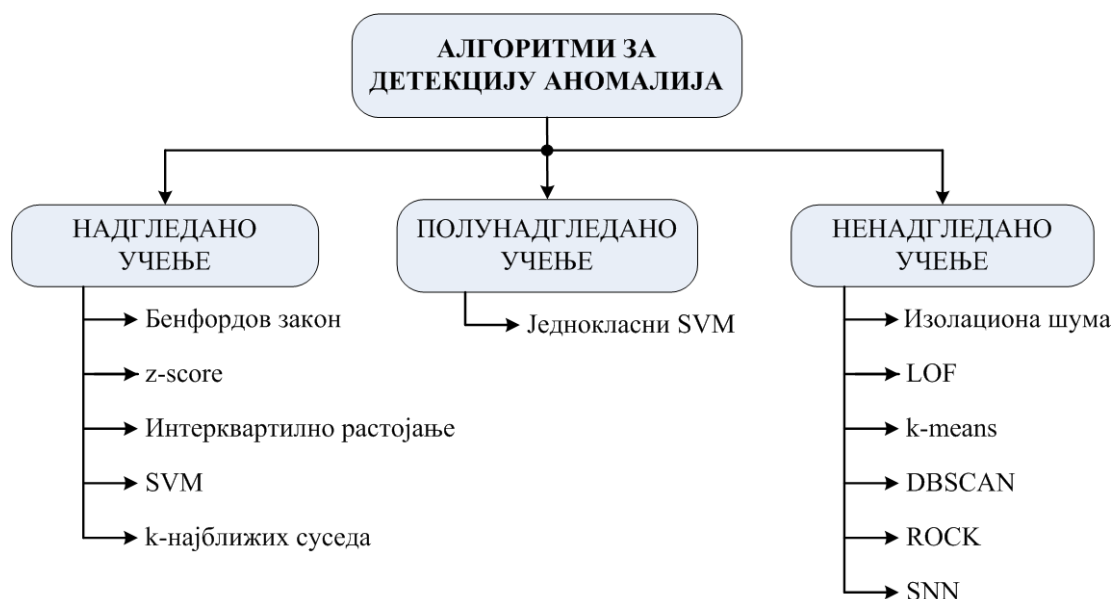
## 6.2 Алгоритми за детекцију аномалија

Аномалије представљају ретке догађаје који значајно одступају од уобичајених образаца понашања. Главни изазов са којим се сусрећу алгоритми за детекцију аномалија је исправно класификовање ових ретких догађаја у присуству великог броја исправних податка. Неадекватно обучени алгоритми за детекцију аномалија могу генерисати велики број лажних упозорења.

Области у којима се примењују алгоритми за детекцију аномалија су:

- финансије - откривање злоупотреба у облику неовлашћених трансакција, превара са кредитним картицама, лажних захтева за повраћај пореза и неуобичајених образаца трговања;
- безбедност рачунарских мрежа – идентификовање неовлашћеног приступа, напада на мреже и злонамерног софтвера;
- јавна безбедност - откривање ризичних ситуација на јавним скуповима и у саобраћају коришћењем видео надзора;
- индустријска производња - идентификовање дефектних производа или недостатака у паковању производа;
- производни системи - откривање грешака како би се обезбедило ефикасно управљање, рано откривање кварова и смањили негативни утицаји на производњу и животну средину.

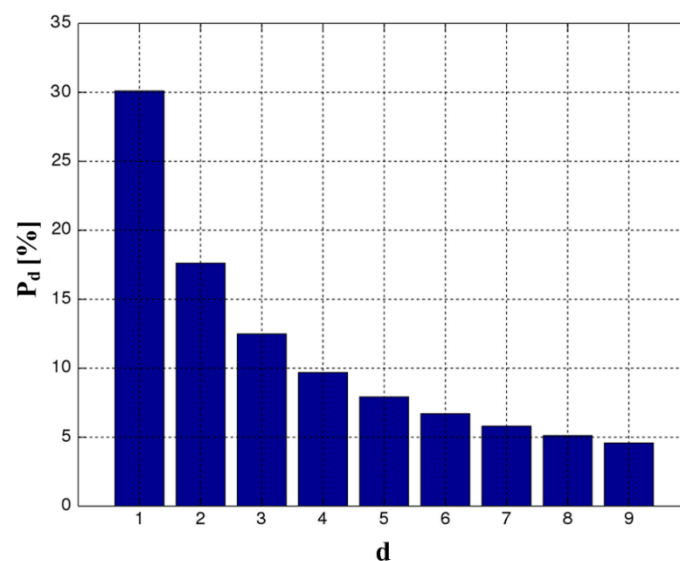
Алгоритми за детекцију аномалија могу се обучавати помоћу техника надгледаног (енг. *supervised*), полунадгледаног (енг. *semi-supervised*) или ненадгледаног (енг. *unsupervised*) учења, у зависности од структуре расположивог скупа података. Код надгледаног учења проблем се јавља због неуравнотежености скупа података јер је број инстанци, означених као аномалије, знатно мањи од броја нормалних инстанци. Полунадгледано учење користи мањи део скупа података са придруженим ознакама за нормалне инстанце и аномалије, док је већина података у скупу неозначена. Ненадгледано учење открива аномалије на основу особина већинског дела скупа података који представља нормалне инстанце. Алгоритми за детекцију аномалија могу се класификовати према подели приказаној на Слици 45.



Слика 45. Класификација алгоритама за детекцију аномалија

Статистички алгоритми за детекцију аномалија ослањају се на креирање статистичког модела за нормалне инстанце података како би се утврдило да ли се нове инстанце уклапају у креиран модел. Инстанце података које одступају од креираног модела се означавају као аномалије. Ови алгоритми су посебно ефикасни у откривању аномалија у моноваријабилним (енг. *univariate*) подацима.

Бенфордов (енг. *Benford*) закон, познат као и закон прве цифре, дефинише расподелу водећих цифара у бројевима из стварних података. Ова расподела није нормална, већ прати специфичан образац. Овај закон се може применити на нумеричке скуповете података прикупљених из извора података из стварног живота. Према Бенфордовом закону вероватноћа да ће водећа цифра броја бити јединица износи 30%, док се веће цифре појављују као водеће, са све мањом учесталošћу. Најређе се као водећа цифра појављује цифра девет, која је водећа само у 4.6% случајева [112]. У случају децималног бројног система са основом десет, вероватноћа појаве прве цифре приказана је графиком на Слици 46.



Слика 46. Вероватноћа појаве водеће цифре према Бенфордовом закону [113]

Ову расподелу прате многи природни и друштвени процеси, што омогућава њену примену у детекцији аномалија. На пример, приликом анализе неког финансијског извештаја, ако су подаци кривотворени, расподела водећих цифара неће пратити Бенфордов закон, који је дефинисан формулом (33):

$$P(d) = \log_b(d+1) - \log_b d = \log_b \left(1 + \frac{1}{d}\right) \quad (33)$$

где је,

$b$  - основа коришћеног бројног система;

$d$  - водећа цифра из опсега од 1 до  $b-1$ .

Статистичка *z-score* метода детектује аномалије тако што идентификује вредности која одступају за унапред дефинисани број стандардних девијација од средње вредности. Она представља вредности атрибута у нормализованом облику тако што се из скупа података израчунава средња вредност  $\mu$  и стандардно одступање  $\sigma$ , а затим се за сваки податак  $x$  израчунава *z-score* према формули (34):

$$z = \frac{x - \mu}{\sigma} \quad (34)$$

где је:

$\mu$  – средња вредност скупа података;

$\sigma$  – стандардно одступања скупа података;

$x$  – вредност посматраног атрибута.

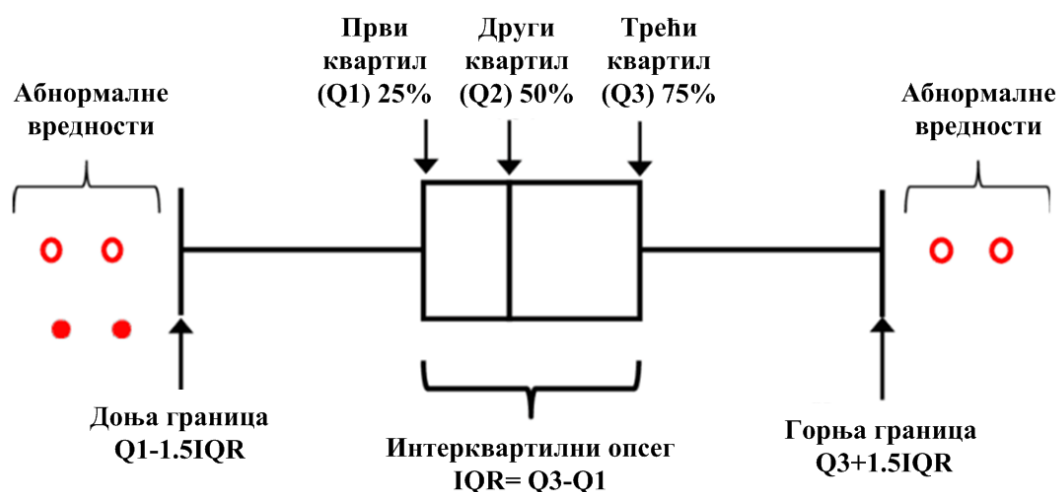
Метода  $6\sigma$  се најчешће користи за детекцију аномалија у области контроле квалитета. Ова метода је заснована на нормалној дистрибуцији података, при чему се 99.7% инстанци нормалних података налази у опсегу  $\pm 3\sigma$  од средње вредности. Уколико се неко опажање налази на растојању већем од  $\pm 3\sigma$  у односу на средњу вредност оно се проглашава аномалијом. У другим областима примене, граница за детекцију аномалија може се прилагодити у зависности од заступљености аномалија у односу на већину нормалних инстанце података.

Модификована  $z$ -score метода користи се у случајевима када дистрибуција вредности података не прати нормалну расподелу. Ова метода посматра одступања у односу на медијану, што је чини мање осетљивом на утицај аномалија које могу утицати на средњу вредност дистрибуције, као што је случај са класичном  $z$ -score методом.

Детекција аномалија коришћењем интерквартилног опсега IQR (енг. *Interquartile range*) ослања се на поделу посматраног скупа података на кватиле (енг. *quartile*). Тиме се дефинишу следеће категорије:

- први кватил  $Q1$  обухвата 25% података са најмањом вредношћу,
- други кватил  $Q2$  обухвата податке од првог кватила до медијане,
- трећи кватил  $Q3$  обухвата податке од медијане до 75% података и
- четврти кватил  $Q4$  обухвата преосталих 25% података са највећим вредностима.

Интерквартилни опсег IQR представља растојање између трећег и првог кватила ( $IQR = Q3 - Q1$ ) и користи се за одређивање граница изван којих се сва опажања могу сматрати аномалијама. Типична вредност константе за интерквартилни опсег је 1.5, на основу које се одређују горња и доња граница за детекцију аномалија. Горња граница се одређује ако је вредност опажања већа од  $Q3 + 1.5IQR$ , док се доња граница одређује ако је вредност опажања мања од  $Q1 - 1.5IQR$ , као што је приказано на Слици 47.



Слика 47. Интерквартилно растојање [114]

За откривање аномалија у мултиваријантним подацима неопходно је користити меру раздаљине у вишедимензионалном простору. У зависности од односа између вредности појединачних димензија, користе се одговарајуће метрике раздаљине као што су Еуклидска, Менхетен, косинусна и Махаланобисово растојање.

Алгоритми засновани на методама мерења близине утврђују присуство аномалије на основу различитости у односу на нормалне инстанце података. Мера различитости може се утврдити мерењем удаљености, мерењем густине или груписањем.

Најједноставнији алгоритам заснован на мерењу удаљености одређује растојање тачке од свих осталих тачака и сумира измерена растојања. Тачка чија је сума растојања до осталих тачака највећа, проглашава се аномалијом. Недостатак овог алгоритма је што може детектовати само глобалне појединачне (тачкасте) аномалије које су значајно удаљене од скупа података, а не може детектовати локалне аномалије које су непосредно у близини нормалних података [115].

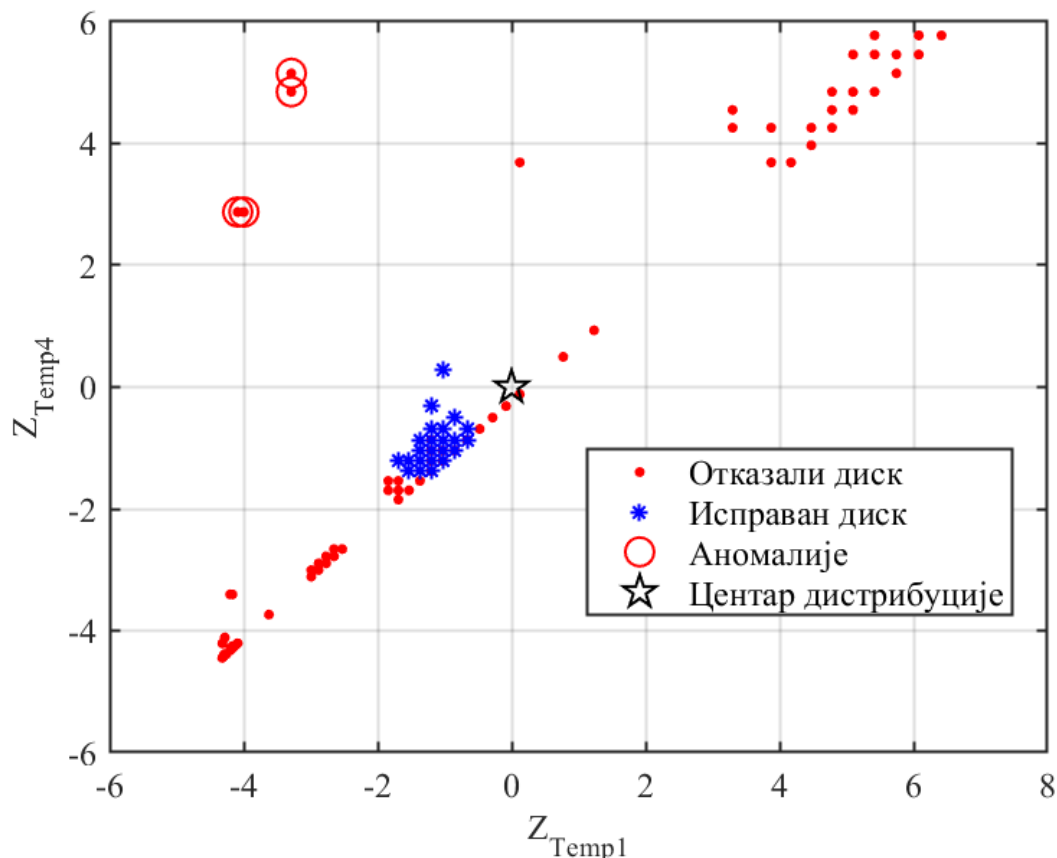
Алгоритам који користи растојање до најближег суседа за детекцију аномалија, израчунава удаљеност до најближе суседне тачке и проглашава аномалијом ону тачку која има највећу удаљеност од свог најближег суседа. Овај алгоритам је погодан за детекцију глобалних аномалија, али може имати потешкоћа са аномалијама које се налазе у областима са великом густином. Унапређена верзија овог алгоритма, позната као метода *k*-најближих суседа (енг. *k-Nearest Neighbors*) [116] дефинише меру удаљености на основу растојања до *k*-најближих суседа. Ова метода заснована је на претпоставци да ће аномалије имати веће удаљености до својих *k*-најближих суседа у односу на удаљеност нормалних тачака података до њихових *k*-најближих суседа. Мера удаљености може бити сума или средња вредност удаљености до *k*-најближих суседа. Такође, мера удаљености може бити и број најближих суседа који су удаљени преко унапред дефинисаног прага. Проблематика са којом се ова метода сусреће огледа се у потреби да се измери удаљеност до свих преосталих тачака скупа података како би се издвојило *k*-најближих суседа, што је рачунарски захтевно за велике скупове података. С тога се скуп података дели на подскупове како би се мериле удаљености тачке од најближег подскупа и тиме смањила комплексност израчунавања удаљености. Метода *k*-најближих суседа погодна је за налажење глобалних аномалија које су појединачне или груписане, док се са овом методом не могу откривати локалне аномалије.

Методе засноване на густини, као што је LOF (енг. *Local Outlier Factor*) метода [117], упоређују густину око посматране тачке са густином око њених суседа. Ако је густина око посматране тачке знатно нижа у односу на густину њених *k*-најближих суседа, та тачка се класификује као аномалија. Густина се одређује на основу раздаљине до најудаљеније тачке која описује круг око посматране тачке у којем се налази свих *k*-најближих суседа. Што је раздаљина до најудаљенијег суседа већа, густина ће бити мања. Добијена густина се пореди са средњом вредношћу густине за сваког од *k*-најближих суседа, која се мери за њихове најближе суседе. Ако је однос густине мањи од 1 то показује да је посматрано опажање аномалија. LOF метода погодна је за откривање како локалних, тако и глобалних појединачних аномалија, а може откривати и глобалне груписане аномалије ако је њихов број у групи мањи од *k*.

Алгоритми изолационе шуме (енг. *Isolation Forest*) [118], креирају бинарна стабла са циљем да изолују појединачне тачке података. С обзиром да је процес креирања стабла насумичан, потребно је креирати велики број стабала како би се формирала шума. Основна претпоставка ове методе је да је аномалије много лакше изоловати, јер су удаљена од нормалних тачака. Као резултат тога, аномалије ће се обично налазити на мањој просечној дубини у бинарном стаблу у односу на нормалне тачке.

Алгоритми за детекцију аномалија засновани на кластеровању користе се у случају ненадгледаног учења и они теже да групишу сличне инстанце података у кластере. Тачке података које представљају аномалије обично неће припадати ни једном кластеру или ће бити удаљене од центра најближих кластера. Најчешће коришћени алгоритми у овој области су DBSCAN (енг. *Density-Based Spatial Clustering of Applications with Noise*) и ROCK (енг. *Robust Clustering Algorithm for Categorical Attributes*). Иако су ови алгоритми првенствено дизајнирани за кластеровање, могу се користити и за детекцију аномалија. Појединачне (тачкасте) аномалије се могу открити мерењем раздаљине до центра најближег кластера, јер ће се налазити на већој удаљености у односу на нормалне инстанце података. Групне аномалије могу се идентификовати мерењем густине кластера, јер оне теже да припадају кластерима са малом густином и малим бројем тачака, за разлику од кластера нормалних инстанци који имају велику густину и велики број тачака.

На Слици 48 приказане су нормализоване вредности два атрибута температуре за исправни и отказали диск из CMRR скупа података. Применом *Matlab* уграђене функције *lof* креиран је LOF модел за детекцију аномалија који је обучен са инстанцама исправних дискова. Овај модел је у стању да открије аномалије на основу њихове густине коришћењем мере удаљености. Применом функције *isanomaly* над инстанцама исправног и отказалог диска детектоване су четири аномалије код инстанци отказалог диска које су приказане на Слици 48.



Слика 48. Детекција аномалија у подацима два атрибута температуре  $Temp1$  и  $Temp4$

## 7 Модел за предикцију отказа магнетних дискова заснован на детекцији аномалија

У овом поглављу представљен је модел за предвиђање отказа магнетних дискова на бази алгоритма за детекцију аномалија. Реализовани модел ослања се на низ алгоритама представљених у претходним поглављима, који се користе за трансформацију скупа података како би алгоритам за детекцију аномалија исправно предвиђао отказе магнетних дискова. Кључни параметар предиктивних перформанси представљеног модела је постизање високе стопе предикције отказа при чему се минимизује појава лажних детекција отказа магнетних дискова.

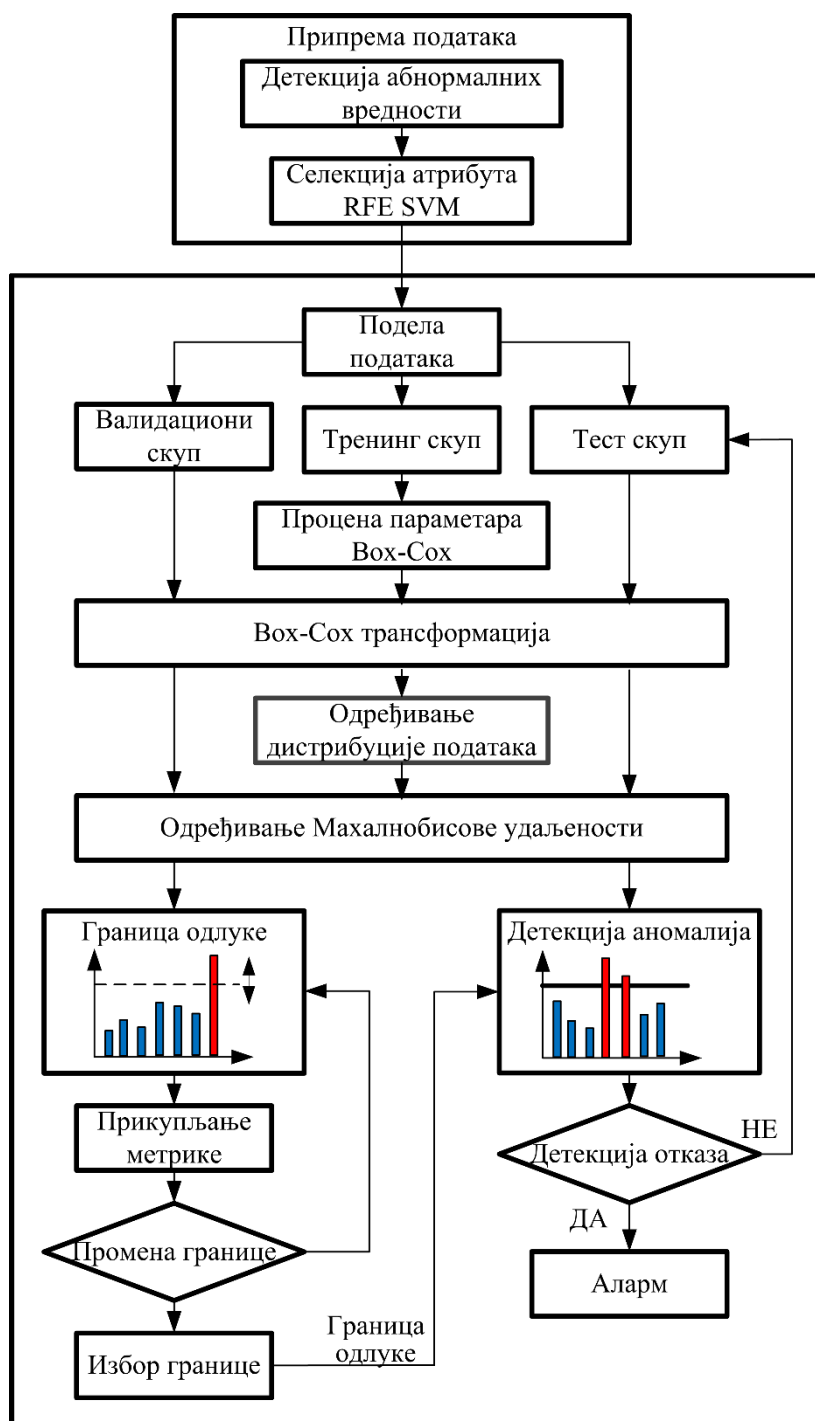
### 7.1 Структура предложеног модела

Под отказом се подразумева немогућност система или компоненте да изврши своје функције у складу са дефинисаним захтевима перформанси [119]. Савремени системи и компоненте пројектују се са високим степеном поузданости, због чега су откази ретки догађаји све док систем ради у прописаним условима. Стога отказе можемо посматрати као аномалије, јер у тим случајевима систем или компонента манифестују сасвим другачије понашање од пројектованог.

