

Univerzitet „Džemal Bijedić“

Fakultet Informacijskih Tehnologija

FINALNI DOKUMENT

PREDMET: Mašinsko učenje

Profesor: Nina Bijedić

Student: Medisa Šatara, IB170012

## OPIS PROBLEMA

Ovaj dokument je prikaz zadataka iz mašinskog učenja. Kroz zadatke prolazimo od prečišćavanja, pripreme podataka, metoda i algoritama predviđanja, te detaljnije obrade jedne od metoda. Podatke učitavamo iz txt file, koji su stvarni

U sljedećim poglavljima su prikazana objašnjenja za svaki zadatak.

## ZADAĆA 1 – Priprema podataka

Da bi smo mogli raditi određene taskove pravilno, potrebno je podatke koje dobijemo pripremiti na pravilan i odgovarajući način. Podatke smo učitali iz txt file u Excel.

Sadrže podatke o studentima od 1-4 godine, njihovo prethodno školovanje, završetak školovanje te informacije o trenutnom statusu studenata tih godina. File sadrži 5720 podataka.

Način prečišćavanja podatak odrađen u zadaći 1:

1. Nazivi kolona „BrojKredita“ zamjenjen nazivom ECTS, datum\_diplomiranja -> DatumDiplomiranja, nacin\_studiranja -> NacinStudiranja, opstina\_prebivalista -> Opstina prebivalista, opstina\_rodjenja -> OpstinaRodjenja, Srednja\_skola -> SrednjaSkola, datum\_upisa -> DatumUpisa
2. Promjena tipa podataka u koloni datum diplomiranja, položio datum, i datum upisa
3. Vrijednosti NULL zamjenjene praznim stringom
4. Kolona Priznat, vrijednosti 0 i 1, zamjenjene sa NE i DA
5. Spajanje kolone – Kolone Ocjena i Opisna, spojeno u jednu kolonu ocjena, gdje umjesto NULL pisemo priznato iz tabele opisna, a ako NULL nema nikakvu vrijednost u tabeli opisna pisemo 5, a ostale ocjene prepisemo

Nakon odbrane zadace, i ukazanih grešaka od strane profesorice, izmjenjeni su podaci, što je urađeno u zadaći 3.

## ZADAĆA 2 – UPITI ZA PREDIKCIJU

Na osnovu podataka koje imamo u tabeli za studente, napisali smo upite za predikcije, koje smatramo da bi mogli predviditi.

Predikcija je sposobnost da se predvide izlazne vrijednosti na osnovu ulaznih podataka preko treninga skupa podataka. Iz tog razlog smo prvo dobili skupljene podatke o studentima, zatim smo pripremili podatke.

Upiti koje sam navela kao predikcija su:

1. Predikcija budućeg uspjeha studenta - studenti sa visokim ocjenama, vjerovatno, će imati veću šansu za uspjeh u školskim aktivnostima, pa i u karijeri. Na osnovu ocjena, datuma diplomiranja, načinu studiranja može se predvidjeti uspjeh studenta na tržištu nakon završetka studija.

2. Predikcija prosječnog trajanja studija na temelju načina studiranja – način studiranja može biti DL i redovan, te se na osnovu načina studiranja i položenih predmeta, može predvidjeti koji studenti će prije završiti studij.
3. Predikcija interesovanja studenta – na osnovu ocjena studenta i datuma polaganja, može se predvidjeti interesovanje i sposobnost studenta za određenu oblast, te ukazati na sklonost ka toj naučnoj disciplini.
4. Predikcija napuštanja studija – na osnovu ocjena, datuma polaganja, datuma upisa, može se predvidjeti koji broj studenata bi mogao napustiti studij. Studenti s rijetko položenim predmetima, velikim brojem izlaska na isti, te dužim trajanjem studija, imaju veću mogućnost napuštanja studija.