Циљ овог истраживања заснован је на примени алгоритама за детекцију аномалија да би се предвидели откази магнетних дискова. Магнетни дискови су веома поуздане рачунарске компоненте, израђене са високим степеном прецизности који им омогућава непрекидан дугогодишњи рад у рачунарсим системима за складиштење података. У системима центара података магнетни дискови раде у оптималним условима што им омогућава дуг животни век. Ипак, повремено долази до кварова који доводе до отказа магнетних дискова, са стопом отказа која износи око 1% на годишњем нивоу. На пример, у популацији од 100.000 дискова који раде у центру података у режиму 24/7, очекује се 1000 отказа дискова годишње, што у просеку значи три отказа диска на дневном нивоу. Имајући ово у виду може се закључити да су откази дискова ретке појаве и да се могу сматрати аномалијама. Праћењем рада дискова помоћу SMART атрибута могуће је идентификовати неправилан рад магнетног диска и благовремено креирати резервну копију података на другом магнетном диску пре него што наступи отказ.

У оквиру ове дисертације развијен је модел за предвиђање отказа магнетних дискова, заснован на алгоритму за детекцију аномалија који користи мерење удаљености [120]. Алгоритамска структура модела приказана је на *Слици 49*, и састоји се из фазе припреме података која претходи фази детекције аномалија.

Прва фаза подразумева припрему података, која обухвата учитавање скупа података, уклањање абнормалних вредности и рангирање атрибута по значају. У овој фази изводе се статистички тестови за детекцију вредности ван граница (енг. *outliers*) како би се елиминисале абнормалне вредности SMART атрибута, које би негативно утицале на тачност предвиђања отказа. Идентификоване абнормалне вредности SMART атрибута у овом кораку замењене су средњом вредношћу атрибута у целокупном скупу података. У ову сврху коришћена је *outlier* библиотека из програмског пакета R.



Слика 49. Алгоритамска структура модела за предвиђање отказа [120]

Атрибути који се налазе у скупу података имају различит утицај на предикцију отказа. Неки атрибути имају значајан утицај у односу на друге, док постоје и они атрибути који немају никакав утицај на предвиђање отказа. Из тог разлога је неопходно извршити рангирање атрибута према његовом утицају на предикцију отказа.

У представљеном моделу користи се метода рекурзивне елиминације атрибута SVM-RFE која итеративно из скупа података уклања један атрибут са најмањим утицајем на предвиђање отказа. За предвиђање отказа коришћен је SVM класификатор са линеарним кернелом, који се обучава на скупу атрибута додељујући тежине сваком од њих. SVM са линеарним кернелом је изабран у односу на друге типове кернела јер експериментално посматрање показује да су подаци приближно линеарно одвојиви.

Атрибут који у процесу евалуације има најмању тежину се елиминише из скупа података, а процес се итеративно понавља све док у скупу података не остане само један, најутицајнији атрибут. Процес рекурзивне селекције атрибута реализован је у програмском пакету R коришћењем библиотеке *svm-rfe* и *svm* функције из библиотеке *e1071*.

Након што је скуп података претходно обрађен и листа атрибута уређена, следећи корак је подела скупа података на три подскупа: скуп за обуку (тренинг скуп), скуп за валидацију и скуп за тестирање. Ова три подскупа су одабрана тако да између њих нема преклапања како би се обезбедила прецизна процена перформанси модела предвиђања и избегло прекомерно уклапање модела.

- Скуп за обуку садржи податке од 60% насумично одабраних здравих дискова и у њему нису присутне инстанце неисправних дискова.
- Скуп за валидацију садржи податке од 20% насумично одабраних здравих дискова, и податке са 50% насумично одабраних неисправних дискова.
- Скуп за тестирање укључује податке од 20% насумично одабраних здравих дискова, и податке са 50% насумично одабраних неисправних дискова.

Вредности података из скупа за обуку се користе за одређивање референтних вредности за сваки SMART атрибут на основу вредности података исправних дискова. Добијени скуп референтних вредности користи се како би се дефинисало нормално понашање магнетног диска, што подразумева да диск ради у оквиру постављених спецификација перформанси.

С обзиром да је рад магнетних дискова праћен мултиваријабилним SMART атрибутима, као меру удаљености неопходно је користити Махаланобисово растојање. Референтне вредности се користе за одређивање центра масе дистрибуције података у односу на који се одређује Махаланобисово растојање за сваку инстанцу скупа података.

Предуслов за одређивање Махаланобисовог растојања је да подаци прате нормалну расподелу. С обзиром да су вредности SMART атрибута типично непараметарски распоређени, неопходно је трансформисати вредности података сваког атрибута како би оне пратиле нормалну дистрибуцију. За трансформацију података у нормалну дистрибуцију је примењена *Box-Cox* трансформација. Ова трансформација на основу скупа за обуку за сваки SMART атрибут одређује одговарајући експонент  $\lambda$  који стабилизује варијансу трансформисаних података. Трансформација скупа података извршена је применом уграђене *Matlab* функције *boxcox* са општом синтаксом  $[transdat, \lambda] = boxcox(data)$ , којој се као параметар *data* прослеђује непараметарски скуп података. Функција као повратне вредности враћа низ експонената  $\lambda$  са којим је сваки атрибут из скупа података *transdat* трансформисан у нормализован облик. У овом случају, скуп података за обуку прослеђен је као непараметарски скуп података функцији *boxcox*, која је одредила оптималну вредност низа експонената  $\lambda$   $[\lambda_1, \lambda_2, \dots, \lambda_m]$ . С обзиром да *Box-Cox* трансформација дозвољава само позитивне вредности које су веће од нуле, неопходно је вредности сваког атрибута увећати за један како би се избегли проблеми са атрибутима који имају нулте вредности. На основу добијених вредности низа експонената  $\lambda$ , трансформисан је целокупан скуп података, тако да су вредности сваког атрибута постале приближно нормално дистрибуиране. Након примене *Box-Cox* трансформације, скуп података се стандардизује помоћу *z-score* на основу средње вредности  $\mu$  и стандардног одступања  $\sigma$  за вредности атрибута исправних дискова из скупа за обуку, како би се елиминисао ефекат скалирања.

Трансформисани подаци из скупа за обуку се даље користе за одређивање центра масе података исправних дискова, како би се измерило Махаланобисово растојање свих

тачака података у друга два подскупа података. Махаланобисово растојање је растојање између одређене тачке у мултиваријантном простору и тачке центра масе. На основу скупа за обуку, Махаланобисово растојање се израчунава за све податке из скупа, а свака инстанца података је представљена једном вредношћу Махаланобисовог растојања у односу на центар масе.

Махаланобисово растојање одређено је помоћу уграђене *Matlab* функције *mahal* са општом синтаксом *mahal(Y,X)*, која враћа квадрат Махаланобисовог растојања за сваку инстанцу у скупу  $Y$  у односу на референтни скуп инстанци  $X$ . У конкретном случају функцији *mahal* прослеђен је целокупан нормализован скуп података као параметар  $Y$ , док је за параметар  $X$  прослеђен нормализован скуп података за обуку. Пошто скуп података садржи неке атрибуте са константним вредностима за исправне дискове, додаје им се минимална вредност Гаусовог шума како би се избегли проблеми са сингуларношћу приликом израчунавања матрице коваријансе.

Мера удаљености на основу које се препознаје аномалија утврђује се у поступку надгледаног учења, при чему се модел обучава над валидационом скупу података док се не постигне задовољавајући ниво перформанси у предикцији отказа. Алгоритам за детекцију аномалија обучен је коришћењем вредности Махаланобисовог растојања инстанци из скупа за валидацију. Скуп за валидацију користи се за подешавање хиперпараметара модела машинског учења са циљем да се максимизира нека од метрика коришћених за процену перформанси предикције. Овај скуп садржи инстанце података за исправне дискове (означени као стварно негативни) и отказале магнетне дискове (означени као стварно позитивни) и они се користе како би се одредила гранична вредност Махаланобисовог растојања. Све инстанце података које се налазе на растојању мањем од граничне вредности биће класификоване као исправне, док ће инстанце са већом раздаљином од граничне вредности бити класификоване као откази.

Избор граничне вредности имаће различит утицај на прецизност и осетљивост, кључне метрике које се користе код предикције отказа магнетних дискова. Прецизност представља меру способности модела да не детектује лажне отказе диска, како би се избегле непотребне акције над дисковима који су исправни а модел их класификује као отказале. У идеалном случају, са стопом лажних отказа једнаком нули прецизност модела износи 100%. Осетљивост представља способност модела да предвиди све отказе и у идеалном случају она износи 100%. С обзиром да инстанце исправних и отказалих дискова нису јасно разграничене, неминовно је да ће у процесу предикције доћи до грешака у виду предикције лажних отказа (лажно позитивни) и недетекованих отказа (лажно негативни). Предикција лажних отказа ће негативно утицати на прецизност, док ће недетековани откази негативно утицати на осетљивост. Када се граница одлучивања постави јако ниско, тада ће модел успети да детектује скоро све отказе, што ће резултирати високом осетљивошћу. Међутим, то може довести до детекције великог броја лажних отказа па ће прецизност модела бити ниска. У супротном, ако се граница одлучивања постави довољно високо да модел не детектује лажне отказе, прецизност ће бити блиска 100%, међутим одређени број отказа неће бити детектован, па ће осетљивост бити ниска.

Метрика која се користи за процену перформанси предикције отказа магнетних дискова је осетљивост, која се одређује при максималној прецизности (ниска стопа лажних отказа која је блиска нули). Оптимална гранична вредност одређује се итеративно у одређеном опсегу вредности тако што се за сваку њену вредност испитују предиктивне перформансе. На основу параметара предиктивних перформанси који је

остварио највећу прецизност, бира се граница одлучивања. Граница одлучивања  $\varepsilon$  се одређује претраживањем интервала у две фазе:

- У првој фази грубом претрагом проналази се најбоља граница одлучивања у логаритамској скали у опсегу  $[10^{-3} \div 10^{12}]$ , како би се убрзала претрага у широком опсегу вредности. Вредност границе одлучивања за коју се не детектује ниједан лажно позитиван отказ (0% FAR) бира се као груба вредност за границу одлуке  $\varepsilon_c$ .
- Друга фаза подразумева креирање нове линеарне скале, у опсегу између грубе граничне вредности  $\varepsilon_c$  и њеног претходног елемента  $\varepsilon_{c-1}$ . Линеарна скала подељена је на 1000 елемената, при чему се први елемент за који се постигне максимална прецизност (0% FAR) бира као коначна вредност за границу одлуке  $\varepsilon$ .

Одабрана вредност границе одлуке  $\varepsilon$  се користи за процену перформанси предложеног алгоритма за детекцију отказа на скупу за тестирање. Скуп података за тестирање представља независан део скупа података који се користи само за процену перформанси обученог модела машинског учења. Он не садржи инстанце које се користе у процесу обуке и валидације како би се избегао проблем претераног уклапања (енг. *overfitting*). Циљ коришћења независног скупа за тестирање је да се предиктивне перформансе обученог модела непристрасно тестирају. Скуп за тестирање користи атрибуте сваке инстанце како би модел предвидео очекивани исход, који се пореди са стварним исходом за ту инстанцу. Добијени резултати поређења се представљају у форми матрице конфузије, из којих се израчунавају жељени параметри предиктивних перформанси обученог модела. Параметри који су генерисани при тестирању модела су прецизност, стопа лажних аларма, осетљивост и вредност F0.5 мере. Алгоритми за предвиђање отказа диска су обучени на такав начин да смање број лажно позитивних класификација (лажних аларма) и максимизирају прецизност, што је условна метрика која се комбинује са осетљивошћу. Веће вредности осетљивости се могу постићи снижавањем границе одлучивања, али то ће резултирати смањеном прецизношћу и већим бројем лажних аларма. Како би се остварила равнотежа између прецизности и осетљивости у области предикције отказа магнетних дискова користи се F0.5 мера, која двоструко више вреднује прецизност у односу на осетљивост.

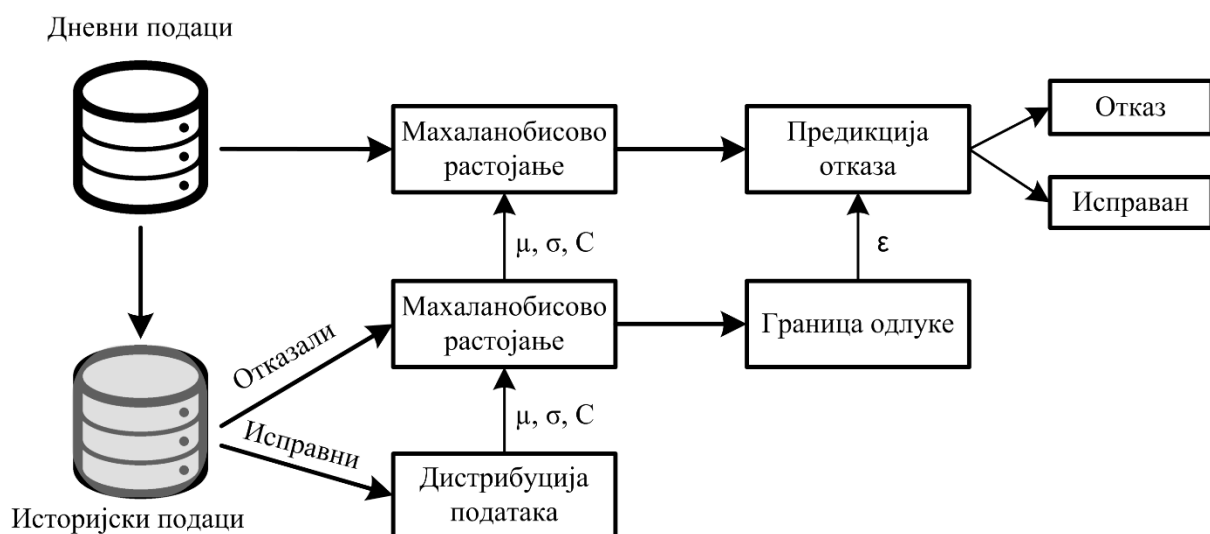
Предикција отказа магнетног диска са практичног становишта оправдана је ако се отказ предвиди довољно рано пре наступа стварног отказа. Време између предвиђеног и стварног отказа, познато као време до отказа, треба да буде довољно дуго да се подаци могу мигрирати са неисправног диска на резервни исправан диск. Ово време би требало да буде најмање 24 часа како би се омогућила израда резервне копија података. Овакав приступ спречава губитак корисничких података а у центрима података елиминише потребу за реконструкцијом изгубљених података која може потрајати неколико дана.

Инстанце које припадају једном диску представљене су вредношћу Махаланобисовог растојања, груписаног у временску серију. Представљени модел је у стању да сигнализира тренутак када вредност Махаланобисовог растојања пређе границу одлуке. Мерењем времена између тог тренутка и тренутка стварног отказа диска могуће је утврдити време до отказа за сваки диск и на основу тога израчунати средње време до отказа.

Предложени модел у овој дисертацији се може имплементирати у реалном сценарију за предикцију отказа у центрима података, где би се користио за проактивно сервисирање кварова. Кораци неопходни за имплементацију модела су:

1. Рангирање и селекција значајних SMART атрибута магнетних дискова;
2. Прикупљање података за скуп исправних дискова и њихова трансформација у нормално дистрибуирани скуп података;
3. Формирање центра масе нормално дистрибуираних података описаног вектором средње вредности, стандардног одступања и матрицом коваријансе;
4. Инстанце тренутних SMART атрибута се користе за израчунавање Махаланобисовог растојања у односу на центар масе података (описаног вектором средње вредности, стандардног одступања и матрицом коваријансе);
5. Када наступи отказ диска, користе се његове претходне вредности Махаланобисовог растојања, на основу којег се одређује граница одлуке која ће омогућити детектовање отказа без активирања лажних аларма;
6. Појавом нових отказа, граница одлуке се прилагођава да би се обучила на новонасталим и претходним отказима, чиме се онемогућује предвиђање будућих отказа.

Предложена имплементација алгоритма за детекцију отказа магнетних дискова у центру података приказана је на *Слици 50*. Подаци за сваки диск прикупљају се на дневном нивоу и архивирају се у складишту током одређеног периода (месец, квартал или година). На основу историјских података, инстанце исправних дискова се користе за одређивање центра дистрибуције исправних података на основу којих се израчунава вектор средње вредности и стандардног одступања за сваки атрибут, као и матрица коваријансе која описује међусобну зависност атрибута. Коришћењем ових параметара одређује се Махаланобисово растојање за инстанце отказалих дискова, на основу којих се одређује граница одлуке. Граница одлуке представља праг вредности Махаланобисовог растојања изнад којег диск класификује као отказао, а испод којег диск остаје класификован као исправан. Исти поступак се примењује на актуелне дневне податке, чија се вредност Махаланобисовог растојања упоређује са границом одлуке како би се предвидео потенцијални отказ магнетног диска. Предложена имплементација је у стању да се прилагођава постепеним променама SMART атрибута које настају као последица старења, деградирања перформанси и смањења поузданости.



Слика 50. Имплементација алгоритма у центру за складиштење података

## **7.2 Примена реализованог модела у предикцији отказа полупроводничких SSD дискова**

Реализовани модел се, поред своје основне области примене може користити и у другим областима истраживања где су откази система узроковани постепеном деградацијом радних параметара. Модел се може применити у предикцији отказа полупроводничких SSD дискова, других делова рачунарских система као и на широк спектар техничких система (индустријске машине, мотори, возила и сл.). Међутим, примена модела на наведеним системима захтева одређени ниво модификације како би алгоритам за предикцију отказа задржао висок степен предиктивних перформанси. Разлог за то је чињеница да се радни параметри ових система значајно разликују од SMART атрибута који се користе за праћење стања магнетних и SSD дискова.

Реализовани модел је примењен и за предвиђање отказа полупроводничких SSD дискова над јавно доступним скупом података [121]. Механизам отказа полупроводничких SSD дискова значајно се разликује од магнетних дискова, јер SSD дискови не поседују покретне електромеханичке склопове који се временом хабају. Већина компоненти SSD дискова има тенденцију да отказује непредвидиво, без икаквих претходних упозорења. Компоненте SSD диска које су подложне „хабању“ су *FLASH* меморијске ћелије, које имају ограничен број уписа, након чега се поузданост складиштења података убрзано смањује [122]. Параметри перформанси који описују процес складиштења података у *FLASH* меморијским ћелијама уједно су и најзначајнији индикатори предстојећег отказа SSD диска.

Модел за предвиђање отказа SSD диска заснован је на представљеном основном моделу за предикцију отказа магнетних дискова, са којим дели заједнички алгоритам за детекцију аномалија [123]. Главна разлика у односу на основни модел огледа се у томе што је уместо SVM-RFE алгоритма за рекурзивну елиминацију атрибута, имплементиран алгоритам секвенцијалног избора атрибута унапред. Ова модификација модела је била потребна како би се алгоритам за селекцију атрибута директно интегрисао у модел за детекцију аномалија, за разлику од претходног решења, које је селекцију атрибута изводило независно, што је у неким случајевима доводило до мање оптималних резултата. Алгоритам секвенцијалног избора атрибута функционише тако што прво креира подскуп који садржи само један атрибут за који се тестирају предиктивне перформансе модела. Овај поступак се понавља за све појединачне атрибуте, а онај који омогућава да модел постигне најбоље перформансе проглашава се најзначајнијим атрибутом. У следећем кораку формира се подскуп који садржи два атрибута, најзначајнији атрибут и још један атрибут за који се испитују перформансе као у претходном кораку. Поступак се наставља све док се не изврши комплетно рангирање свих атрибута, при чему последњи атрибут има најмањи утицај. Ова модификација омогућила је да се користе само најутицајнији атрибути што значајно поједностављује процес предикције и побољшава предиктивне перформансе модела. Алгоритамска шема реализованог модела за предвиђања отказа полупроводничких SSD дискова приказана је на *Слици 51*.



## 8 Представљање резултата истраживања и дискусија

У овом поглављу представљени су резултати истраживања и извршена је анализа предиктивних перформанси предложеног модела за детекцију отказа магнетних дискова на CMRR скупу података. Реализовани модел упоређен је са другим моделима за предикцију отказа који су испитани над истим скупом података. Модели су међусобно упоређени на основу стопе детекције отказа, при чему се захтевао минималан број лажних аларма. Поред стопе детекције, модели су међусобно упоређени према времену за које су могли предвидети отказ унапред, у распону од 12 до 48 часова пре његове стварне појаве.

### 8.1 Приказ резултата предикције отказа магнетних дискова

У првом кораку извршен је унос CMRR скупа података у *.arff* формату у програмски пакет *Matlab*. Овај скуп података састоји се од 369 дискова са укупно  $n=68411$  инстанци података, при чему је свака инстанца описана са  $m=64$  атрибута. Атрибуте инстанце чине четири општа атрибута (серијски број, редни број записа, време рада и време до отказа), 59 SMART атрибута и индикатор отказа диска (0 за исправан диск, 1 за отказали диск).