### ZADAĆA 3 – DORADA ZADAĆE 1 (PRIPREMA PODATAKA)

Pripremljeni podaci u zadaci 1 nisu bili uredni, jer fokus treba biti student, dakle jedan red je jedan student. To se dobiva kreiranjem pivot tabele na osnovu studenta, te dodavanje dodatnih informacija potrebne za deskriptivne metode.

Podaci s prve zadaće su prečišćeni i doradeni na sljedeći način:

1. Sortiran studentId – sortiranje studentId od najmanjeg ka najvećem
2. Promjena podataka za DatumUpisa, PoložioDatum
3. Spajanje tabele Ocjena i opisna u jednu tabelu Ocjene, na način da se iz tabele Ocjene prepisuju ocjene, a ako je opisna ocjena priznat, vrijednost će biti 7, a ako je zadovoljava biti će 10, a ako je opisna ocjena prazna, onda je to ocjena 5
4. Dodavanje nove kolone Diplomirao na osnovu tabele DatumDiplomiranja, uneseni datum u red, predstavljen je sa 1, a u suprotnom sa 0
5. Promjena naziva kolona

U finalnim podacima su predstavljene kolone studentId, ocjene po predmetima, te dodatne informacije poput datuma upisa, srednja skola, godina zavrsetka, nacin studiranja, da li je diplomirao, spol i godina studija.

### ZADAĆA 4 – DESKRIPTIVNE METODE

Nakon preciscavanja podataka, podaci se spasavaju kao .csv file, te se pusta deskriptivna metoda. U ovom slucaju, odluceno je za klasterizaciju.

Deskriptivna metoda znaci da sistem koristi podatke kako bi objasnio sta se dogodilo. Kod klasterizacija, cilj je grupisati tačke podataka koje su slične. Iz organizacionih razloga, zgodno je imati različite klase podataka. Ljudima može biti lakše pregledati podatke ako su unaprijed

grupno kategorizirani. U ovoj zadaci koristio se Visual Studio 2022 za K-Means klasterovanje. Nakon što se napravi projekat, instalira se ML.NET za klasterizaciju.

K-Means klasterovanje je deskriptivna metoda za grupisanje tačaka podataka u slične klaster. Proces klasterizacije se odvija tako što se prvo biraju početni centri klastera iz skupa podataka. Zatim se dodjeljuje svaka tačka podataka najbližem klaster centru na osnovu metričke udaljenosti. Centar klastera se izračunava prosječnim vrijednostima svih tačaka podataka dodjeljenih svakom klasteru..

```
public class StudentsData
{
    [LoadColumn(1)]
    public float P104 {get;set;}
    [LoadColumn(2)]
    public float P12 { get; set; }
    [LoadColumn(3)]
    public float P140 { get; set; }
    [LoadColumn(4)]
    public float P141 { get; set; }
    [LoadColumn(5)]
    public float P142 { get; set; }
    [LoadColumn(6)]
    //..(ostatak kolone
    [LoadColumn(10)]
    public float P88 { get; set; }
}
```

Klasa „StudentData“ sadrži sve kolone iz skupa podataka.

```
public class StudentPrediction
{
    [ColumnName("PredictedLabel")]
    public uint PredictedClusterId;

    [ColumnName("Score")]
    public float[]? Distances;
}
```