Вредности атрибута су тестиране на присуство абнормалних вредности које су замењене средњом вредношћу атрибута. Укупан број уклоњених абнормалних вредности атрибута износи 36. Након тога, уклоњени су SMART атрибути који имају константне вредности у целом скупу података како би се смањила сложеност рачунања у даљим корацима. Од укупног броја атрибута, идентификовано је 13 SMART атрибута, чије вредности су константне у целом скупу података, који су уклоњени тако да скуп података након овог корака садржи 46 атрибута. Листа уклоњених атрибута приказана је у Табели 9.

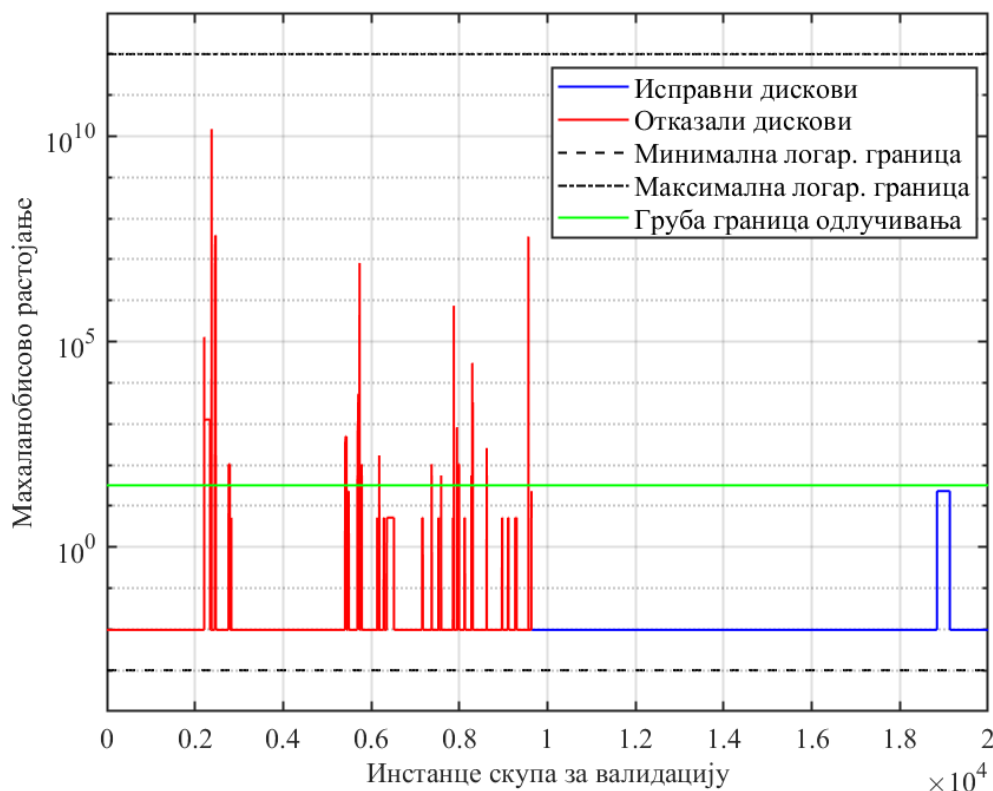
У наредном кораку извршено је рангирање атрибута коришћењем SVM-RFE алгоритма и као резултат рангирања, добијена је уређена листа са 46 атрибута, који су распоређени према њиховој важности, а која је приказана у Табели 10.

CMRR скуп података садржи 191 отказалих дискова и 178 исправних дискова, који је подељен у три насумично креирана подскупа:

- Скуп за обуку садржи податке 106 насумично одабраних исправних дискова, који чине 59.56% исправних дискова;
- Скуп за валидацију садржи податке 36 насумично одабраних исправних дискова (20.22% исправних дискова) и 95 насумично одабраних отказалих дискова (49.73% отказалих дискова);
- Скуп за тестирање укључује податке 36 насумично одабраних исправних дискова (20.22% исправних дискова) и податке 96 насумично одабраних отказалих дискова (50.26% отказалих дискова).

Како би се одредио оптимални подскуп атрибута потребних за постизање најбољих предиктивних перформанси предложеног модела за детекцију аномалија, итеративно је додаван један по један атрибут из уређене листе атрибута добијене помоћу SVM-RFE алгоритма. За формирање подскупа атрибута одређен је центар дистрибуције, на основу којег је одређено Махаланобисово растојање за сва три подскупа података. Након тога се на скупу за валидацију проналази оптимална граница одлуке која постиже најмањи

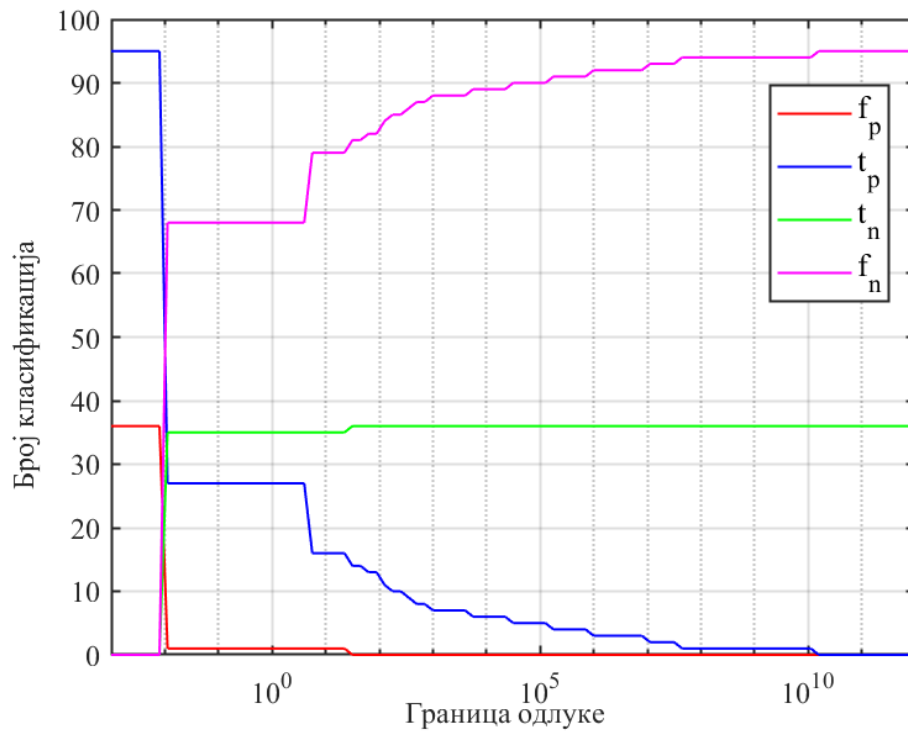
број лажно позитивних детекција. Пример проналажења границе одлучивања за најутицајнији атрибут *GList1* приказан је на *Слици 52*. За све инстанце валидационог скупа приказана је вредност Махаланобисовог растојања за исправне и отказале дискове, као и минимална и максимална вредност границе одлуке, између којих се врши претрага како би се пронашла најоптималнија граница одлуке. Све инстанце чија је вредност Махаланобисовог растојања већа од границе одлуке класификоване су као откази, док су преостале инстанце класификоване као исправне. Ова граница мора бити изабрана тако да не би дошло до лажних детекција отказа, тј. да инстанце исправних дискова не буду погрешно класификоване као откази.



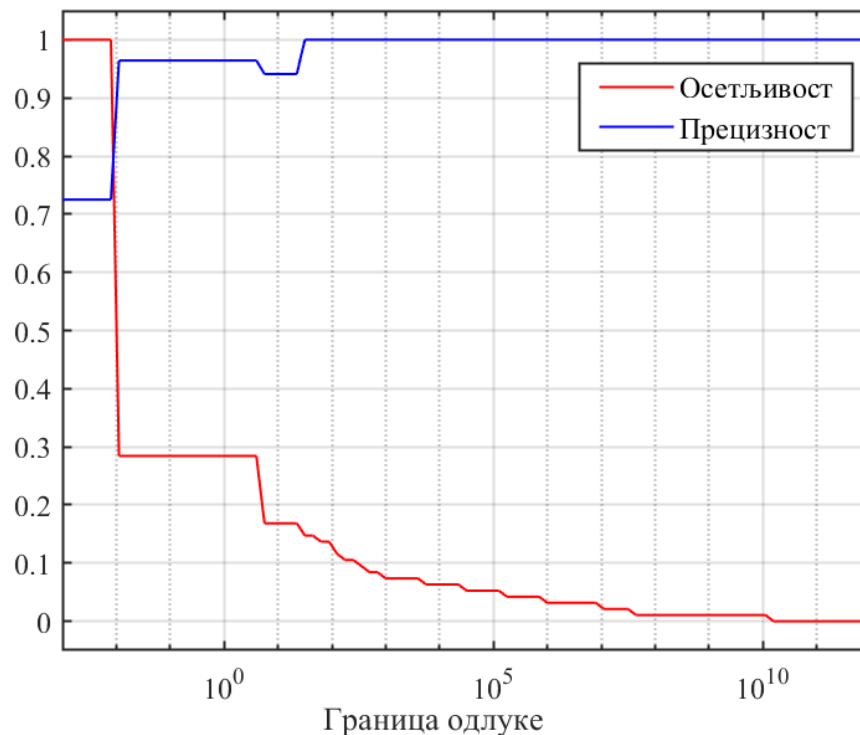
Слика 52. Одређивање грубе границе одлучивања за најутицајнији атрибут *GList1*

Граница одлуке претражује се у две фазе. Најпре се проналази груба граница одлуке на ширем интервалу вредности у логаритамској скали, након чега се оптимална граница одлуке проналази претрагом ужег интервала у околини изабране грубе границе одлуке. Груба граница одлуке се претражује на интервалу између минималне и максималне границе одлуке у логаритамској скали у опсегу  $[10^{-3} \div 10^{12}]$ . Када је граница одлуке блиска минималној вредности све инстанце валидационог скупа класификују се као откази. У том случају је број исправно позитивних детекција  $t_p$  максималан и износи 95, али је истовремено и број лажно позитивних детекција  $f_p$  такође максималан и износи 36, као што је приказано на *Слици 53*. Са даљим повећавањем границе одлучивања број инстанци које су класификоване као откази се постепено смањује. Када вредност границе одлучивања достигне своју максималну вредност, ниједна инстанца неће бити класификована као отказ. У том случају број исправно позитивних детекција  $t_p$  је једнак нули (сви откази су класификовани као погрешно негативни  $f_n$ ), док је број погрешних детекција  $f_p$  такође једнак нули (сви исправни дискови су тачно класификовани  $t_n$ ). Груба граница одлучивања је она вредност за коју је број погрешних детекција  $f_p=0$ , а у овом случају она износи  $\varepsilon=31.62$  за атрибут *GList1*. За ову вредност границе одлучивања број исправно класификованих отказа износи  $t_p=14$ , док је број неисправно класификованих отказа  $f_n=81$ . Избор границе одлуке може се заснивати и на другим метричким

критеријумима, као што је тачност, позитивни однос вероватноће и други, у зависности од области примене модела. У случају предикције отказа SSD дискова користи се метрика максималне тачности. Перформансе класификатора у зависности од вредности границе одлуке приказане су у облику прецизности и осетљивости (Слика 54). Модел који користи најзначајнији атрибут *GList1* са грубом границом одлучивања постиже осетљивост од 14.74% уз прецизност од 100% на валидационом скупу.

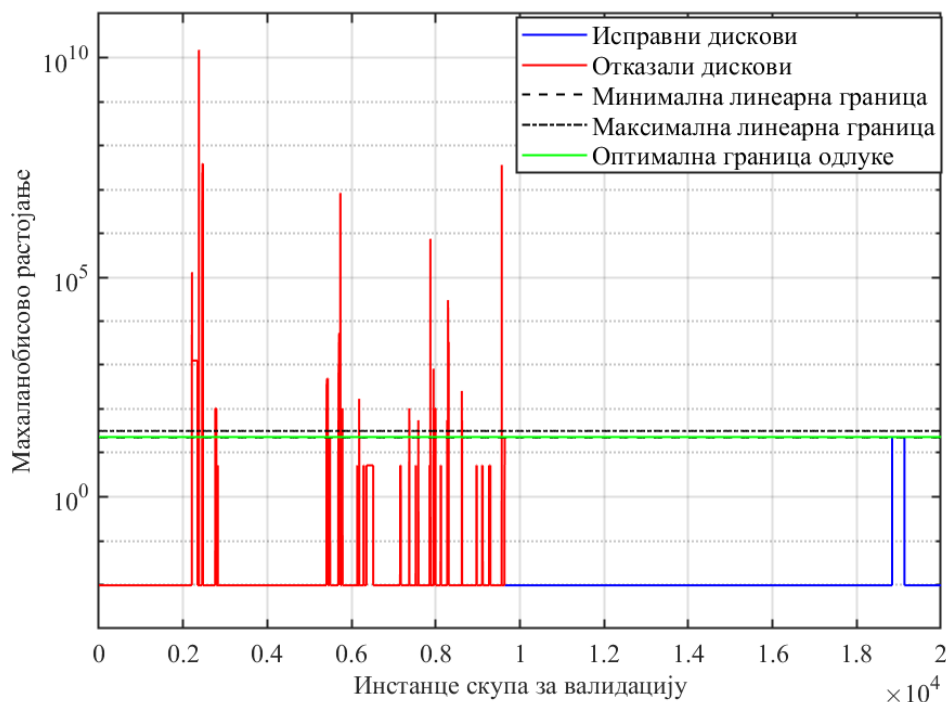


Слика 53. Утицај границе одлуке на број исправних и погрешних класификација за атрибут *GList1*



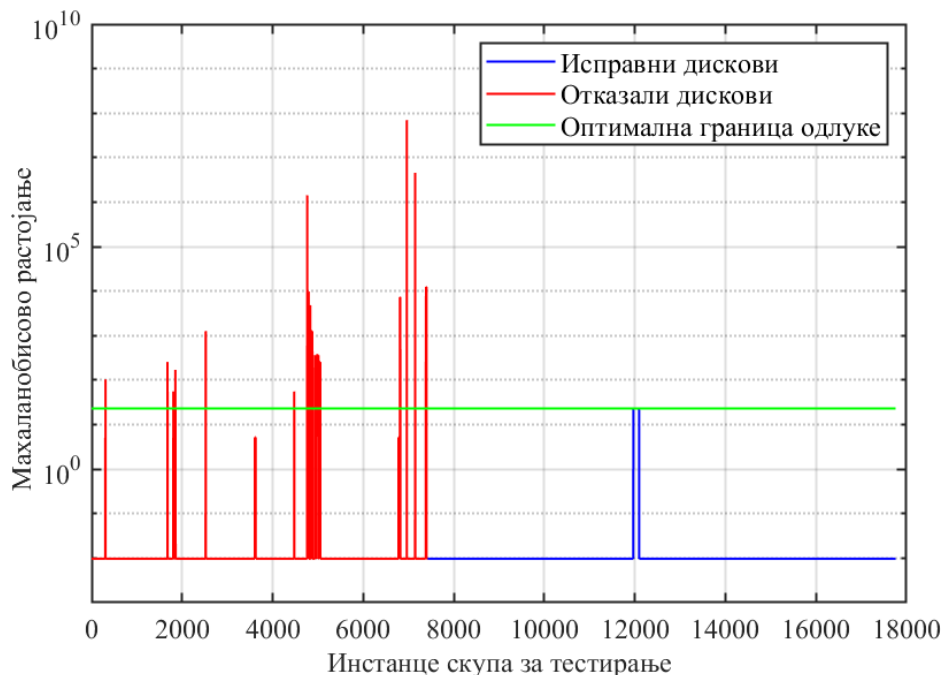
Слика 54. Перформансе класификације у зависности од вредности границе одлуке за атрибут *GList1*

Претрагом границе у логаритамској скали проналази се груба вредност границе одлуке, након чега се вредност оптималне границе одлуке проналази претрагом у линеарној скали у околини грубе границе. На *Слици 55* приказан је пример избора вредности оптималне границе одлучивања за атрибут *GList1* која износи  $\varepsilon=22.39$ .



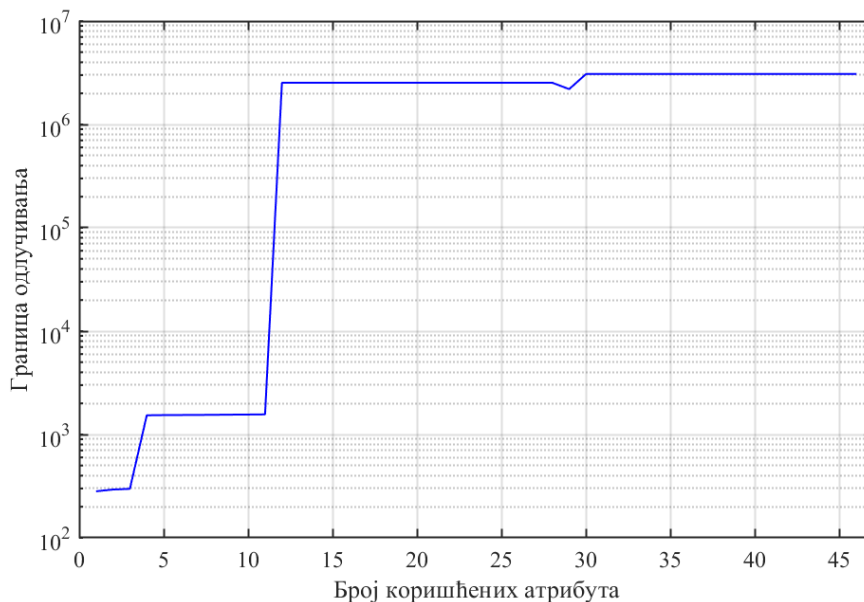
Слика 55. Одређивање оптималне границе одлучивања за атрибут *GList1*

Предиктивне перформансе модела за оптималну границу одлучивања одређене су на независном скупу за тестирање. Све вредности Махаланобисовог растојања за инстанце из скупа за тестирање су упоређене са оптималном границом одлучивања како би се извршила класификација отказа као што је приказано на *Слици 56*. При томе, модел достиже стопу детекције отказа магнетног диска од 16.67% уз прецизност од 100% користећи само најзначајнији атрибута *GList1*.



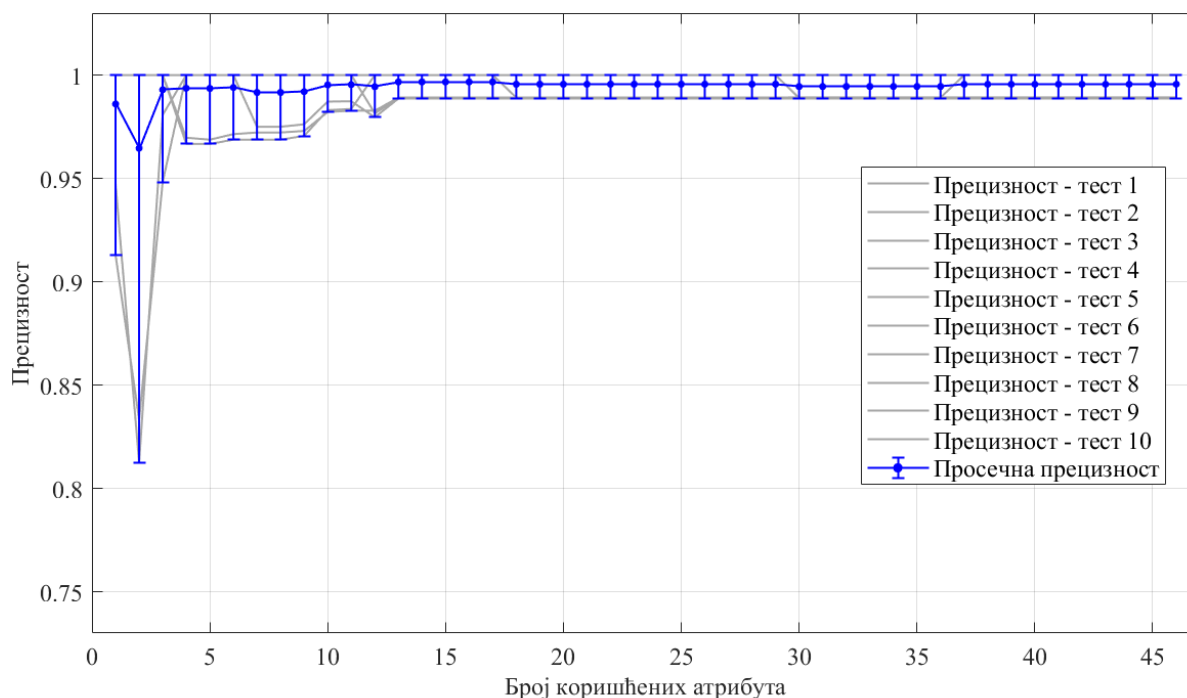
Слика 56. Примена оптималне границе одлучивања на тест скупу за атрибут *GList1*

Следећи корак укључује додавање наредног атрибута из рангиране листе и понављање поступка одређивања оптималне границе одлуке, као и процене предиктивних перформанси на скупу за тестирања. Обуком модела на скупу за валидацију одређена је оптимална граница одлучивања за сваки додати атрибут, а резултати су приказани на *Слици 57*.



Слика 57. Граница одлучивања код модела са *SVM-RFE* рангираном листом атрибута

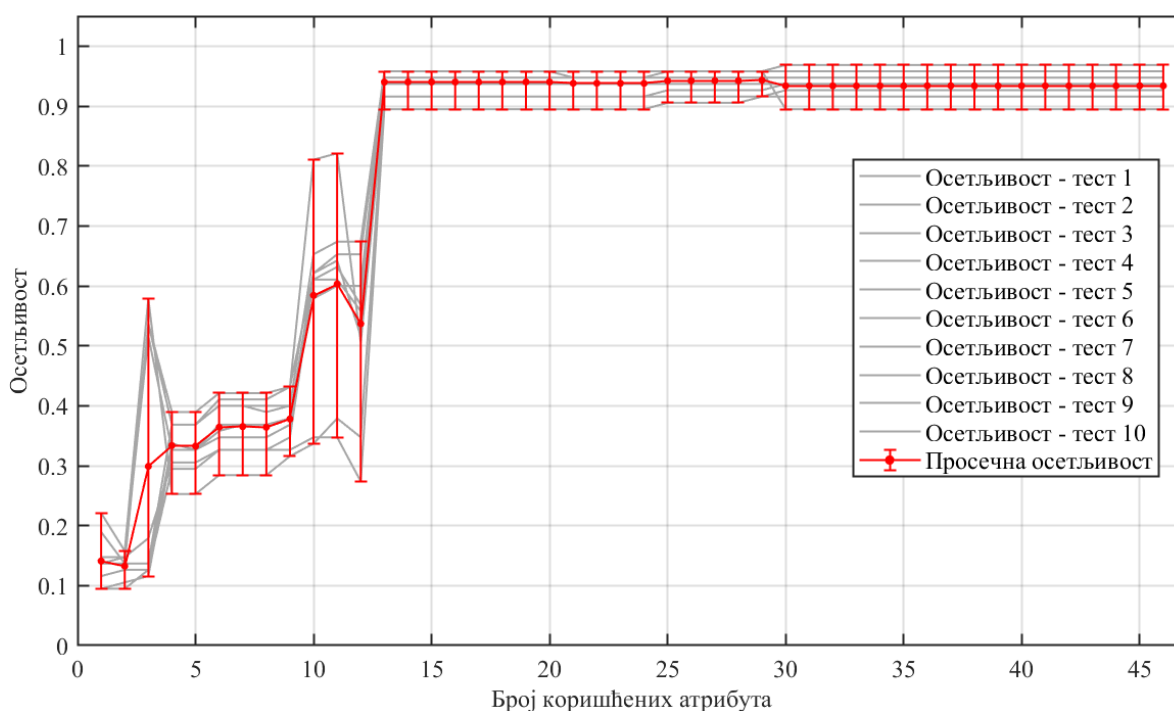
Поступак обуке модела завршава се када се достигне жељени ниво перформанси или када се сви атрибути из рангиране листе укључе у модел. У овом случају перформансе модела су одређене за свих 46 атрибута из рангиране листе. Како би се умањио утицај поделе скупова података на прецизност и осетљивост модела приликом предвиђања отказа, извршено је десет експеримената са насумично формираним скуповима података. На основу десет тестова, модел је остварио прецизност приказану на *Слици 58*.



Слика 58. Прецизност модела у зависности од броја коришћених атрибута

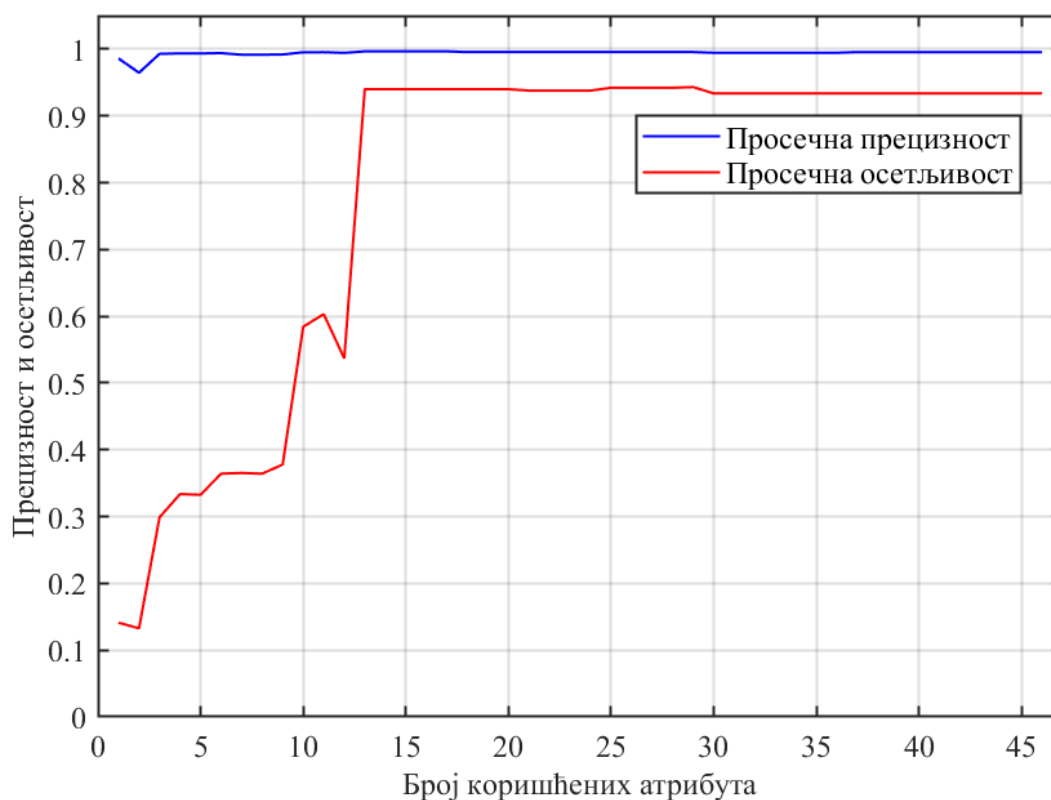
Варијације у прецизности између тестова последица су различитих подела података, због чега је одређена и просечна прецизност остварена током тестирања. На основу добијених резултата може се закључити да реализовани модел постиже високу прецизност са малим бројем одабраних атрибута, што је и основни захтев код предикције отказа магнетних дискова. Модел достиже максималну прецизност од 99.67% коришћењем 13 најзначајнијих атрибута, при чему је стопа лажних аларма 0.33%. Додавање нових атрибута не доприноси даљем побољшању прецизности модела, па њихово укључивање у модел није оправдано јер повећава рачунарску комплексност модела.

Предиктивне перформансе реализованог модела представљене су помоћу метрике осетљивости, а резултати добијени извршавањем десет тестова приказани су на *Слици 59*. Осетљивост модела се постепено повећавала додавањем нових атрибута, док су варијације у резултатима последица различите поделе скупова података између тестова. Просечна осетљивост одређена је као средња вредност резултата тестова и она достиже максималну вредност од 94% коришћењем 13 најзначајнијих атрибута. Укључивање нових атрибута није оправдано јер значајније не доприноси повећању осетљивости модела.



Слика 59. Осетљивост модела у зависности од броја коришћених атрибута

Прецизност и осетљивост модела међусобно су упоређени у односу на број коришћених атрибута. Као што је приказано на *Слици 60*, обе метрике достижу своје максималне вредности при коришћењу 13 атрибута. Листа 13 најзначајнијих атрибута, који омогућавају постизање максималне прецизности и осетљивости приказане су у *Табели 12*.



Слика 60. Прецизност и осетљивост модела у зависности од броја коришћених атрибута

Табела 12. Листа тринаест најзначајнијих атрибута за реализовани модел

| Ранг атрибута | Назив атрибута | Утицај на прецизност | Утицај на осетљивост |
|---------------|----------------|----------------------|----------------------|
| 1.            | GList1         | 0%                   | + 14.11%             |
| 2.            | GList3         | - 2.15%              | - 0.84%              |
| 3.            | Servo10        | + 2.83%              | + 16.36%             |
| 4.            | ReadError18    | + 0.07%              | + 3.47%              |
| 5.            | Servo9         | - 0.01%              | - 0.105%             |
| 6.            | FlyHeight1     | + 0.05%              | + 3.16%              |
| 7.            | FlyHeight2     | - 0.24%              | + 0.1%               |
| 8.            | Temp4          | 0%                   | - 0.1%               |
| 9.            | FlyHeight11    | + 0.04%              | + 1.37%              |
| 10.           | Servo5         | + 0.33%              | + 20.63%             |
| 11.           | FlyHeight10    | + 0.01%              | + 1.89%              |
| 12.           | ReadError19    | - 0.10%              | - 6.63%              |
| 13.           | Servo6         | + 0.23%              | + 40.32%             |

Остварене предиктивне перформансе модела са 13 укључених атрибута приказане су за свих десет тестова у Табели 13. Реализовани модел је на седам тестова постигао максималну очекивану прецизност од 100%, уз просечну осетљивост од 93.98%. На преостала три теста остварена је просечна прецизност 98.89%, осетљивост од 94.04%, а стопа лажних аларма 1.11%. У просеку модел је постигао прецизност од 99.67%, осетљивост од 94% и стопу лажних аларма 0.33%.

Табела 13. Резултати теста за SVM-RFE рангиране атрибуте

| Број теста    | Осетљивост    | Прецизност    | FAR          |
|---------------|---------------|---------------|--------------|
| 1.            | 95.79%        | 100%          | 0%           |
| 2.            | 95.79%        | 100%          | 0%           |
| 3.            | 91.58%        | 100%          | 0%           |
| 4.            | 89.47%        | 100%          | 0%           |
| 5.            | 94.74%        | 98.90%        | 1.10%        |
| 6.            | 91.58%        | 98.86%        | 1.14%        |
| 7.            | 95.79%        | 100%          | 0%%          |
| 8.            | 93.68%        | 100%          | 0%           |
| 9.            | 95.79%        | 98.91%        | 1.09%        |
| 10.           | 95.79%        | 100%          | 0%           |
| <b>Просек</b> | <b>94.00%</b> | <b>99.67%</b> | <b>0.33%</b> |

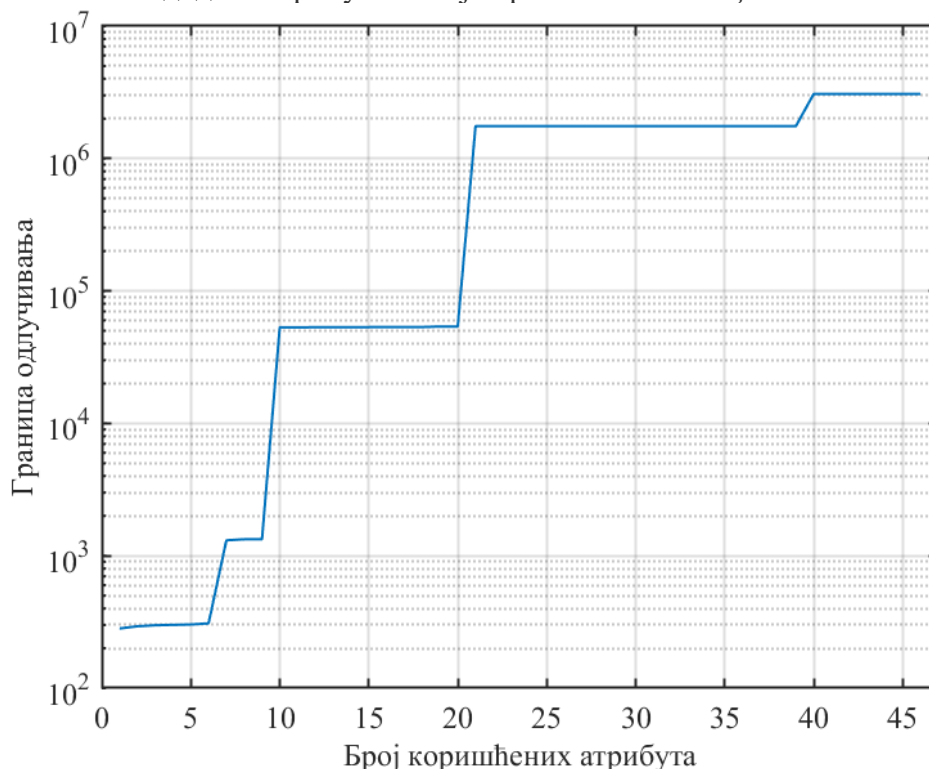
Анализом утицаја атрибута на прецизност и осетљивост уочено је да неки атрибути који су високо ранжирани од стране SVM-RFE, негативно утицали на перформансе модела за предикцију отказа. С обзиром да су атрибути *FlyHeight2*, *Servo9*, *ReadError18*, *ReadError19*, негативно утицали на осетљивости модела, премештени су на крај листе ранжираних атрибута, као што је приказано у Табели 14.

Табела 14. Листа ранжираних атрибута SVM-RFE након оптимизације [120]

| Ранг атрибута | Назив атрибута | Ранг атрибута | Назив атрибута |
|---------------|----------------|---------------|----------------|
| 1.            | GList1         | 24.           | ReadError2     |
| 2.            | GList3         | 25.           | ReadError5     |
| 3.            | Servo10        | 26.           | ReadError9     |
| 4.            | FlyHeight1     | 27.           | FlyHeight7     |
| 5.            | Temp4          | 28.           | FlyHeight4     |
| 6.            | FlyHeight11    | 29.           | Servo7         |
| 7.            | Servo5         | 30.           | Temp3          |
| 8.            | FlyHeight10    | 31.           | Servo3         |
| 9.            | Servo6         | 32.           | ReadError1     |
| 10.           | FlyHeight6     | 33.           | ReadError7     |
| 11.           | ReadError4     | 34.           | ReadError6     |
| 12.           | Servo1         | 35.           | ReadError11    |
| 13.           | FlyHeight12    | 36.           | GList2         |
| 14.           | Temp1          | 37.           | ReadError12    |
| 15.           | FlyHeight13    | 38.           | Servo8         |
| 16.           | Reads          | 39.           | ReadError10    |
| 17.           | FlyHeight3     | 40.           | ReadError8     |
| 18.           | FlyHeight5     | 41.           | FlyHeight2     |
| 19.           | CSS            | 42.           | ReadError19    |
| 20.           | PList          | 43.           | ReadError18    |
| 21.           | Servo2         | 44.           | Servo9         |
| 22.           | Writes         | 45.           | FlyHeight9     |
| 23.           | ReadError3     | 46.           | FlyHeight8     |

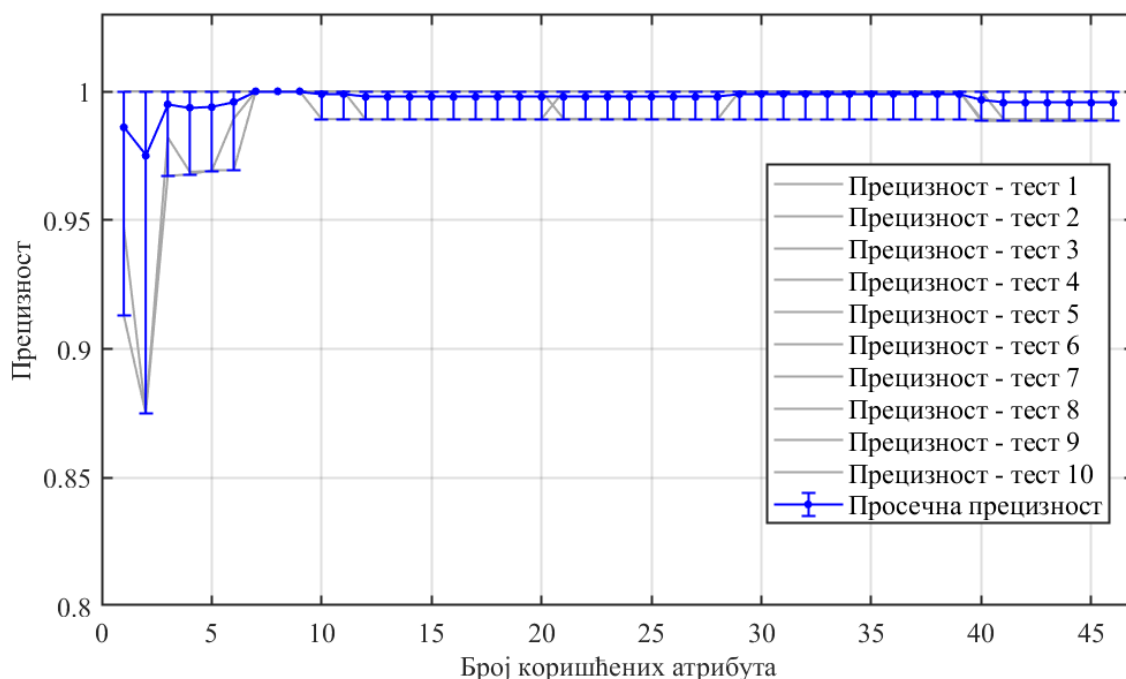
Након оптимизације листе атрибута ранжираних од стране SVM-RFE, тестови су поновљени итеративним додавањем једног по једног атрибута са оптимизоване листе

атрибута. Обуком модела на валидационом скупу дефинисана је оптимална граница одлучивања за сваки додати атрибут и она је приказана на *Слици 61*.



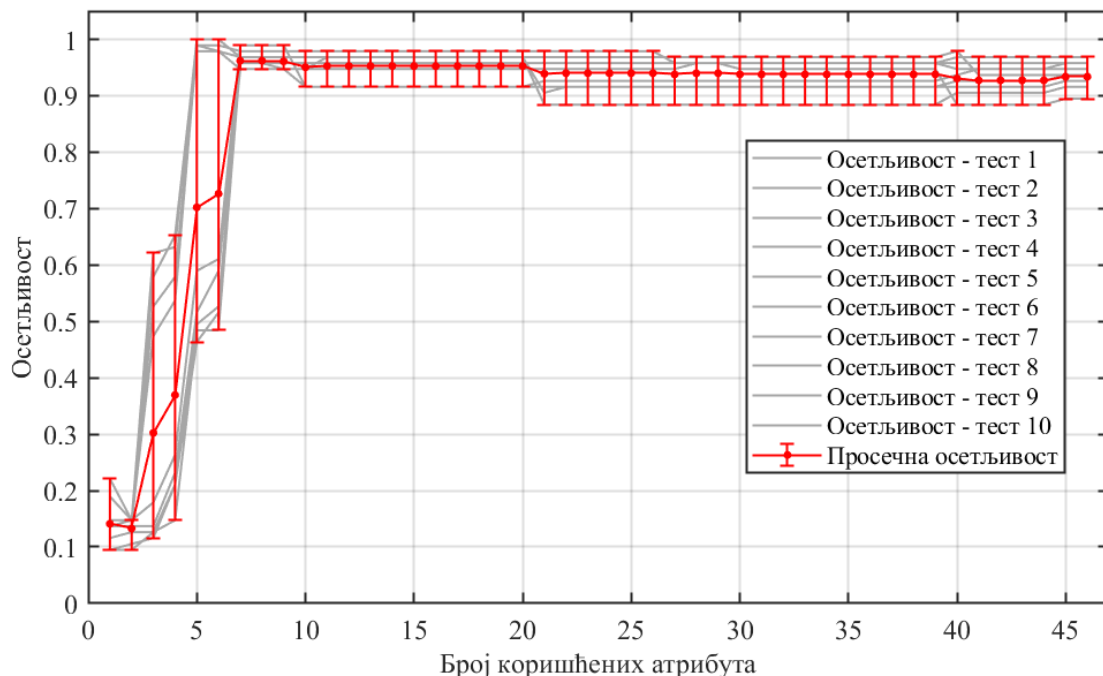
Слика 61. Граница одлучивања код модела са оптимизованом листом атрибута

Прецизност модела са оптимизованом листом атрибута тестирана је у десет независних тестова, уз итеративно додавање једног по једног атрибута. Модел је у свих десет тестова достигао прецизност од 100% након додавања седам атрибута, као што је приказано на *Слици 62*. Додавање осталих атрибута је даље негативно утицало на прецизност, па њихово укључивање у модел није било оправдано.



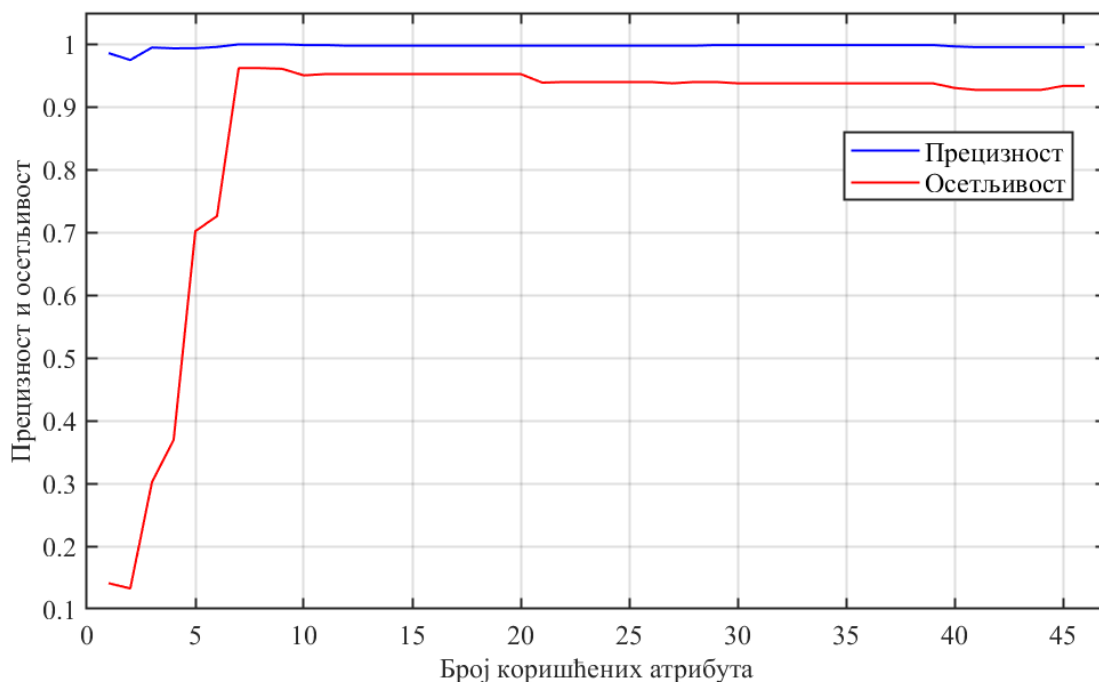
Слика 62. Прецизност оптимизованог модела у зависности од броја атрибута

Осетљивост модела након оптимизације листе атрибута достигла је максималну вредност од 96.11% након додавања седам атрибута (Слика 63). Овај максимум представља средњу вредност осетљивости свих десет тестова. Додавањем нових атрибута долази до благог пада осетљивости, тако да њихово укључивање у модел није потребно.



Слика 63. Осетљивост оптимизованог модела у зависности од броја атрибута

Прецизност и осетљивост модела одређене су као средње вредности свих тестова и приказане су на Сlici 64, на којој се уочава да модел достиже максималне предиктивне перформансе са седам најзначајнијих атрибута, при чему не детектује лажне аларме.