Klasa „StudentPrediction“ sadrži Id klastera i niz udaljenosti od centra klastera.

```
private static void Main(string[] args)
{
    StudentsData newSample = new StudentsData();

    KMeansTrainer trainer = new KMeansTrainer(numberOfClusters: 5);

    TrainEvaluatePredict(trainer, newSample);

    Console.WriteLine("Press any key to exit...");
    Console.ReadKey();
}
```

```

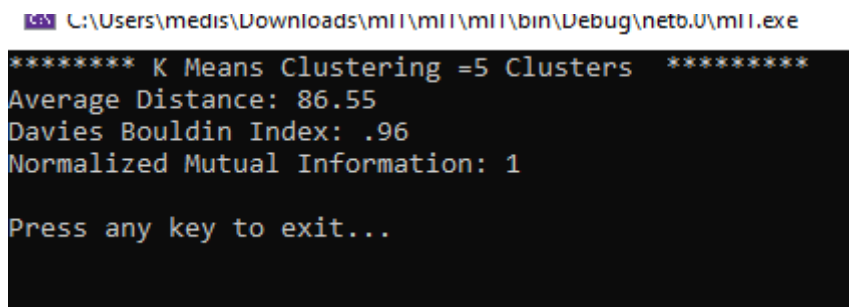
    }

    static void TrainEvaluatePredict(ITrainerBase trainer, StudentsData newSample)
    {
        Console.WriteLine("***** " + $"{trainer.Name}" + " *****");
        trainer.Fit("C:\\Users\\medis\\Desktop\\podaciML.csv");
        var modelMetrics = trainer.Evaluate();

        Console.WriteLine($"Average Distance:
{modelMetrics.AverageDistance:###}{Environment.NewLine}" +
            $"Davies Bouldin Index:
{modelMetrics.DaviesBouldinIndex:###}{Environment.NewLine}" +
            $"Normalized Mutual Information:
{modelMetrics.NormalizedMutualInformation:###}{Environment.NewLine}");
        trainer.Save();
    }

```

Ovaj dio koda upravlja treningom, evaluacijom i predikcijom za K-means klasterovanje.



```

C:\Users\medis\Downloads\mll\mll\mll\bin\Debug\net6.0\mll.exe
***** K Means Clustering =5 Clusters *****
Average Distance: 86.55
Davies Bouldin Index: .96
Normalized Mutual Information: 1
Press any key to exit...

```

Na osnovu rezultat s prethodne slike: Average Distance od 86.55 ukazuje da su tačke u klasterima blizu njihovom centru. Davies Bouldin Index od 0.96 ukazuje na razdvajanje između klastera i kompaktnost unutar klastera, koje se može poboljšati. Normalized Mutual Information ukazuje da su klasteri slični stvarnim klasama ako su dostupni.

## ZADAĆA 5 – LOGISTIČKA MULTINOMIJALNA REGRESIJA

Multinomijalna logistička regresija je vrsta logističke regresije koja se koristi kada imamo više od dve kategorije za predviđanje.

Prvo smo odabrali upit iz prethodnih upita za predikciju:

- **UPIT:** Predikcija interesovanja studenta – na osnovu ocjena studenta i datuma polaganja, može se predvidjeti interesovanje i sposobnost studenta za određenu oblast, te ukazati na sklonost ka toj naučnoj disciplini.

Zatim, odabrali smo metodu:

- **METODA:** Metoda koja se koristi za ovaj upit je multinomijalna logistička regresija, obzirom da se predviđa interesovanje i sposobnosti studenta.

Potrebno je odrediti ulazne podatke i ciljnu varijablu, tj vrijednost koju želimo predviditi. Odabirom modela pisemo algoritam ili matematičku funkciju koja će napraviti predviđanje.

Obzirom da nemam konkretnu ciljanu varijablu tj interesovanja studenta, konstruirano je intresovanje na osnovu dostupnih podataka. Interesovanje se konsturisalo, na osnovu grupisanja predmeta po određenim oblastima.

Prvo, se uzimaju prečišćeni podaci iz zadatke 3, odabrati kolone potrebne za predviđanje interesovanja studenta. Odabrana je prva godina studija, tj predmeti s prve godine studija s ocjenama (**P-219, P-221, P-175, P-176, P-149, P-12, P-246, P-222, P-220, P-218, P-142, P-58**).

Dodatne informacije pored studentId su

Godina	DatumUpisa	SrednjaSkola	GodinaZavrsetka	NacinStudiranja	Diplomirao	Spol
--------	------------	--------------	-----------------	-----------------	------------	------

Da bi se konstruisalo interesovanje na osnovu gurpisanja predmeta po određenim oblastima, potrebno je znati ECTS bodove za svaki predmet, datum polaganja predmeta, te prosječna ocjena za svaki predmet koji se dobija dijeljenjem ocjena i ECTS-ova.