Слика 64. Прецизност и осетљивост оптимизованог модела у зависности од броја коришћених атрибута

Осетљивост и прецизност оптимизованог модела приказане су у Табели 15 за свих десет тестова, уз коришћење седам најзначајнијих атрибута. Просечна осетљивост оптимизованог модела износи 96.11%, док је прецизност 100%.

Табела 15. Резултати теста након оптимизације SVM-RFE листе атрибута

| Број теста    | Осетљивост    | Прецизност  | FAR       |
|---------------|---------------|-------------|-----------|
| 1.            | 94.74%        | 100%        | 0%        |
| 2.            | 98.94%        | 100%        | 0%        |
| 3.            | 96.84%        | 100%        | 0%        |
| 4.            | 94.74%        | 100%        | 0%        |
| 5.            | 96.84%        | 100%        | 0%        |
| 6.            | 95.79%        | 100%        | 0%        |
| 7.            | 94.74%        | 100%        | 0%        |
| 8.            | 95.79%        | 100%        | 0%        |
| 9.            | 97.89%        | 100%        | 0%        |
| 10.           | 94.79%        | 100%        | 0%        |
| <b>Просек</b> | <b>96.11%</b> | <b>100%</b> | <b>0%</b> |

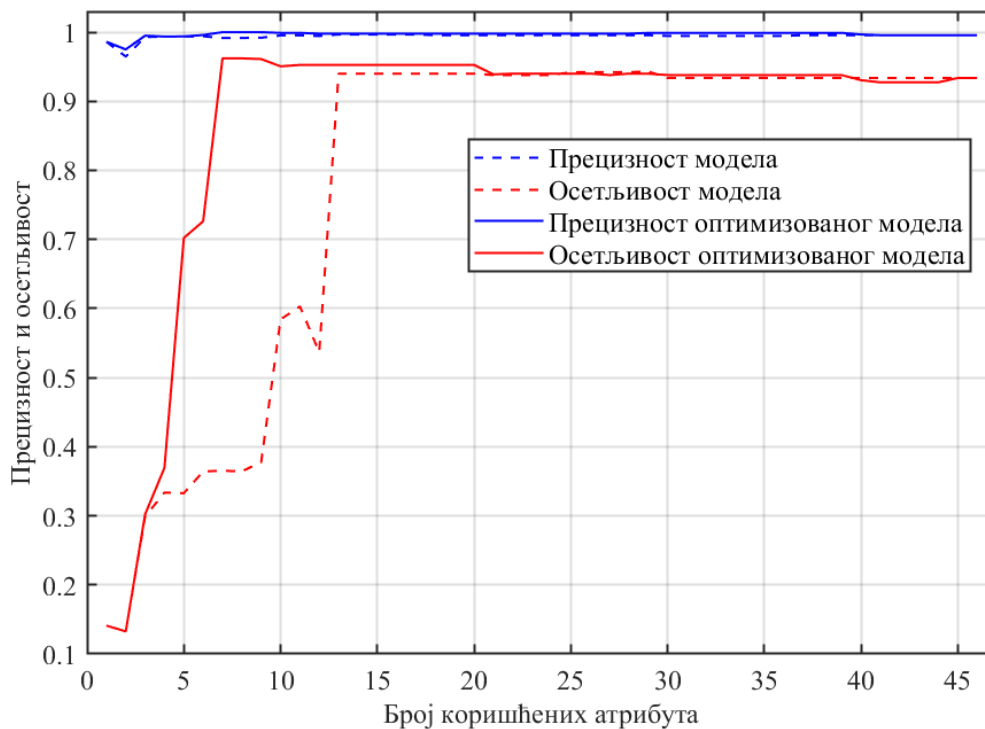
Листа атрибута са којом је оптимизовани модел постигао максималне предиктивне перформансе приказана је у Табели 16.

Табела 16. Листа седам најзначајнијих атрибута за оптимизовани модел

| Ранг атрибута | Назив атрибута | Утицај на прецизност | Утицај на осетљивост |
|---------------|----------------|----------------------|----------------------|
| 1.            | GList1         | 0%                   | + 14.11%             |
| 2.            | GList3         | - 1.1%               | - 0.84%              |
| 3.            | Servo10        | + 1.99%              | + 16.95%             |
| 4.            | FlyHeight1     | - 0.13%              | + 6.74%              |
| 5.            | Temp4          | + 0.02%              | + 33.26%             |
| 6.            | FlyHeight11    | + 0.2%               | + 2.42%              |
| 7.            | Servo5         | + 0.41%              | + 23.58%             |

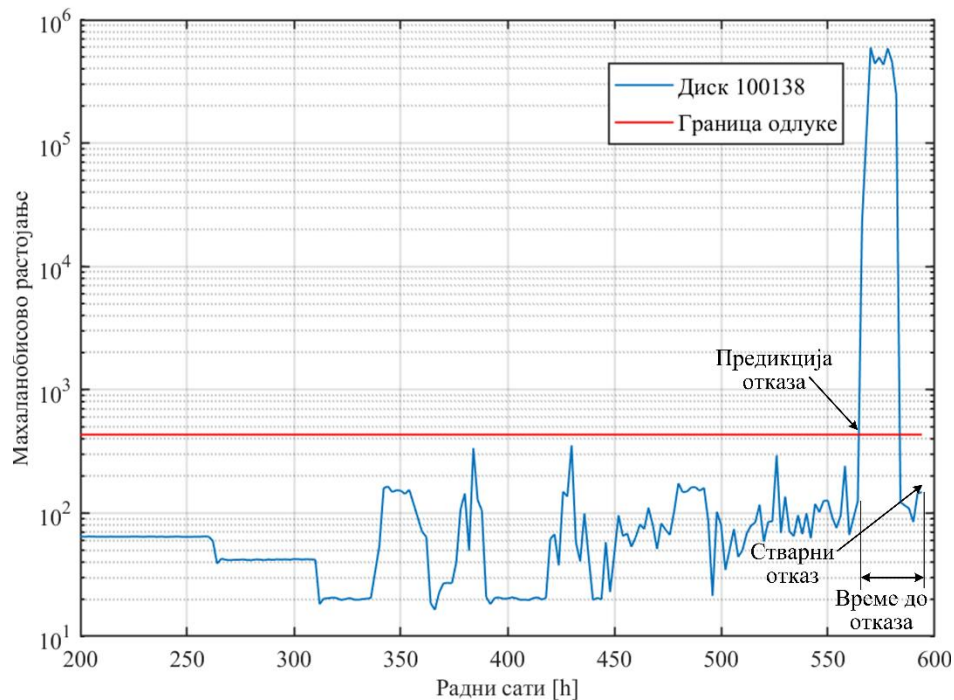
Предиктивне перформансе представљених модела међусобно су упоређене на Слици 65. Резултати показују да је оптимизација ранг листе користи двоструко мањи број атрибута потребних за достизање максималне осетљивости. Поред мањег броја атрибута оптимизовани модел је постигао осетљивост већу за 2.11% у односу на референтни модел, а при томе је у потпуности елиминисао лажне детекције отказа.

Поред тачне класификације отказа магнетних дискова, модел је у стању да предвиђа отказ унапред, пре него што се они заиста и догоде. Предвиђање отказа унапред омогућава примену проактивног одржавања рачунарског система што се може спровести правовременом израдом резервне копије података. Све док је вредност Махаланобисовог растојања нижа од границе одлуке, инстанца диска се сматра као исправна. Када вредност Махаланобисовог растојања пређе дефинисану границу одлуке, активира се аларм који сигнализира предстојећи отказ диска.



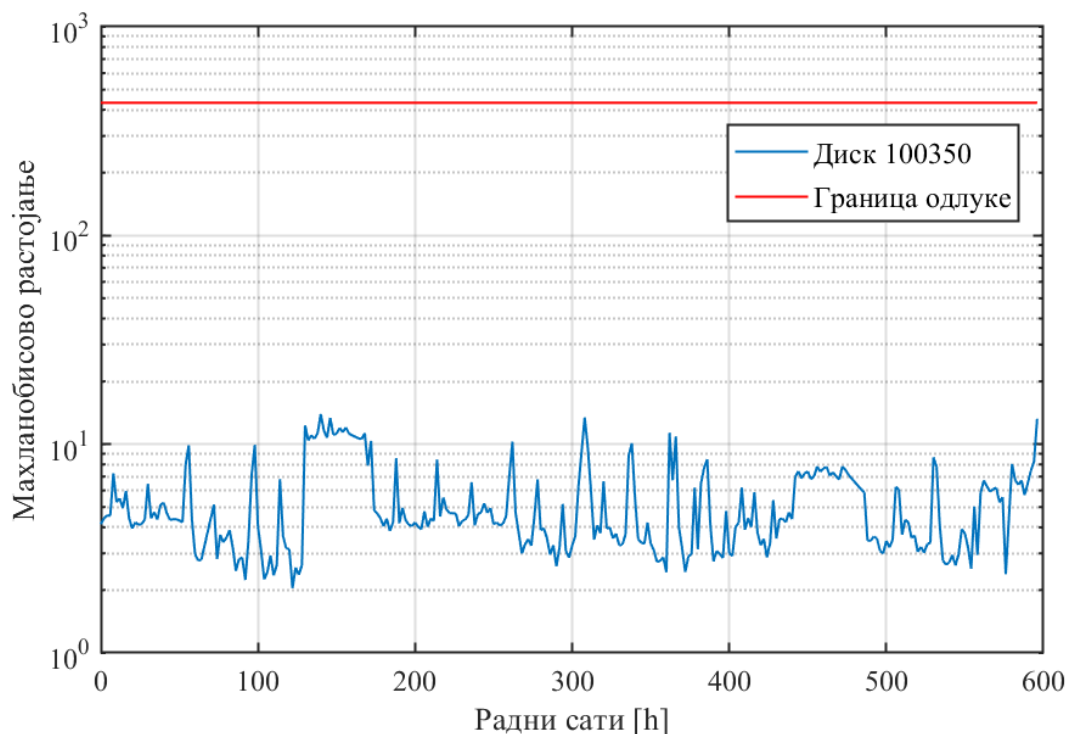
Слика 65. Прецизност и осетљивост пре и после оптимизације модела

На Слици 66. приказана је временска серија Махаланобисовог растојања за магнетни диск са серијским бројем 100138. У првом интервалу (300h) магнетни диск ради у оквиру референтних вредности, а његово Махаланобисово растојање се не мења значајно током времена. Након тога долази до значајног одступања у раду диска, што се огледа у брзим променама Махаланобисовог растојања. У 564 сату долази до прекорачења границе одлуке, чиме се активира аларм за предстојећи отказ. Диск наставља са непоузданим радом и до отказа долази у 594 сату. Алгоритам је успешно предвидео отказ 30 часова пре његовог настанка, што оставља довољно времена кориснику да предузме превентивне активности.



Слика 66. Детекција отказа за магнетни диск са серијским бројем 100138

Рад исправног диска са серијским бројем 100350 приказан је на Слици 67, на којој се уочава да је вредност Махаланобисовог растојања током посматраног интервала далеко нижа од границе одлуке, што значи да неће доћи до активације аларма.



Слика 67. Махаланобисово растојање за исправан диск са серијским бројем 100350

### 8.2 Поређење добијених резултата са постојећим истраживањима

Научни допринос ове дисертације представљен је поређењем остварених резултата са резултатима досадашњих истраживања у овој области. Најрелевантније поређење резултата остварује се када се алгоритми за предикцију отказа магнетних дискова упореде на истом скупу података. У досадашњим истраживањима најчешће је коришћен јавно доступни CMRR скуп података.

У Табели 17 приказан је преглед ранијих резултата предикције отказа магнетних дискова, који су тестирани на CMRR скупу података и упоређени са оствареним резултатима у овој дисертацији. Алгоритми су међусобно упоређени према броју атрибута CMRR скупа који су коришћени у процесу обуке. Генерално, алгоритми који користе мањи број атрибута су рачунарски мање комплексни, брже се извршавају и заузимају мање рачунарске меморије. Такође, приказано је и поређење предиктивних перформанси помоћу следећих метрика: прецизност, стопа аларма, осетљивост и F0.5 мера.

Аутори рада [14] први су употребили алгоритме машинског учења за предикцију отказа на CMRR скупу података. У свом истраживању представили су нови алгоритам заснован на наивном Бајесовом (енг. *Naïve Bayes*) класификатору са вишеструким инстанцама, mi-NB (енг. *multiinstance Naïve Bayes*). Овај алгоритам упоређен је са SVM алгоритмом на подскупу са 25 значајних атрибута. mi-NB алгоритам постигао је високу прецизност од 99% са стопом аларма од 1%, при чему је детектовао 34.5% отказа. SVM алгоритам постигао је знатно боље перформансе од предложеног mi-NB, остваривши 100% прецизности, без лажних аларма уз стопу детекције 50.6%.

Табела 17. Преглед резултата предикције отказа магнетних дискова на CMRR скупу

| Алгоритам                      | Број атрибута | Прецизност TNR | Стопа аларма FAR | Осетљивост FDR | F0.5 мера     |
|--------------------------------|---------------|----------------|------------------|----------------|---------------|
| mi-NB [14]                     | 25            | 99%            | 1%               | 34.5%          | 72.06%        |
| SVM [14]                       | 25            | 100%           | 0%               | 50.6%          | 83.66%        |
| Rank-sum [14]                  | 4             | 99.3%          | 0.7%             | 52.8%          | 84.42%        |
| HMM [18]                       | 2             | 100%           | 0%               | 52%            | 84.42%        |
| HMMD [47]                      | 8             | 99.44%         | 0.56%            | 63%            | 89.13%        |
| FSMD [19]                      | 31            | 100%           | 0%               | 67%            | 91.03%        |
| TSP [20]                       | 47            | 100%           | 0%               | 68%            | 91.40%        |
| GMM [124]                      | 8             | 100%           | 0%               | 80.59%         | 95.40%        |
| Fusion AD [71]                 | /             | 97.1%          | 2.9%             | 90.58%         | 95.72%        |
| GMFD [23]                      | 14            | 100%           | 0%               | 92.21%         | 98.34%        |
| <b>Реализовани модел [120]</b> | <b>7</b>      | <b>100%</b>    | <b>0%</b>        | <b>96.11%</b>  | <b>99.20%</b> |

Такође, SVM алгоритам је постигао вредност F0.5 мере од 83.66% насупрот вредности од 72.06% коју је остварио mi-NB алгоритам. Аутори су применили статистички Rank-sum тест користећи четири најзначајнија атрибута и постигли сличне резултате са SVM алгоритмом, приказане у Табели 17. Сличан ниво перформанси постигнут је од стране аутора рада [18] који су користили алгоритам заснован на скривеним Марковљевим моделима (енг. *Hidden Markov Models*) са два атрибута и остварили стопу детекције од 52% без лажних аларма. У раду [47] аутори су користили метод заснован на мерењу Махаланобисовог растојања са осам најутицајнијих атрибута идентификованих FMMEA методом. Алгоритам је постигао стопу детекције од 63% на CMRR скупу података са стопом лажних аларма од 0.56%, при чему је остварена вредност F0.5 мере од 89.13%. У истраживању [19] иста група аутора је представила модел који је користио 31 атрибут, са којим је остварио осетљивост од 67%, без детекције лажних аларма и вредност F0.5 мере од 91.03%. TSP алгоритам двостепене предикције отказа [20] користио је Вох-Сох трансформацију за нормализацију вредности 47 атрибута и постигао осетљивост од 68%, без лажних аларма и остварио вредност F0.5 мере од 91.40%. Аутори рада [71] представили су модел који је са стопом лажних аларма од 2.9% остварио стопу детекције 90.58% и вредност F0.5 мере од 95.72%. Коришћењем GMM модела [124] на CMRR скупу података са осам одабраних атрибута постигли су осетљивост 80.59% са прецизношћу од 100% и вредност F0.5 мере од 95.40%. Исти аутори су у раду GMFD [23] унапредили претходни модел који са 14 атрибута остварује осетљивост од 92.21%, без присуства лажних аларма, док вредност F0.5 мере достиже 98.34%.

Модел представљен у овој дисертацији постигао је просечну стопу детекције отказа од 96.11% са 0% лажних аларма, при чему је остварена вредност F0.5 мере 99.20%. Анализом резултати представљених у Табели 17 може се закључити да је реализовани модел овог истраживања постигао најбоље предиктивне перформансе од свих представљених алгоритама на CMRR скупу података. Штавише, овај модел је захтевао само седам атрибута, што је двоструко мањи број атрибута које захтева GMFD модел и само четвртина атрибута које захтевају TSP и FSMD модели. Са смањеним бројем

атрибута, представљени модел је био једноставнији за практичну имплементацију у поређењу са другим моделима. Резултати представљени у Табели 18 показују да је реализовани модел у стању да предвиди 86.53% отказа најмање 12 сата пре појаве стварног отказа магнетног диска, док се 80.53% отказа може предвидети један дан унапред. Три четвртине отказа се могу унапред предвидети у року од 36 сати, а две трећине отказа се могу предвидети два дана унапред. Реализовани алгоритам представљен у овој дисертацији у стању је да предвиди отказ у просеку 140 сати, односно око шест дана пре појаве стварног отказа магнетног диска. Добијени резултати времена предвиђања отказа су упоређени у Табели 18 са резултатима постојећих алгоритама за предвиђање отказа. Представљени резултати су упоредиви са алгоритмом GMFD, док остварују знатно боље време предикције отказа у односу на TSP и FSMД алгоритме.

Табела 18. Поређење времена предикције отказа магнетних дискова [120]

| Модел<br>Време<br>до отказа | Стопа детекције      |        |        |        |
|-----------------------------|----------------------|--------|--------|--------|
|                             | Реализовани<br>модел | GMFD   | FSMD   | TSP    |
| Није детектован отказ       | 3.89%                | 7.79%  | 32.98% | 31.58% |
| $\geq 0$                    | 96.11%               | 92.21% | 67.02% | 68.42% |
| $\geq 12$ h                 | 86.53%               | 86.39% | 60.21% | 50.00% |
| $\geq 24$ h                 | 80.53%               | 81.88% | 52.66% | 46.05% |
| $\geq 36$ h                 | 74.95%               | 76.55% | 45.12% | 43.42% |
| $\geq 48$ h                 | 67.89%               | 73.19% | 37.58% | 32.89% |

### 8.3 Приказ резултата предикције отказа полупроводничких SSD дискова

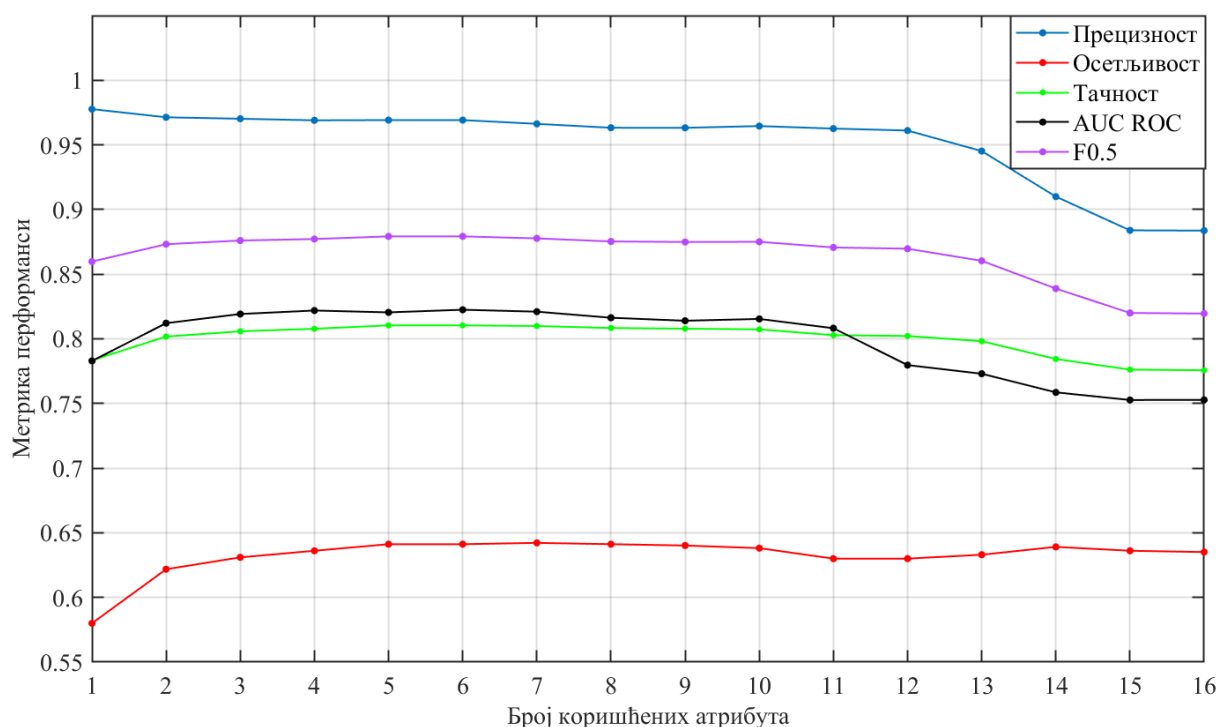
Предложени модел за предикцију отказа полупроводничких SSD дискова тестиран је на јавно доступном *Alibaba* SSD скупу података који се састоји од SMART атрибута прикупљених у њиховом центру података за период од две године. Током овог периода прикупљени су подаци са пола милиона SSD дискова који су произведени од стране три произвођача, означених као MA, MB, MC. Подаци у центру података прикупљани су свакодневно у виду SMART атрибута за сваки оперативни SSD диск, при чему је формирана и листа проблема у којој су евидентирана неуобичајена одступања дискова од нормалних вредност или откази дискова. За потребе истраживања издвојени су подаци за диск са ознаком MC1 који је најбројнији диск у наведеном скупу података. Овај модел диска има капацитет од 1.92TB и реализован је коришћењем 3D TLC NAND технологије. Током посматраног периода у трајању од годину дана забележено је 1962 отказа од укупног броја од 150972 коришћених дискова. Како би се обезбедило да скуп података буде избалансиран, извршено је насумично узорковање већинске класе исправних дискова и скуп података је редукован на 4308 исправних дискова. У наведеном скупу доступно је 21 SMART атрибута, при чему је пет атрибута уклоњено јер је у директној корелацији са осталим атрибутима у скупу података. Скуп података је подељен у три подскупа у којима је заступљеност исправних и отказалих дискова приказана у Табели 19.