Primjer:

<b>P-104</b>	ECTS_P-104	PolozioDatum_P-104	ProsjecnaOcjena_P-140
--------------	------------	--------------------	-----------------------

```

import pandas as pd
import numpy as np

# Učitavanje podataka iz CSV fajla
data = pd.read_csv('18178812_MedisaSatara_podaci_ME.csv', encoding='latin1')

# Definisanje grupa predmeta (ovo je samo primer, prilagodite prema stvarnim podacima)
predmeti_grupe = {
    'Matematika': ['P-219', 'P-221'],
    'Programiranje': ['P-175', 'P-176', 'P-149'],
    'Engleski jezik': ['P-12', 'P-246'],
    'Ekonomija': ['P-222', 'P-220'],
    'Informatika': ['P-218'],
    'Mreza': ['P-142', 'P-58']
}

# Računanje prosečnih ocena po grupama
for grupa, predmeti in predmeti_grupe.items():
    ocene_kolone = [f'ProsječnaOcena_{predmet}' for predmet in predmeti]

    # Osiguravanje da su sve vrednosti numeričke
    for kolona in ocene_kolone:
        data[kolona] = pd.to_numeric(data[kolona], errors='coerce')

    data[f'ProsječnaOcena_{grupa}'] = data[ocene_kolone].mean(axis=1)

# Definisanje interesovanja kao oblast sa najvišom prosečnom ocenom
# Zamenjujemo NaN vrednosti pre nego što primenimo idmax
prosječne_ocene_grupe = [f'ProsječnaOcena_{grupa}' for grupa in predmeti_grupe.keys()]
data[prosječne_ocene_grupe] = data[prosječne_ocene_grupe].fillna(-1)
definirano_interesovanje = data[prosječne_ocene_grupe].idxmax(axis=1)
data['Interesovanje'] = definirano_interesovanje.apply(lambda x: x.split('_')[1])

# Sada imamo kolonu 'Interesovanje' koja se može koristiti kao ciljna varijabla
print(data[['StudentId', 'Interesovanje']].head(40))

# Kodiranje kategorijalnih varijabli
label_encoders = {}
for column in data.select_dtypes(include=['object']).columns:
    label_encoders[column] = LabelEncoder()
    data[column] = label_encoders[column].fit_transform(data[column])

# Selektovanje karakteristika i ciljne varijable
X = data.drop(columns=['Interesovanje']) # Navesti sve kolone osim ciljne varijable
y = data['Interesovanje']

# Podela podataka na trening i test skupove
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Treniranje multinomijalne logističke regresije
model = LogisticRegression(multi_class='multinomial', solver='lbfgs', max_iter=200)
model.fit(X_train, y_train)

# Predikcija na test skupu
y_pred = model.predict(X_test)

# Evaluacija modela
print(classification_report(y_test, y_pred))
print("Tačnost modela:", accuracy_score(y_test, y_pred))

# Dodatno: Prikaz koeficijenata modela
coefficients = pd.DataFrame(model.coef_, columns=X.columns)
print(coefficients)

```

Predmeti su definisani u sljedeće grupe:

Matematika:( **Matematika I P-219, Matematika II P-221**)  
Programiranje:(**Programiranje I P-175, Programiranje II P-176, Web razvoj i dizajn P-149**)  
Engleski jezik:(**Engleski jezik I P-12, Engleski jezik II P-246**)  
Ekonomija:( **Uvod u marketing P-222, Digitalna logika P-220**)  
IT:( **Računarstvo i informatika P-218**)  
Mreze:(**Arhitektura računara P-142, Operativni sistemi P-58**)

U kodu je navedeno da ispise 40 prvih, koji imaju predviđena interesovanja.