Табела 19. Партиционисање Alibaba SSD скупа података

| Подскуп података           | Подскуп за тренирање | Подскуп за валидацију | Подскуп за тестирање | Укупан број дискова |
|----------------------------|----------------------|-----------------------|----------------------|---------------------|
| Исправни дискови           | 2346                 | 981                   | 981                  | <b>4308</b>         |
| Отказали дискови           | -                    | 981                   | 981                  | <b>1962</b>         |
| <b>Укупан број дискова</b> | <b>2346</b>          | <b>1962</b>           | <b>1962</b>          | <b>6270</b>         |

Обуком модела за предикцију отказа SSD дискова, атрибути су додавани у подскуп атрибута према њиховом утицају на перформансе модела. Након додавања последњег атрибута, завршава се процес обуке модела, а предиктивне перформансе модела приказане су на *Слици 68*.

Модел је постигао максимум перформанси користећи шест најзначајнијих атрибута, што доприноси смањењу његове рачунарске комплексности. Применом модела за предикцију отказа на овом скупу података, постигнута је стопа детекције од 64%, при чему је прецизност износила 96.7%, а вредност F0.5 мере 88%. Генерализоване класификационе перформансе модела у виду површине испод ROC криве, износе 0.8225, што се сматра задовољавајуће добрим нивоом класификационих перформанси.



Слика 68. Предиктивне перформансе модела у зависности од броја коришћених атрибута MC1 модела SSD диска [123]

Добијени резултати су упоређени у Табели 20 са претходно реализованим моделима које су развили други истраживачи. Поред предиктивних перформанси, модели су међусобно упоређени према броју коришћених атрибута. Једна од предности реализованог модела је што користи знатно мањи број атрибута у односу на остале моделе, што га чини погоднијим за практичну имплементацију.

Табела 20. Поређење алгоритама на *Alibaba SSD* скупу података [123]

| Модел                                         | Број атрибута | Тачност    | Прецизност   | Осетљивост | F0.5 мера  |
|-----------------------------------------------|---------------|------------|--------------|------------|------------|
| Логистичка регресија [125]                    | 10            | 30.9%      | 40.8%        | 24.0%      | 35.8%      |
| Случајна шума [125]                           | 10            | 46.3%      | 48.5%        | 41.3%      | 46.9%      |
| LSTM [125]                                    | 5             | 60.9%      | 66.1%        | 61.6%      | 65.1%      |
| Асемблирано учење [126]                       | 7             | 98%*       | 100%         | 21%        | 57%        |
| FFN [127]                                     | 13            | 58%        | 47%          | 69%        | 50%        |
| WEFR [128]                                    | 168           | 47%*       | 49%          | 18%        | 36%        |
| MVTRF Alibaba MB1[129]                        | 16            | 88%*       | 89%          | 76%        | 86%        |
| MSFRD Alibaba MB2 [130]                       | 14            | 86%*       | 87 %         | 27%        | 60%        |
| <b>Реализовани модел за SSD дискове [123]</b> | <b>6</b>      | <b>81%</b> | <b>96.7%</b> | <b>64%</b> | <b>88%</b> |

\*- пристрасност због високог дисбаланса класа

У поређењу са алгоритмима за предвиђање отказа представљеним у раду [125] који су такође тестирани на истом скупу података за MC1 модел диска, остварени резултати су показали значајно побољшање у погледу тачности, прецизности и осетљивости. Представљени модел постиже боље резултате и у поређењу са LSTM моделом, који је рачунарски сложенији. Аутори у раду [126] су такође користили *Alibaba SSD* скуп података у модификованом моделу асемблираног учења и постигли су стопу осетљивости од 21% са 100% прецизношћу. Исти скуп података коришћен је у истраживању [127] где је FFN (енг. *Feed Forward Neural Network*) мрежа постигла стопу осетљивости од 69% са прецизношћу од 47%, што је ниже од перформанси представљеног модела у овом истраживању. Аутори рада [128] су представили су WEFR (енг. *Wear-out-updating Ensemble Feature Ranking*) модел који је тестиран на MC1 диску и постигли су осетљивост од 18%, док је прецизност износила 49%. Аутори рада [129] развили су алгоритам случајне шуме MVTRF (енг. *Multi-View and multi-Task Random Forest*), који су тестирали на *Alibaba* скупу података, али на моделу диска MB1, где су постигли прецизност од 89% и осетљивост од 76%. Истраживачи у раду [130] су представили MSFRD (енг. *Mutation Similarity based Failure Rating and Diagnosis*) модел који је тестиран на MB2 моделу диска који је постигао прецизност од 87%, уз осетљивост од 27%, што је знатно мање у поређењу са резултатима оствареним у овом истраживању. Резултати за метрику тачности, означени у Табели 20. са звездом (\*), постигнути су са високо неуравнотеженим скуповима података, што за последицу има високу пристрасност услед утицаја већинске негативне класе. У предложеном моделу извршено је балансирање скупа података коришћењем случајног узорковања већинске класе, тако да је остварена метрика тачности поузданија. Резултат F0.5 мере од 88% је међу највишим, што показује добар баланс између прецизности и осетљивости, тако да је реализовани модел погодан за практичну примену у предикцији отказа SSD дискова.

## 9 Закључна разматрања

Магнетни дискови представљају доминантан медијум на којима се чува готово целокупна количина дигиталних података које је створило човечанство. С обзиром да се количина ових података из године у годину убрзано повећава, магнетни дискови ће и у наредној деценији чинити окосницу за масовно складиштење у центрима података. Захваљујући иновативним технологијама, густина записа на магнетним дисковима се константно повећава, тако да нови модели магнетних дискова успевају да одговоре на тржишне захтеве у погледу капацитета и цене. Свој доминантни статус на тржишту, магнетни дискови задржавају високом густином записа и ниском ценом по ускладиштену биту, која је неколико пута нижа у односу на конкурентну технологију полупроводничких SSD дискова.

Савремена технологија магнетних дискова успела је да нађе решење за наизглед међусобно супростављене захтеве, масовном производњом високо прецизних машинских склопова, који користе микрометарску резолуцију за складиштење података. Магнетни дискови се производе као високо поуздани уређаји код којих се ултра-глатке плоче константно ротирају на растојању од неколико микрометара од уписно-читајућих глава. Упркос томе, магнетни диск је једина електромеханичка компонента рачунарског система и његови високо прецизни механички склопови су временом подложни хабању, које негативно утиче на радне перформансе. Када магнетни диск није у стању да испуни постављене функционалне захтеве у оквиру специфицираних перформанси, проглашава се његов отказ. Стопа отказа магнетних дискова на годишњем нивоу креће се од 0.3% до 3%.

За разлику од других компоненти рачунарског система, отказ диска најчешће доводи и до губитка ускладиштених података и врло често економска штета проузрокована губитком података вишеструко превазилази цену самог диска. Како би се у некој мери спречила штета узрокована губитком података, која настаје услед отказа диска развијени су реактивни и проактивни приступи у сервисирању последица отказа диска. Реактивни приступи започињу примену мера у тренутку када наступи отказ магнетног диска и они се ослањају на постајање уграђене редундантности података, који остају доступни кориснику, чак и ако се део података изгуби отказом магнетног диска. Овакав приступ захтева складиштење података у RAID низовима који укључују већи број дискова. Проактивни приступ започиње примену мера знатно раније од тренутка када наступи отказ диска, што се постиже праћењем низа радних параметара изражених у виду SMART атрибута. Када вредности неког од SMART атрибута одступе од граница дефинисаних од стране произвођача, корисник се упозорава о могућем отказу магнетног диска, како би на време креирао резервну копију података и спречио њихов губитак. Како би избегли често генерисање лажних аларма који би резултирали непотребним рекламацијама, произвођачи дискова постављају границе за SMART атрибуте на високе вредности, па се у пракси детектује испод 10% стварних отказа, док у 90% случајева долази до отказа диска без претходног упозорења.

Из наведеног разлога област истраживања предикције отказа магнетних дискова актуелна је дужи низ година. Прва истраживања била су заснована на статистичкој анализи појединачних SMART атрибута и њиховог утицаја на појаву отказа магнетног диска. Развој области машинског учења и примена његових алгоритама допринела је развоју мултиваријантних модела који су знатно повећали стопу предикције отказа

магнетних дискова. Свеобухватном анализом сагледана су досадашња истраживања везана за ову проблематику на основу које је уочен простор за даља истраживања, која би побољшала стопу предикције отказа магнетних дискова у односу на резултате претходних истраживања. Област истраживања предикције отказа дискова и даље је актуелна, а томе доприноси све већа доступност отворених скупова података који потичу из центара података, што није био случај у претходном периоду.

У овој докторској дисертацији представљен је модел за предикцију отказа магнетних дискова заснован на детекцији аномалија. С обзиром да су магнетни дискови високо поуздане рачунарске компоненте откази дискова су ретке појаве које се у овом истраживању посматрају као аномалије, односно одступања од нормалних радних параметара. Развојем модела за детекцију аномалија заснованог на мерењу удаљености, омогућено је да се препознају откази дискова на основу мерења Махаланобисовог растојања у односу на центар дистрибуције који чине исправни дискови.

Кроз претходна поглавља представљене су методе коришћене за развој предложеног модела предикције отказа који је потом реализован и тестиран на CMRR скупу података и упоређени су остварени резултати са претходним истраживањима. У уводном поглављу дисертације описан је предмет истраживања у области предикције отказа магнетних дискова и дефинисан је циљ дисертације као и основне хипотезе. Такође, у наставку поглавља представљена је детаљна анализа досадашњих истраживања у области предикције отказа магнетних дискова. Структура и принцип рада магнетних дискова и анализа узрока који доводе до њихових отказа детаљно је описана у другом поглављу. У наставку је дат преглед класификације алгоритама машинског учења као и фазе тока креирања предиктивног модела. На крају поглавља приказани су алгоритми који су најчешће коришћени у досадашњим истраживањима у области предикције отказа магнетних дискова. Алгоритми машинског учења користе скупове података који су описани у трећем поглављу. Посебна пажња у овом поглављу, посвећена је CMRR скупу података који је у овој дисертацији коришћен у процесу обуке и тестирања реализованог модела.

Скупови података састоје се из низа атрибута који својим вредностима описују инстанце података, и они на различити начин доприносе предикцији отказа диска. У четвртном поглављу приказан је процес селекције атрибута којом се као резултат добија листа најзначајнијих атрибута који имају највећи утицај на предикцију отказа. За потребе овог истраживања примењен је алгоритам рекурзивне селекције атрибута који је генерисао уређену листу у којој су рангирани најзначајнији атрибути CMRR скупа података. Овај поступак омогућио је обуку модела са подскупом најзначајнијих атрибута чиме се значајно побољшале предиктивне перформансе, а самим тим се и избегава проблем претераног уклапања модела. На основу претходно наведеног, може се сматрати да су **прва и друга хипотеза** докторске дисертације потврђене.

Инстанце података са вредностима својих атрибута описују радно стање магнетног диска у вишедимензионалном простору. Како би се измерило растојање између нормалног и тренутног радног стања диска у простору, неопходно је коришћење одговарајуће метрике за мерење удаљености. С обзиром да SMART атрибути представљају вредности параметара различитих по типу и скали, примена традиционалних метрика није погодна у овој области истраживања. На основу наведеног, у петом поглављу представљена је метрика заснована на Махаланобисовом растојању која је погодна за мерење удаљености у мултиваријабилном простору. Примена ове метрике захтева да вредности атрибута прате нормалну расподелу, што је редак случај код вредности SMART атрибута. Из тог разлога било је неопходно

трансформисати податке помоћу Vox-Cox трансформације како би вредности атрибута скупа података имале приближно нормалну расподелу.

Аномалије представљају опажања која значајно одступају од очекиваних вредности. У шестом поглављу приказана је класификација аномалија и приказани су алгоритми који се користе за њихово детектовање. У наредном поглављу дисертације детаљно је представљен реализовани модел за предикцију отказа заснованог на детекцији аномалија на основу мерења Махаланобисовог растојања. Модел је представљен у виду алгоритма при чему су детаљно описане све фазе развоја модела. Поред основне намене модела, анализирана је и могућност примене модела за предикцију отказа за полупроводничке SSD дискове и представљен је реализовани модел. Такође, дат је предлог имплементације реализованог модела за предикцију отказа дискова у центрима података.

У завршном делу дисертације приказани су резултати обуке и тестирања реализованог модела на CMRR скупу података. Реализовани модел постигао је просечну стопу предикције отказа од 96.11%, са 0% лажних аларма, чиме је потврђена **трећа хипотеза** докторске дисертације. Резултати тестирања показали су да је модел остварио боље перформансе на идентичном CMRR скупу података, у поређењу са GMFD моделом који је остварио стопу детекције од 92.21% FDR, као и TSP и FSDM моделима који су постигли стопу детекције од 68.42% FDR и 67.02% FDR, респективно. Поред тога, реализовани модел користи само седам SMART атрибута, што је у двоструко мањи број атрибута које захтева GMFD модел и четири пута мање атрибута које користе TSP и FSDM модели. Мањи број атрибута олакшава обуку, имплементацију и коришћење модела, а самим тим и значајно смањује рачунарску комплексност у поређењу са другим моделима. Реализовани модел је успео да предвиди 86.53% отказа, најмање 12 сати пре наступа стварног отказа магнетног диска. Просечно време предвиђања отказа реализованог модела је шест дана унапред, чиме је потврђена и **четврта хипотеза** ове докторске дисертације.

На основу добијених резултата представљених у овој докторској дисертацији и наведеним у закључном разматрању може се констатовати да је реализовани модел испунио дефинисане циљеве истраживања и остварио значајни научни допринос постигавши висок ниво предиктивних перформанси.

Наставак истраживања биће усмерен на имплементацију предложеног модела за предикцију отказа у реалном времену у центрима података. Планиран је развој модела за NVMe полупроводничке дискове, као и примена модела у RAID низовима дискова.

## Литература

- [1] I. Zhou, I. Makhdoom, N. Shariati, M. A. Raza, R. Keshavarz, J. Lipman, M. Abolhasan, and A. Jamalipour, “Internet of Things 2.0: Concepts, Applications, and Future Directions”, *IEEE Access*, vol. 9, pp. 70961–71012, 2021, doi: 10.1109/ACCESS.2021.3078549.
- [2] A. S. Al-Ahmad and H. Kahtan, “Cloud Computing Review: Features And Issues”, in *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, Shah Alam: IEEE, Jul. 2018, pp. 1–5. doi: 10.1109/ICSCEE.2018.8538387.
- [3] G. Habib, S. Sharma, S. Ibrahim, I. Ahmad, S. Qureshi, and M. Ishfaq, “Blockchain Technology: Benefits, Challenges, Applications, and Integration of Blockchain Technology with Cloud Computing”, *Future Internet*, vol. 14, no. 11, Nov. 2022, doi: 10.3390/fi14110341.
- [4] Statista, “Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028”, Accessed: May 12, 2024. [Online]. Available: <https://www.statista.com/statistics/871513/worldwide-data-created/>
- [5] Coughlin Associates, “Digital Storage Technology Newsletter”, Accessed: May 12, 2024. [Online]. Available: <https://blocksandfiles.com/2024/05/13/coughlin-associates-hdd/>
- [6] R. Wood, “Future hard disk drive systems”, *Journal of Magnetism and Magnetic Materials*, vol. 321, no. 6, pp. 555–561, Mar. 2009, doi: 10.1016/j.jmmm.2008.07.027.
- [7] S. Đurašević, U. Pešović, B. Đorđević, and V. Luković, “Hard disk drive failure rates in datacenters”, in *Proceedings of the 18th Conference for Informatics and Information Technology*, Skopje, North Macedonia: Faculty of Computer Science and Engineering, 6-7 May 2021, pp. 43–46, ISBN: 978-608-4699-11-8.
- [8] Y. Zhu, K. C. How, H. J. Wu, and Q. Cao, “AI-based Proactive Storage Failure Management in Software-Defined Data Centres”, in *Proceedings of the 2023 6th International Conference on Information Science and Systems*, Edinburgh, United Kingdom, Aug. 2023, pp. 231–237. doi: 10.1145/3625156.3625190.
- [9] D. A. Patterson, P. Chen, G. Gibson, and R. H. Katz, “Introduction to redundant arrays of inexpensive disks (RAID)”, in *Digest of Papers. COMPCON Spring 89. Thirty-Fourth IEEE Computer Society International Conference: Intellectual Leverage*, San Francisco, CA, USA: IEEE Comput. Soc. Press, 1989, pp. 112–117. doi: 10.1109/CMPCON.1989.301912.
- [10] F. Salfner, M. Lenk, and M. Malek, “A survey of online failure prediction methods”, *ACM Computing Surveys (CSUR)*, vol. 42, no. 3, pp. 1–42, Mar. 2010, doi: 10.1145/1670679.1670680.
- [11] A. De Santo, A. Galli, M. Gravina, V. Moscato, and G. Sperli, “Deep Learning for HDD Health Assessment: An Application Based on LSTM”, *IEEE Transactions on Computers*, vol. 71, no. 1, Jan. 2022, pp. 69–80, doi: 10.1109/TC.2020.3042053.
- [12] J. Zhao, Y. He, H. Liu, J. Zhang, B. Liu, J. Zhang, W. Lv, A. Zhou, F. Jiang, J. Liu and A. Nishi, “Disk Failure Early Warning Based on the Characteristics of Customized SMART”, in *2020 19th IEEE Intersociety Conference on Thermal and*

*Thermomechanical Phenomena in Electronic Systems (ITherm)*, Orlando, FL, USA: IEEE, Jul. 2020, pp. 1282–1288. doi: 10.1109/ITherm45881.2020.9190324.

- [13] E. Pinheiro, W.D. Weber, and L. A. Barroso, “Failure Trends in a Large Disk Drive Population”, in *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, San Jose, CA, USA, 2007, doi: 10.5555/1267903.126790
- [14] J. F. Murray, J. Org, G. F. Hughes, and K. Kreutz-Delgado, “Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application”, *The Journal of Machine Learning Research*, Volume 6, 2005, pp. 783 – 816, doi: 10.5555/1046920.1088699
- [15] G. Hamerly and C. Elkan, “Bayesian approaches to failure prediction for disk drives”, in *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2001, pp. 202–209, doi: 10.5555/645530.655825
- [16] G. F. Hughes, J. F. Murray, K. Kreutz-Delgado, and C. Elkan, “Improved disk-drive failure warnings”, *IEEE Transactions on Reliability*, vol. 51, no. 3, pp. 350–357, Sep. 2002, doi: 10.1109/TR.2002.802886.
- [17] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, “Hard drive failure prediction using non-parametric statistical methods,” in *Proceedings of ICANN/ICONIP 2003*, Istambul, Turkey, June 26 - 29, 2003, ISBN: 978-3-540-40408-8
- [18] Y. Zhao, X. Liu, S. Gan, and W. Zheng, “Predicting Disk Failures with HMM- and HSMM-Based Approaches”, in *Advances in Data Mining. Applications and Theoretical Aspects*, P. Perner, Ed., in Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, vol. 6171, 2010, pp. 390–404. doi: 10.1007/978-3-642-14400-4\_30.
- [19] Y. Wang, Q. Miao, E. W. M. Ma, K.-L. Tsui, and M. G. Pecht, “Online Anomaly Detection for Hard Disk Drives Based on Mahalanobis Distance”, *IEEE Transactions on Reliability*, vol. 62, no. 1, pp. 136–145, Mar. 2013, doi: 10.1109/TR.2013.2241204.
- [20] Y. Wang, E. W. M. Ma, T. W. S. Chow, and K.-L. Tsui, “A Two-Step Parametric Method for Failure Prediction in Hard Disk Drives”, *IEEE Transactions Industrial Informatics*, vol. 10, no. 1, pp. 419–430, Feb. 2014, doi: 10.1109/TII.2013.2264060.
- [21] B. Zhu, G. Wang, X. Liu, D. Hu, S. Lin, and J. Ma, “Proactive drive failure prediction for large scale storage systems”, in *2013 IEEE 29th Symposium on Mass Storage Systems and Technologies (MSST)*, Long Beach, CA, USA: IEEE, May 2013, pp. 1–5. doi: 10.1109/MSST.2013.6558427.
- [22] J. Qian, S. Skelton, J. Moore, and H. Jiang, “P3: Priority based proactive prediction for soon-to-fail disks”, in *2015 IEEE International Conference on Networking, Architecture and Storage (NAS)*, Boston, MA, USA: IEEE, Aug. 2015, pp. 81–86. doi: 10.1109/NAS.2015.7255224.
- [23] L. P. Queiroz, F. C. M. Rodrigues, J. P. P. Gomes, F. T. Brito, I. C. Chaves, M. R. P. Paula, M. R. Salvador, and J. C. Machado, “A Fault Detection Method for Hard Disk Drives Based on Mixture of Gaussians and Nonparametric Statistics,” *IEEE Transactions Industrial Informatics*, vol. 13, no. 2, pp. 542–550, Apr. 2017, doi: 10.1109/TII.2016.2619180.
- [24] M. L. Lesser and J. W. Haanstra, “The RAMAC data-processing machine: system organization of the IBM 305”, in *Papers and discussions presented at the December 10-12, 1956, eastern joint computer conference: New developments in computers on - AIEE-*

- IRE '56 (Eastern)*, New York, New York: ACM Press, 1957, pp. 139–146. doi: 10.1145/1455533.1455565.
- [25] National Academies of Sciences, Engineering, and Medicine, *Rapid Expert Consultation on Archival Data Storage Technologies for the Intelligence Community*. Washington, D.C.: National Academies Press, 2024, pp. 27445. doi: 10.17226/27445.
- [26] T. Kamigaichi, F. Arai, H. Nitsuta, M. Endo, K. Nishihara, T. Murata, H. Takekida, T. Izumi, K. Uchida, T. Maruyama, I. Kawabata, Y. Suyama, A. Sato, K. Ueno, H. Takeshita, Y. Joko, S. Watanabe, Y. Liu, H. Meguro, A. Kajita, Y. Ozawa, T. Watanabe, S. Sato, H. Tomiie, Y. Kanamaru, R. Shoji, C.H. Lai, M. Nakamichi, K. Oowada, T. Ishigaki, G. Hemink, D. Dutta, Y. Dong, C. Chen, G. Liang, M. Higashitani, and J. Lutze, “Floating Gate super multi level NAND Flash Memory Technology for 30nm and beyond”, in *2008 IEEE International Electron Devices Meeting*, San Francisco, CA, USA: IEEE, Dec. 2008, pp. 1–4. doi: 10.1109/IEDM.2008.4796825.
- [27] S. Ikegawa, F. B. Mancoff, J. Janesky, and S. Aggarwal, “Magnetoresistive Random Access Memory: Present and Future”, *IEEE Transactions Electron Devices*, vol. 67, no. 4, pp. 1407–1419, Apr. 2020, doi: 10.1109/TED.2020.2965403.
- [28] P. Rai, A. Prasad, S. M. Reddy, and A. Chinmay, “Evolution of Optical Storage in Computer Memory”, in *Advances in Data Science and Management*, S. Borah, V. Emilia Balas, and Z. Polkowski, Eds., in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 37, 2020, pp. 489–495, doi: 10.1007/978-981-15-0978-0\_48
- [29] M. Chris, “Analyst forecasts continued HDD sales beyond 2028”, Accessed: May 13, 2024. [Online]. Available: <https://blocksandfiles.com/2024/05/13/coughlin-associates-hdd/>
- [30] B. Marchon and T. Olson, “Magnetic Spacing Trends: From LMR to PMR and Beyond”, *IEEE Transactions on Magnetics*, vol. 45, no. 10, pp. 3608–3611, Oct. 2009, doi: 10.1109/TMAG.2009.2023624.
- [31] “Seagate x16” Seagate, 2019. Accessed: Jul. 12, 2024. [Online]. Available: <https://www.seagate.com/www-content/product-content/enterprise-hdd-fam/exos-x-16/en-us/docs/100845789f.pdf>
- [32] M. H. Kryder, E. C. Gage, T. W. McDaniel, W. A. Challener, R. E. Rottmayer, G. Ju, Y.-T. Hsia, and M. F. Erden, “Heat Assisted Magnetic Recording”, *Proceedings of IEEE*, vol. 96, no. 11, pp. 1810–1835, Nov. 2008, doi: 10.1109/JPROC.2008.2004315.
- [33] Seagate, “Seagate Mozaic 3+”, 2024. Accessed: Dec. 16, 2024. [Online]. Available: <https://www.seagate.com/content/dam/seagate/assets/resources/enterprise/white-paper/what-is-mozaic-3plus-how-does-it-work-and-what-can-it-do-for-me.pdf>
- [34] IDEMA, “HDD Projection”, 2022. Accessed: Dec. 17, 2024. [Online]. Available: <https://www.asrc.idema.org>
- [35] A. al Mamun, G. Guo, and C. Bi, *Hard disk drive: mechatronics and control*. in *Automation and control engineering*, no. 23. Boca Raton, Fla.: CRC Press, 2007. ISBN: 978-0849372537
- [36] T. Atsumi, S. Nakamura, M. Furukawa, I. Naniwa, and J. Xu, “Triple-Stage-Actuator System of Head-Positioning Control in Hard Disk Drives” *IEEE Transactions on Magnetics*, vol. 49, no. 6, pp. 2738–2743, Jun. 2013, doi: 10.1109/TMAG.2013.2260136.