Prvi slučaj prolazi kroz sve oblasti, gdje se vidi da najveću prosječnu ocjenu student generalno imaju iz engleskog jezika, te je interesovanje studenata prve godine usmjereno ka engleskom jeziku.

Rezultati dobijeni nakon ovog prvog slučaja su:

	StudentId	Interesovanje
0	16320.0	Engleski jezik
1	16321.0	Engleski jezik
2	16322.0	Engleski jezik
3	16323.0	Engleski jezik
4	16324.0	Engleski jezik
5	16325.0	Engleski jezik
6	16326.0	Engleski jezik
7	16327.0	Engleski jezik
8	16328.0	Engleski jezik
9	16329.0	Engleski jezik
10	16330.0	Engleski jezik
11	16331.0	Engleski jezik
12	16332.0	Engleski jezik
13	16333.0	Engleski jezik
14	16334.0	Engleski jezik
15	16335.0	Engleski jezik
16	16336.0	Engleski jezik
17	16337.0	Engleski jezik
18	16338.0	Engleski jezik
19	16339.0	Engleski jezik
20	16340.0	Engleski jezik
21	16341.0	Engleski jezik
22	16342.0	Engleski jezik
23	16343.0	Engleski jezik
24	16344.0	Engleski jezik
25	16345.0	Engleski jezik
26	16348.0	Engleski jezik
27	16349.0	Engleski jezik
28	16350.0	Engleski jezik
29	16351.0	Engleski jezik
30	16352.0	Engleski jezik
31	16353.0	Engleski jezik
32	16354.0	Engleski jezik
33	16355.0	Engleski jezik
34	16356.0	Engleski jezik
35	16357.0	Engleski jezik
36	16358.0	Engleski jezik
37	16359.0	Engleski jezik
38	16360.0	Engleski jezik
39	16361.0	Engleski jezik

Drugi slučaj, uključuje oblasti iz struke, dakle izuzela se oblast engleski jezik. Prikazana su interesovanja studenata iz oblasti iz struke.

	StudentId	Interesovanje
0	16320.0	Informatika
1	16321.0	Mreze
2	16322.0	Informatika
3	16323.0	Ekonomija



4	16324.0	Ekonomija
5	16325.0	Ekonomija
6	16326.0	Ekonomija
7	16327.0	Ekonomija
8	16328.0	Ekonomija
9	16329.0	Informatika
10	16330.0	Informatika
11	16331.0	Informatika
12	16332.0	Informatika
13	16333.0	Informatika
14	16334.0	Informatika
15	16335.0	Mreže
16	16336.0	Informatika
17	16337.0	Informatika
18	16338.0	Ekonomija
19	16339.0	Ekonomija
20	16340.0	Informatika
21	16341.0	Informatika
22	16342.0	Informatika
23	16343.0	Informatika
24	16344.0	Informatika
25	16345.0	Informatika
26	16348.0	Informatika
27	16349.0	Informatika
28	16350.0	Informatika
29	16351.0	Informatika
30	16352.0	Informatika
31	16353.0	Ekonomija
32	16354.0	Informatika
33	16355.0	Mreže
34	16356.0	Informatika
35	16357.0	Informatika
36	16358.0	Informatika
37	16359.0	Informatika
38	16360.0	Mreže
39	16361.0	Informatika

## LITERATURA

[DEUSCHLE-SENIORTHESIS-2019.pdf \(harvard.edu\)](#)

[MLBOOK.pdf \(stanford.edu\)](#)

[Machine learning, explained | MIT Sloan](#)

## SADRŽAJ

OPIS PROBLEMA .....	2
ZADAĆA 1 – Priprema podataka.....	2
ZADAĆA 2 – UPITI ZA PREDIKCIJU .....	2
ZADAĆA 3 – DORADA ZADAĆE 1 (PRIPREMA PODATAKA) .....	3
ZADAĆA 4 – DESKRIPTIVNE METODE .....	3
ZADAĆA 5 – LOGISTIČKA MULTINOMIJALNA REGRESIJA.....	5
LITERATURA .....	9