- [37] F. Forensic, “Bad Sector Recovery”, Accessed: Jan. 21, 2025. [Online]. Available: <https://www.forensicfocus.com/articles/bad-sector-recovery/>
- [38] M. S. Manasse, C. A. Thekkath, A. Silverberg, and C. Thekkath, “A Reed-Solomon Code for Disk Storage, and Efficient Recovery Computations for Erasure-Coded Disk Storage”, *Proc. Inform.*, Accessed: Jan. 11, 2025, [Online]. Available: <https://www.microsoft.com/en-us/research/publication/a-reed-solomon-code-for-disk-storage-and-efficient-recovery-computations-for-erasure-coded-disk-storage/>
- [39] “Advanced Format.” Western Digital, 2018. Accessed: Dec. 21, 2024. [Online]. Available: [https://documents.westerndigital.com/content/dam/doc-library/en\\_us/assets/public/western-digital/collateral/white-paper/white-paper-advanced-format.pdf](https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/collateral/white-paper/white-paper-advanced-format.pdf)
- [40] D. Nosachev, *Advanced format (4Kib) HDD sector*. Feb. 11, 2025. [Online]. Available: [https://commons.wikimedia.org/wiki/File:Advanced\\_format\\_\(4Kib\)\\_HDD\\_sector.svg](https://commons.wikimedia.org/wiki/File:Advanced_format_(4Kib)_HDD_sector.svg)
- [41] A. Silberschatz, P. B. Galvin, and G. Gagne, “Operating System Concepts”, 8th Edition, Wiley, 2008, ISBN: 978-0470128725
- [42] W. Stallings, *Computer organization and architecture: designing for performance*, 10th edition. Boston: Pearson-Prentice Hall, 2016, ISBN: 978-0136073734
- [43] Western Digital, “SMR Technology”, Accessed: Dec. 27, 2024, [Online]. Available: [https://documents.westerndigital.com/content/dam/doc-library/en\\_us/assets/public/western-digital/collateral/white-paper/white-paper-shingled-magnetic-recording-hdd-technology.pdf](https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/collateral/white-paper/white-paper-shingled-magnetic-recording-hdd-technology.pdf)
- [44] B. Schroeder and G. A. Gibson, “Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?”, *ACM Transactions on Storage (TOS)*, Volume 3, Issue 3., 2007, doi: 10.1145/1288783.1288785
- [45] *Bathtub curve* [Online]. Accessed: Dec. 30, 2024, [Online]. Available: [https://upload.wikimedia.org/wikipedia/commons/7/78/Bathtub\\_curve.svg](https://upload.wikimedia.org/wikipedia/commons/7/78/Bathtub_curve.svg)
- [46] S. Đurašević and B. Đorđević, “Anomaly detection using SMART indicators for hard disk drive failure prediction”, *IcETRAN 2017 conference proceedings*, Beograd, Belgrade: Društvo za ETRAN : Akademska misao : ETRAN Society : Academic Mind, Kladovo, Serbia, June 5-8, 2017, pp. RT11.6.1-4., ISBN 978-86-7466-692-0
- [47] Y. Wang, Q. Miao, and M. Pecht, “Health monitoring of hard disk drive based on Mahalanobis distance”, in *2011 Prognostics and System Health Managment Confernece*, Shenzhen, China: IEEE, May 2011, pp. 1–8. doi: 10.1109/PHM.2011.5939558.
- [48] G. Wiederhold and J. McCarthy, “Arthur Samuel: Pioneer in Machine Learning”, *IBM Journal of Research and Development*, vol. 36, no. 3, pp. 329–331, May 1992, doi: 10.1147/rd.363.0329.
- [49] T. M. Mitchell, *Machine Learning*. McGraw-Hill series in computer science. New York: McGraw-Hill, 1997, ISBN: 978-0071154673
- [50] X. Jin and J. Han, “K-Means Clustering”, in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds., Boston, MA: Springer US, 2011, pp. 563–564. doi: 10.1007/978-0-387-30164-8\_425.
- [51] D. Deng, “DBSCAN Clustering Algorithm Based on Density”, in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, Hefei, China: IEEE, Sep. 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.

- [52] F. Nielsen, “Hierarchical Clustering”, in *Introduction to HPC with MPI for Data Science*, in Undergraduate Topics in Computer Science, Cham: Springer International Publishing, 2016, pp. 195–211. doi: 10.1007/978-3-319-21903-5\_8.
- [53] M. Al-Maolegi and B. Arkok, “An Improved Apriori Algorithm For Association Rules”, *Int. J. Nat. Lang. Comput.*, vol. 3, no. 1, pp. 21–29, Feb. 2014, doi: 10.5121/ijnlc.2014.3103.
- [54] H.-J. Jang, Y. Yang, J. S. Park, and B. Kim, “FP-Growth Algorithm for Discovering Region-Based Association Rule in the IoT Environment”, *Electronics*, vol. 10, no. 24, pp. 3091, Dec. 2021, doi: 10.3390/electronics10243091.
- [55] V. Srinadh, “Evaluation of Apriori, FP growth and Eclat association rule mining algorithms”, *International Journal of Health Sciences*, vol. 6(S2), pp. 7475–7485, Apr. 2022, doi: 10.53730/ijhs.v6nS2.6729.
- [56] G. I. Webb, “Overfitting”, in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds., Boston, MA: Springer US, 2017, pp. 947–948. doi: 10.1007/978-1-4899-7687-1\_960.
- [57] K. P. Burnham and D. R. Anderson, Eds., *Model Selection and Multimodel Inference*. New York, NY: Springer New York, 2002. ISBN: 978-0-387-95364-9
- [58] M. Radaković, “Audio Signal Preparation Process for Deep Learning Application Using Python”, in *Proceedings of the International Scientific Conference - Sinteza 2021*, Beograd, Serbia: Singidunum University, 2021, pp. 146–152. doi: 10.15308/Sinteza-2021-146-152.
- [59] S. Cheng and M. Pecht, “Using cross-validation for model parameter selection of sequential probability ratio test”, *Expert Syst. Expert Systems with Applications*, vol. 39, no. 9, pp. 8467–8473, Jul. 2012, doi: 10.1016/j.eswa.2012.01.172.
- [60] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection”, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, in IJCAI’95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143, doi: 10.5555/1643031.1643047
- [61] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation”, *PeerJ Computer Science*, vol. 7(3), pp. e623, 2021, doi: 10.7717/peerj-cs.623
- [62] D. Santos Martins, “Biometric recognition based on the texture along palmprint principal lines”, Master thesis, Faculty of Engineering, University of Porto, Porto, 2011.
- [63] W. Li, “Hard disk drive failure detection with recurrence quantification analysis”, Master thesis, Northeastern University, 2020. doi: 10.17760/D20385579.
- [64] T.-T. Teoh, S.-Y. Cho, and Y.-Y. Nguwi, “Hidden Markov Model for hard-drive failure detection”, in *2012 7th International Conference on Computer Science & Education (ICCSE)*, Melbourne, Australia: IEEE, Jul. 2012, pp. 3–8. doi: 10.1109/ICCSE.2012.6295014.
- [65] Y. Wang, S. Jiang, L. He, Y. Peng, and T. W. Chow, “Hard Disk Drives Failure Detection Using A Dynamic Tracking Method”, in *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, Helsinki, Finland: IEEE, Jul. 2019, pp. 1473–1477. doi: 10.1109/INDIN41052.2019.8972108.

- [66] Y. Wang, L. He, S. Jiang, and T. W. S. Chow, "Failure Prediction of Hard Disk Drives Based on Adaptive Rao–Blackwellized Particle Filter Error Tracking Method", *IEEE Transactions Industrial Informatics*, vol. 17(2), pp. 913–921, 2021, doi: 10.1109/TII.2020.3016121.
- [67] A. Harish, G. Prakash, R. R. Nair, V. Bhaskaran Iyer, B. R. Mohan, and M. Das, "Hidden Markov Model for Hard Disk Drive Failure Detection", in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, Gwalior, India: IEEE, Mar. 2024, pp. 1–5. doi: 10.1109/IATMSI60426.2024.10502467.
- [68] V. Agarwal, C. Bhattacharyya, T. Niranjana, and S. Susarla, "Discovering Rules from Disk Events for Predicting Hard Drive Failures", in *2009 International Conference on Machine Learning and Applications*, Miami, FL, USA: IEEE, Dec. 2009, pp. 782–786. doi: 10.1109/ICMLA.2009.62.
- [69] B. Y. Vineetha and A. Borah, "Enhanced Rules Framework for Predicting Disk Drives Failures", *International Journal of Computer Science and Mobile Computing*, vol. 5, no. 7, pp. 177–185, 2016, ISSN 2320–088X
- [70] S. Kirdponpattara, P. Sooraksa, and V. Boonjing, "Building a Rule-Based Expert System to Enhance the Hard Disk Drive Manufacturing Processes", *IEEE Access*, vol. 12, pp. 29558–29570, 2024, doi: 10.1109/ACCESS.2024.3369443.
- [71] Y. Wang, K. Tsui, E. W. M. Ma, and M. Pecht, "A fusion approach for anomaly detection in hard disk drives", in *Proceedings of the IEEE 2012 Prognostics and System Health Management Conference (PHM-2012 Beijing)*, Beijing, China: IEEE, May 2012, pp. 1–5. doi: 10.1109/PHM.2012.6228814.
- [72] M. Natsumeda, "Failure Prediction of Hard Disk Drives in Redundant Arrays Using Disk-Level Performance Metrics", *PHM Soc. Asia-Pacific Conference*, vol. 4, no. 1, Sep. 2023, doi: 10.36001/phmap.2023.v4i1.3626.
- [73] F. L. F. Pereira, D. N. Teixeira, J. P. P. Gomes, and J. C. Machado, "Evaluating One-Class Classifiers for Fault Detection in Hard Disk Drives", in *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, Salvador, Brazil: IEEE, Oct. 2019, pp. 586–591. doi: 10.1109/BRACIS.2019.00108.
- [74] W. Yang, D. Hu, Y. Liu, S. Wang, and T. Jiang, "Hard Drive Failure Prediction Using Big Data", in *2015 IEEE 34th Symposium on Reliable Distributed Systems Workshop (SRDSW)*, Montreal, QC, Canada: IEEE, Sep. 2015, pp. 13–18. doi: 10.1109/SRDSW.2015.15.
- [75] P. H. Franklin, "Predicting disk drive failure using condition based monitoring", in *2017 Annual Reliability and Maintainability Symposium (RAMS)*, Orlando, FL, USA: IEEE, 2017, pp. 1–5. doi: 10.1109/RAM.2017.7889647.
- [76] C. A. Rincon C., J.-F. Paris, R. Vilalta, A. M. K. Cheng, and D. D. E. Long, "Disk failure prediction in heterogeneous environments", in *2017 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, Seattle, WA: IEEE, Jul. 2017, pp. 1–7. doi: 10.23919/SPECTS.2017.8046776.
- [77] M. M. Botezatu, I. Giurgiu, J. Bogojeska, and D. Wiesmann, "Predicting Disk Replacement towards Reliable Data Centers", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 39–48. doi: 10.1145/2939672.2939699.

- [78] N. Aussel, S. Jaulin, G. Gandon, Y. Petetin, E. Fazli, and S. Chabridon, “Predictive Models of Hard Drive Failures Based on Operational Data”, in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Cancun, Mexico: IEEE, Dec. 2017, pp. 619–625. doi: 10.1109/ICMLA.2017.00-92.
- [79] J. Xiao, Z. Xiong, S. Wu, Y. Yi, H. Jin, and K. Hu, “Disk Failure Prediction in Data Centers via Online Learning”, in *Proceedings of the 47th International Conference on Parallel Processing*, Eugene OR USA: ACM, Aug. 2018, pp. 1–10. doi: 10.1145/3225058.3225106.
- [80] C.J. Su and S.F. Huang, “Real-time big data analytics for hard disk drive predictive maintenance”, *Computers & Electrical Engineering*, vol. 71, pp. 93–101, Oct. 2018, doi: 10.1016/j.compeleceng.2018.07.025.
- [81] M. Züfle, C. Krupitzer, F. Erhard, J. Grohmann, and S. Kounev, “To Fail or Not to Fail: Predicting Hard Disk Drive Failure Time Windows”, in *Measurement, Modelling and Evaluation of Computing Systems*, vol. 12040, H. Hermanns, Ed., in Lecture Notes in Computer Science, vol. 12040, Cham: Springer International Publishing, 2020, pp. 19–36. doi: 10.1007/978-3-030-43024-5\_2.
- [82] N. M. Rezk, M. Purnaprajna, T. Nordstrom, and Z. Ul-Abdin, “Recurrent Neural Networks: An Embedded Computing Perspective”, *IEEE Access*, vol. 8, pp. 57967–57996, 2020, doi: 10.1109/ACCESS.2020.2982416.
- [83] C. Xu, G. Wang, X. Liu, D. Guo, and T.-Y. Liu, “Health Status Assessment and Failure Prediction for Hard Drives with Recurrent Neural Networks”, *IEEE Transactions on Computer.*, vol. 65, no. 11, pp. 3502–3508, Nov. 2016, doi: 10.1109/TC.2016.2538237.
- [84] J. Zhang, J. Wang, L. He, Z. Li, and P. S. Yu, “Layerwise Perturbation-Based Adversarial Training for Hard Drive Health Degree Prediction”, *2018 IEEE International Conference on Data Mining (ICDM)*, Singapore: IEEE, Nov. 2018, pp. 1428–1433. doi: 10.1109/ICDM.2018.00197.
- [85] D. Liu, B. Wang, P. Li, R. J. Stones, T. G. Marbach, G. Wang, X. Liu, and Z. Li, “Predicting Hard Drive Failures for Cloud Storage Systems”, *Algorithms and Architectures for Parallel Processing*, S. Wen, A. Zomaya, and L. T. Yang, Eds., in Lecture Notes in Computer Science, vol. 11944, Cham: Springer International Publishing, 2020, pp. 373–388. doi: 10.1007/978-3-030-38991-8\_25.
- [86] H. Zhou, Z. Niu, G. Wang, X. Liu, D. Liu, B. Kang, Z. Hu, and Y. Zhang, “Proactive Drive Failure Prediction for Cloud Storage System Through Semi-Supervised Learning”, *IEEE Trans. Dependable Secure Computing*, vol. 21, no. 4, pp. 1528–1543, Jul. 2024, doi: 10.1109/TDSC.2023.3286093.
- [87] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A Search Space Odyssey”, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017, doi: 10.1109/TNNLS.2016.2582924.
- [88] A. Bai, M. Chen, S. Peng, G. Han and Z. Yang, “Attention-Based Bidirectional LSTM With Differential Features For Disk RUL Prediction”, *2022 IEEE 5th International Conference on Electronic Information and Communication Technology (ICEICT)*, Hefei, China, 2022, pp. 684-689, doi: 10.1109/ICEICT55736.2022.9908671.
- [89] G. Han, Y. Wang, Q. Hai, Z. Yang, and S. Peng, “Adversarial Domain Generalization Network for New Enabled Hard Disk Drive Failure Prediction”, *IEEE Transaction of Magnetism*, pp. 1–1, 2025, doi: 10.1109/TMAG.2025.3544584.

- [90] M. Zhang, W. Ge, R. Tang, and P. Liu, “Hard Disk Failure Prediction Based on Blending Ensemble Learning”, *Applied Science*, vol. 13, no. 5, pp. 3288, Mar. 2023, doi: 10.3390/app13053288.
- [91] I. C. Chaves, M. R. P. De Paula, L. G. M. Leite, J. P. P. Gomes, and J. C. Machado, “Hard Disk Drive Failure Prediction Method Based On A Bayesian Network”, *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro: IEEE, Jul. 2018, pp. 1–7. doi: 10.1109/IJCNN.2018.8489097.
- [92] I. C. Chaves, M. R. P. De Paula, L. G. M. Leite, J. P. P. Gomes, and J. C. Machado, “Hard Disk Drive Failure Prediction Method Based On A Bayesian Network”, *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro: IEEE, Jul. 2018, pp. 1–7. doi: 10.1109/IJCNN.2018.8489097.
- [93] F. L. F. Pereira, F. D. D. S. Lima, L. G. D. M. Leite, J. P. P. Gomes, and J. D. C. Machado, “Transfer Learning for Bayesian Networks with Application on Hard Disk Drives Failure Prediction” *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, Uberlandia, Brasil, Oct. 2017, pp. 228–233. doi: 10.1109/BRACIS.2017.64.
- [94] Backblaze Inc., “Hard Drive Data and Stats”, Accessed: Jan. 02, 2025. [Online]. Available: <https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data>
- [95] E. Perrier, A. Youssry, and C. Ferrie, “QDataset: Quantum Datasets for Machine Learning”, *Scientific Data* 9, 582 (2022), doi: 10.1038/s41597-022-01639-1.
- [96] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique”, *Journal of Artificial Intelligence Research*, vol. 16(1), pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
- [97] M. Kuhn and K. Johnson, *Feature engineering and selection: a practical approach for predictive models*, Chapman & Hall/CRC data science series. Boca Raton London New York: CRC Press, Taylor & Francis Group, 2021., ISBN: 978-1138079229
- [98] Н. Бродић, “Семантичка сегментација облака тачака добијеног ласерским скенирањем из ваздуха коришћењем метода машинског учења”, докторска дисертација, Универзитет у Београду, 2023.
- [99] Z. Zhou and G. Hooker, “Unbiased Measurement of Feature Importance in Tree-Based Methods”, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 2, pp. 1–21, Apr. 2021, doi: 10.1145/3429445.
- [100] J. O. Ogutu, T. Schulz-Streeck, and H.-P. Piepho, “Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions”, *BMC Proceedings*, vol. 6, no. S2, pp S2-10, Dec. 2012, doi: 10.1186/1753-6561-6-S2-S10.
- [101] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK ; New York: Cambridge University Press, 2004, ISBN: 978-0521833783
- [102] N. Singh, “Soft Margin SVM: Exploring Slack Variables, the ‘C’ Parameter, and Flexibility”, Accessed: Oct. 28, 2023. [Online]. Available: <https://pub.aimind.so/soft-margin-svm-exploring-slack-variables-the-c-parameter-and-flexibility-1555f4834ecc>
- [103] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines”, *Machine Learning*, vol. 46(1–3), 2002, pp. 389–422, doi: 10.1023/A:1012487302797

- [104] J. D. Gibbons and S. Chakraborti, “Nonparametric Statistical Inference”, in *International Encyclopedia of Statistical Science*, M. Lovric, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 977–979. doi: 10.1007/978-3-642-04898-2\_420.
- [105] G. E. P. Box and D. R. Cox, “An Analysis of Transformations”, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, Jul. 1964, doi: 10.1111/j.2517-6161.1964.tb00553.x.
- [106] I.-K. Yeo, “A new family of power transformations to improve normality or symmetry”, *Biometrika*, vol. 87, no. 4, pp. 954–959, Dec. 2000, doi: 10.1093/biomet/87.4.954.
- [107] E. Blanco-Mallo, L. Morán-Fernández, B. Remeseiro, and V. Bolón-Canedo, “Do all roads lead to Rome? Studying distance measures in the context of machine learning”, *Pattern Recognition*, vol. 141, pp. 109646, Sep. 2023, doi: 10.1016/j.patcog.2023.109646.
- [108] S. R. Dandavolu, “Similarity and Dissimilarity Measures in Data Science”, Accessed: Jun. 24, 2024. [Online]. Available: <https://www.analyticsvidhya.com/blog/2024/06/similarity-and-dissimilarity-measures-in-data-science/>
- [109] Н. Ајзенхамер, А. Букаров, and В. Станковић, *Истраживање података*, Accessed: Oct. 24, 2024. [Online]. Available: <https://www.nikolaajzenhamer.rs/assets/pdf/ip.pdf>
- [110] R. Penrose, “A generalized inverse for matrices”, *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 3, pp. 406–413, Jul. 1955, doi: 10.1017/S0305004100030401.
- [111] М. Котлар, “Детекција аномалија коришћењем мета података у аутоматизованим системима за машинско учење”, докторска дисертација, Универзитет у Београду, 2022.
- [112] T. P. Hill, “A Statistical Derivation of the Significant-Digit Law”, *Statistical Science*, vol. 10, no. 4, Nov. 1995, doi: 10.1214/ss/1177009869.
- [113] A. Berger and T. P. Hill, “Benford’s Law Strikes Back: No Simple Explanation in Sight for Mathematical Gem”, *The Mathematical Intelligencer*, vol. 33, no. 1, pp. 85–91, Mar. 2011, doi: 10.1007/s00283-010-9182-3.
- [114] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey”, *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, Jul. 2009, doi: 10.1145/1541880.1541882.
- [115] K. G. Mehrotra, C. K. Mohan, and H. Huang, *Anomaly Detection Principles and Algorithms*. in Terrorism, Security, and Computation. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-67526-8.
- [116] K. Taunk, S. De, S. Verma, and A. Swetapadma, “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification”, in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India: IEEE, May 2019, pp. 1255–1260. doi: 10.1109/ICCS45141.2019.9065747.
- [117] M. M. Breunig, H.P. Kriegel, R. T. Ng, and J. Sander, “LOF: identifying density-based local outliers”, *ACM SIGMOD Records*, vol. 29, no. 2, pp. 93–104, Jun. 2000, doi: 10.1145/335191.335388.
- [118] H. Xu, G. Pang, Y. Wang, and Y. Wang, “Deep Isolation Forest for Anomaly Detection”, *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12591–12604, Dec. 2023, doi: 10.1109/TKDE.2023.3270293.

- [119] *IEEE Standard Glossary of Software Engineering Terminology*, IEEE Standards, 1990, doi: 10.1109/IEEESTD.1990.101064.
- [120] S. M. Djurasevic, U. M. Pesovic, and B. S. Djordjevic, “Anomaly Detection Model for Predicting Hard Disk Drive Failures”, *Applied Artificial Intelligence.*, vol. 35, no. 8, pp. 549–566, Jul. 2021, doi: 10.1080/08839514.2021.1922840.
- [121] S. Han, P. P. C. Lee, F. Xu, Y. Liu, C. He, and J. Liu, “An In-Depth Study of Correlated Failures in Production SSD-Based Data Centers”, 19th USENIX Conference on File and Storage Technologies (FAST '21) February 23–25, 2021, ISBN 978-1-939133-20-5
- [122] S. Đurašević, U. Pešović, V. Luković, and B. Đorđević, “Analysis of write amplification on solid state drives in datacenters”, presented at the The 19th International Conference on Informatics and Information Technologies – CIIT 2022, Skoplje : Faculty of Computer Science and Engineering, 5-6 May 2022, pp. 54–57, ISBN 978-608-4699-15-6.
- [123] V. Luković, Ž. Jovanović, S. Đurašević Pešović, U. Pešović, and B. Đorđević, “Solid state drive failure prediction using anomaly detection”, *Electronics, Computer Scienec and Engineering Section.*, ISSN: 2079-9292, accepted for publication, 19th March 2025.
- [124] L. P. Queiroz, F. C. M. Rodrigues, J. P. P. Gomes, F. T. Brito, I. C. Brito, and J. C. Machado, “Fault Detection in Hard Disk Drives Based on Mixture of Gaussians”, *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, Recife, Brazil: IEEE, Oct. 2016, pp. 145–150. doi: 10.1109/BRACIS.2016.036.
- [125] G. W. M. Valença, F. L. F. Pereira, F. T. Brito, V. A. E. D. De Farias, and J. C. Machado, “Health Levels Modeling for SSD Failure Prediction”, *Anais do XXXIX Simpósio Brasileiro de Banco de Dados (SBB D 2024)*, Brasil: Sociedade Brasileira de Computação - SBC, Oct. 2024, pp. 764–770. doi: 10.5753/sbbd.2024.243219.
- [126] L. Chen, Z. Zhu, A. Li, N. Mashhadi, R. Frickey, J. Ye, and X Guo, “SSD Drive Failure Prediction on Alibaba Data Center Using Machine Learning”, *2022 IEEE International Memory Workshop (IMW)*, Dresden, Germany: IEEE, May 2022, pp. 1–4. doi: 10.1109/IMW52921.2022.9779284.
- [127] Y. Jiang, R. Lu, S. Zhou, and Q. Li, “Machine Learning Based Collaborative Prediction of SSD Failures in the Cloud”, *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, Victoria, Seychelles: IEEE, Feb. 2024, pp. 1–6. doi: 10.1109/ACDSA59508.2024.10467231.
- [128] F. Xu, S. Han, P. P. C. Lee, Y. Liu, C. He, and J. Liu, “General Feature Selection for Failure Prediction in Large-scale SSD Deployment”, *2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, Taipei, Taiwan: IEEE, Jun. 2021, pp. 263–270. doi: 10.1109/DSN48987.2021.00039.
- [129] Zhang Y.; Hao W.; Niu B.; Liu K.; Wang S.; Liu N.; He X.; Gwon Y.; Koh C, “Multi-view Feature-based SSD Failure Prediction: What, When, and Why”, 21st USENIX Conference on File and Storage Technologies, Santa Clara, CA, USA, Feb. 2023, pp. 409 - 424, doi: 10.5555/3585938.3585964
- [130] Y. Zhang, T. Zhang, W. Hao, S. Wang, N. Liu, and X. He, “MSFRD: Mutation Similarity based SSD Failure Rating and Diagnosis for Complex and Volatile Production Environments”, 2024 USENIX Annual Technical Conference, Santa Clara, CA, USA, Jul. 2024, pp. 869 - 884, doi: 10.5555/3691992.3692045

## Списак слика

|                                                                                                                                             |    |
|---------------------------------------------------------------------------------------------------------------------------------------------|----|
| Слика 1. Прогнозирани тренд производње магнетних дискова (HDD), магнетних трака и полупроводничких дискова (SSD) до 2029. године [29] ..... | 9  |
| Слика 2. Планарно (а) и вертикално (б) уписивање података на магнетни диск.....                                                             | 10 |
| Слика 3. План развоја будућих технологија за магнетно складиштење података [34] ...                                                         | 11 |
| Слика 4. Подела диска на стазе и секторе .....                                                                                              | 12 |
| Слика 5. Структура троструког позиционог актуатора уписно-читајуће главе [36] .....                                                         | 13 |
| Слика 6. Структура сектора података [37].....                                                                                               | 13 |
| Слика 7. Поређење структура стандардног и 4К сектора [40].....                                                                              | 14 |
| Слика 8. Организација података код дискова са већим бројем магнетних плоча [41]....                                                         | 15 |
| Слика 9. Организација стаза (а) са размацама и (б) стаза са преклапањем.....                                                                | 16 |
| Слика 10. Стопе отказа у току животног века диска [45] .....                                                                                | 19 |
| Слика 11. Утицај предвидљивих и непредвидљивих перформанси на отказ диска.....                                                              | 21 |
| Слика 12. Поређење (а) традиционалног програмирања и (б) машинског учења .....                                                              | 25 |
| Слика 13. Типови машинског учења .....                                                                                                      | 26 |
| Слика 14. Процес изградње предиктивног модела код надгледаног машинског учења .                                                             | 26 |
| Слика 15. Одабир оптималне вредности уклапања .....                                                                                         | 29 |
| Слика 16. Фазе у процесу машинског учења [58] .....                                                                                         | 30 |
| Слика 17. Метода k-струке унакрсне валидације.....                                                                                          | 31 |
| Слика 18. Матрица конфузије.....                                                                                                            | 32 |
| Слика 19. Утицај границе одлуке на погрешне класификације [62] .....                                                                        | 34 |
| Слика 20. Одређивање оптималне вредности прага класификације помоћу ROC криве                                                               | 34 |
| Слика 21. Поређење модела класификатора на основу површине AUC-ROC криве .....                                                              | 35 |
| Слика 22. Класификација алгоритама машинског учења за предвиђање отказа [63] .....                                                          | 36 |
| Слика 23. Тренд коришћења разних типова алгоритама за предикцију отказа .....                                                               | 38 |
| Слика 24. Приказ радне температуре Temp4 исправног и отказалог диска .....                                                                  | 43 |
| Слика 25. Абнормалне вредности температуре за параметар Temp3 .....                                                                         | 44 |
| Слика 26. Идентификовање абнормалних вредности атрибута .....                                                                               | 45 |
| Слика 27. Број коригованих абнормалних вредности за све атрибуте скупа података ..                                                          | 45 |
| Слика 28. Технике за балансирање скупа податка.....                                                                                         | 46 |
| Слика 29. Поступак селекције атрибута помоћу методе омотача [98] .....                                                                      | 49 |
| Слика 30. Одређивање оптималне хиперравни на основу SVM алгорита .....                                                                      | 49 |
| Слика 31. SVM алгоритам са а) „тврдом” маргином и б) „меком” маргином [102].....                                                            | 51 |

|                                                                                                                                                              |    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Слика 32. Алгоритам рекурзивне селекције атрибута.....                                                                                                       | 53 |
| Слика 33. Коефицијент корелације за различите односе између две променљиве .....                                                                             | 54 |
| Слика 34. Корелација атрибута и отказа магнетног диска.....                                                                                                  | 55 |
| Слика 35. Међусобна корелација атрибута CMRR скупа података .....                                                                                            | 56 |
| Слика 36. Различити облици нормалне дистрибуције.....                                                                                                        | 59 |
| Слика 37. Дистрибуција вредности температуре Temp1 а) за исправне дискове и б) поређење са нормалном дистрибуцијом .....                                     | 59 |
| Слика 38. Дистрибуција вредности висине уписно читајуће главе FlyHeight3 а) за исправне дискове и б) поређење са нормалном дистрибуцијом .....               | 60 |
| Слика 39. Дистрибуција трансформисаних вредности температуре Temp1 а) за исправне дискове и б) поређење са нормалном дистрибуцијом .....                     | 62 |
| Слика 40. Дистрибуција трансформисане вредности висине уписно читајуће главе FlyHeight3 а) за исправне дискове и б) поређење са нормалном дистрибуцијом..... | 62 |
| Слика 41. Метрике за мерење растојања у вишедимензионалном простору [108] .....                                                                              | 64 |
| Слика 42. Дистрибуција нормализованих вредности температуре Temp1 и Temp4.....                                                                               | 65 |
| Слика 43. Временска серија Махаланобисовог растојања за исправан и отказали диск                                                                             | 66 |
| Слика 44. Различити типови аномалија [111].....                                                                                                              | 68 |
| Слика 45. Класификација алгоритама за детекцију аномалија.....                                                                                               | 69 |
| Слика 46. Вероватноћа појаве водеће цифре према Бенфордовом закону [113] .....                                                                               | 70 |
| Слика 47. Интерквартилно растојање [114].....                                                                                                                | 71 |
| Слика 48. Детекција аномалија у подацима два атрибута температуре Temp1 и Temp4                                                                              | 73 |
| Слика 49. Алгоритамска структура модела за предвиђање отказа [120].....                                                                                      | 75 |
| Слика 50. Имплементација алгоритма у центру за складиштење података .....                                                                                    | 79 |
| Слика 51. Алгоритамска структура модела за предвиђање отказа SSD дискова [123] ...                                                                           | 81 |
| Слика 52. Одређивање грубе границе одлучивања за најутицајнији атрибут GList1 .....                                                                          | 83 |
| Слика 53. Утицај границе одлуке на број исправних и погрешних класификација за атрибут GList1 .....                                                          | 84 |
| Слика 54. Перформансе класификације у зависности од вредности границе одлуке за атрибут GList1 .....                                                         | 84 |
| Слика 55. Одређивање оптималне границе одлучивања за атрибут GList1 .....                                                                                    | 85 |
| Слика 56. Примена оптималне границе одлучивања на тест скупу за атрибут GList1 ...                                                                           | 85 |
| Слика 57. Граница одлучивања код модела са SVM-RFE рангираном листом атрибута                                                                                | 86 |
| Слика 58. Прецизност модела у зависности од броја коришћених атрибута.....                                                                                   | 86 |
| Слика 59. Осетљивост модела у зависности од броја коришћених атрибута .....                                                                                  | 87 |
| Слика 60. Прецизност и осетљивост модела у зависности од броја коришћених атрибута .....                                                                     | 88 |
| Слика 61. Граница одлучивања код модела са оптимизованом листом атрибута.....                                                                                | 90 |

|                                                                                                               |    |
|---------------------------------------------------------------------------------------------------------------|----|
| Слика 62. Прецизност оптимизованог модела у зависности од броја атрибута .....                                | 90 |
| Слика 63. Осетљивост оптимизованог модела у зависности од броја атрибута.....                                 | 91 |
| Слика 64. Прецизност и осетљивост оптимизованог модела у зависности од броја коришћених атрибута .....        | 91 |
| Слика 65. Прецизност и осетљивост пре и после оптимизације модела .....                                       | 93 |
| Слика 66. Детекција отказа за магнетни диск са серијским бројем 100138 .....                                  | 93 |
| Слика 67. Махаланобисово растојање за исправан диск са серијским бројем 100350....                            | 94 |
| Слика 68. Предиктивне перформансе модела у зависности од броја коришћених атрибута MC1 модела SSD диска ..... | 97 |

## Списак табела

|                                                                                   |    |
|-----------------------------------------------------------------------------------|----|
| Табела 1. Величине поља стандардног и 4K сектора.....                             | 14 |
| Табела 2. Времена приступа код савремених магнетних дискова [42].....             | 15 |
| Табела 3. Учесталост појаве различитих типова грешки .....                        | 17 |
| Табела 4. Битни SMART параметри који утичу на отказ магнетног диска.....          | 22 |
| Табела 5. Карактеристике различитих RAID конфигурација .....                      | 24 |
| Табела 6. Пример дела скупа CMRR података за магнетне дискове .....               | 39 |
| Табела 7. Поређење различитих скупова података магнетних дискова [63].....        | 41 |
| Табела 8. Заступљеност класа у CMRR скупу података .....                          | 46 |
| Табела 9. Листа уклоњених атрибута са константним вредностима .....               | 53 |
| Табела 10. Листа ранжираних атрибута SVM-RFE.....                                 | 57 |
| Табела 11. Вредности $\lambda$ за Вох-Сох трансформацију атрибута CMRR скупа..... | 61 |
| Табела 12. Листа тринаест најзначајнијих атрибута за реализовани модел.....       | 88 |
| Табела 13. Резултати теста за SVM-RFE рангиране атрибуте .....                    | 89 |
| Табела 14. Листа ранжираних атрибута SVM-RFE након оптимизације .....             | 89 |
| Табела 15. Резултати теста након оптимизације SVM-RFE листе атрибуте .....        | 92 |
| Табела 16. Листа седам најзначајнијих атрибута за оптимизовани модел.....         | 92 |
| Табела 17. Преглед резултата предикције отказа магнетних дискова на CMRR скупу..  | 95 |
| Табела 18. Поређење времена предикције отказа магнетних дискова.....              | 96 |
| Табела 19. Партиционисање Alibaba SSD скупа података .....                        | 97 |
| Табела 20. Поређење алгоритама на Alibaba SSD скупу података.....                 | 98 |

## Биографија

Слађана Ђурашевић Пешовић, рођена је 1985. године у Ивањици, где је завршила основну и средњу школу. Основне четворогодишње академске студије индустријског менаџмента уписује 2004. године на Техничком факултету у Чачку. Признавањем положених испита, уписује се на интегрисане академске студије технике и информатике (300 ЕСПБ) на истом факултету, које завршава 2011. године, са просечном оценом 8.57 и стиче звање дипломирани професор технике и информатике – мастер. Након завршетка студија, изводила је наставу на предметима Техничко и информатичко образовање и Информатика и рачунарство у основној школи. Полагањем диференцијалних испита, 2015. године стиче услов за упис на докторске академске студије на смеру Електротехничко и рачунарско инжењерство, модул Рачунарска техника, на Факултету техничких наука у Чачку. Положила је све предмете прописане планом и програмом докторских академских студија са просечном оценом 9.67.

У звање истраживач-приправник изабрана је 2018. године, на Факултету техничких наука у Чачку. Од 2019. године ангажована је као истраживач на пројекту Министарства просвете, науке и технолошког развоја ТР32043. Након завршетка пројектног циклуса 2020. године, наставила је истраживања уз финансијску подршку Министарства науке, технолошког развоја и иновација. Одлуком Већа за техничко - технолошке науке Универзитета у Крагујевцу, од 10. новембра 2021. године, одобрена јој је тема докторске дисертације под називом „*Модел предикције отказа магнетних дискова заснован на детекцији аномалија*“. У звање истраживач сарадник изабрана је 2022. године, на Факултету техничких наука у Чачку.

Област истраживања обухвата предикцију отказа магнетних дискова применом метода машинског учења и на основу резултата истраживања објавила је два научна рада у часописима на SCI листи, неколико радова у националним часописима и више од двадесет радова које је као аутор и коаутор представила и објавила на међународним и домаћим научним конференцијама. Рецензирала је радове у научним часописима IEEE Transactions on Industrial Informatics и IEEE Transactions on Reliability. Била је члан организационог одбора међународне конференције TIE 2024. године.

Учествовала је на пројекту финансираног од стране Иновационог фонда, члан је пројектног тима билатералне сарадње са Италијом за период 2025-2027 године и члан је тимова на пројектима који су у процесу евалуације. Била је учесник летње школе отворене науке “*University of Maribor Open Science Summer School*”, у оквиру Erasmus+ Blended Intensive Programme (BIP), 2022. и 2023. године.

Члан је Савеза инжењера и техничара Србије (област електротехника и информатика). Поседује сертификате из области машинског учења, база података и програмског језика Java. Као студент била је члан тима који је освојио прву награду на конкурс Развојне агенције Србије са пројектном идејом под називом „*Пчеларство за сваког*“, такође била је члан тима који је са идејом „*SMS Pastir*“ освојио друго место на такмичењу дани иновација ADRIATinn у организацији Факултета инжењерских наука.

**ИЗЈАВА АУТОРА О ОРИГИНАЛНОСТИ ДОКТОРСKE ДИСЕРТАЦИЈЕ**

Изјављујем да докторска дисертација под насловом:

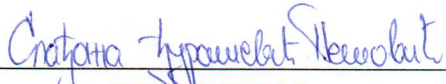
**МОДЕЛ ПРЕДИКЦИЈЕ ОТКАЗА МАГНЕТНИХ ДИСКОВА**  
**ЗАСНОВАН НА ДЕТЕКЦИЈИ АНОМАЛИЈА**

представља оригинално ауторско дело настало као резултат сопственог истраживачког рада.

Овом Изјавом такође потврђујем:

- да сам једини аутор наведене докторске дисертације,
- да у наведеној докторској дисертацији нисам извршио/ла повреду ауторског нити другог права интелектуалне својине других лица,

У Чачку, 24.март 2025. године,

  
потпис аутора

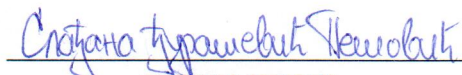
**ИЗЈАВА АУТОРА О ИСТОВЕТНОСТИ ШТАМПАНЕ И ЕЛЕКТРОНСКЕ ВЕРЗИЈЕ  
ДОКТОРСКЕ ДИСЕРТАЦИЈЕ**

Изјављујем да су штампана и електронска верзија докторске дисертације под насловом:

**МОДЕЛ ПРЕДИКЦИЈЕ ОТКАЗА МАГНЕТНИХ ДИСКОВА**  
**ЗАСНОВАН НА ДЕТЕКЦИЈИ АНОМАЛИЈА**

истоветне.

У Чачку, 24.март 2025. године,

  
потпис аутора

**ИЗЈАВА АУТОРА О ИСКОРИШЋАВАЊУ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ**

Ја, Слађана Ђурашевић Пешовић,

☒

дозвољавам

☐

не дозвољавам

Универзитетској библиотеци у Крагујевцу да начини два трајна умножена примерка у електронској форми докторске дисертације под насловом:

**МОДЕЛ ПРЕДИКЦИЈЕ ОТКАЗА МАГНЕТНИХ ДИСКОВА**

**ЗАСНОВАН НА ДЕТЕКЦИЈИ АНОМАЛИЈА**

и то у целини, као и да по један примерак тако умножене докторске дисертације учини трајно доступним јавности путем дигиталног репозиторијума Универзитета у Крагујевцу и централног репозиторијума надлежног министарства, тако да припадници јавности могу начинити трајне умножене примерке у електронској форми наведене докторске дисертације путем *преузимања*.

Овом Изјавом такође

☒

дозвољавам

☐

не дозвољавам<sup>1</sup>

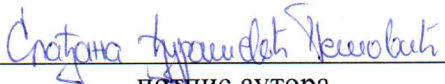
припадницима јавности да тако доступну докторску дисертацију користе под условима утврђеним једном од следећих *Creative Commons* лиценци:

---

<sup>1</sup> Уколико аутор изабере да не дозволи припадницима јавности да тако доступну докторску дисертацију користе под условима утврђеним једном од *Creative Commons* лиценци, то не искључује право припадника јавности да наведену докторску дисертацију користе у складу са одредбама Закона о ауторском и сродним правима.

- 1) Ауторство
- ② Ауторство - делити под истим условима
- 3) Ауторство - без прерада
- 4) Ауторство - некомерцијално
- 5) Ауторство - некомерцијално - делити под истим условима
- 6) Ауторство - некомерцијално - без прерада<sup>2</sup>

У Чачку, 24.март 2025. године,

  
\_\_\_\_\_  
потпис аутора

---

<sup>2</sup> Молимо ауторе који су изабрали да дозволе припадницима јавности да тако доступну докторску дисертацију користе под условима утврђеним једном од *Creative Commons* лиценци да заокруже једну од понуђених лиценци. Детаљан садржај наведених лиценци доступан је на: <http://creativecommons.org.rs/>