**SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU**

**FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK**

**Diplomski studij**

**PREDVIĐANJE TIPA RAKA DOJKE**

**SEMINARSKI RAD**

**RAČUNARSTVO USLUGA I ANALIZA PODATAKA**

**Tomislav Cajbert**

**Matej Milić**

**Osijek, 2019.**

**SADRŽAJ**

[1. UVOD 1](#_Toc12125495)

[2. OPIS 2](#_Toc12125496)

[3. KORIŠTENI PODACI 3](#_Toc12125497)

[4. KORIŠTENI POSTUPCI STROJNOG UČENJA 4](#_Toc12125498)

[5. MODEL STROJNOG UČENJA 5](#_Toc12125499)

[5.1. Planiranje i izrada modela 5](#_Toc12125500)

[5.2. Testiranje i evaluacija dobivenih rezultata 6](#_Toc12125501)

[5.3. Izrada web servisa i korištenje API-a 8](#_Toc12125502)

[5.4. Testiranje izmijenjenog skupa podataka 11](#_Toc12125503)

[6. OPIS PROGRAMSKOG RJEŠENJA 13](#_Toc12125504)

[6.1. Forma za unos parametara i prikaz rezultata 13](#_Toc12125505)

[6.2. Prikaz podataka 14](#_Toc12125506)

[7. ZAKLJUČAK 16](#_Toc12125507)

[LITERATURA 17](#_Toc12125508)

[POVEZNICE 17](#_Toc12125509)

# UVOD

U ovom projektu obrađuje se problem dijagnosticiranja raka dojke pomoću modela strojnog učenja. Aplikacija treba omogućiti unos parametar koji se dobivaju iz digitalizirane slike aspiracijske biopsije tankom iglom, te na temelju unesenih parametara predvidjeti dali je tumor benigni ili maligni i kolika je vjerojatnost da je predviđanje točno.

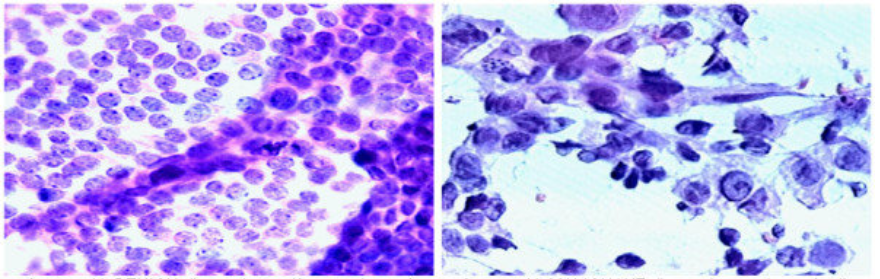
Osim mogućnosti predviđanja aplikacije treba omogućiti grafički prikaz već testiranih vrijednosti koji se uzimaju iz baze podataka na temelju kojih se mogu očitati parametri i vidjeti krajnje predviđanje.

U seminaru će biti opisani korišteni skup podataka, postupci strojnog učenja, izrada modela, objavljivanje web servisa te opis programskog rješenja.

# OPIS

U cijelome svijetu rak dojke je jedan od vodećih uzroka smrti među ženama. U Hrvatskoj će svaka 10. žena oboljeti, a sve je veći broj žena mlađe dobi (čak i mlađih od 30 godina). Kako ne postoji način da bolest spriječimo, jedini je cilj otkriti je u početnoj fazi. [1]

Aspiracijska biopsija tankom iglom je jedan od najbržih i najučinkovitijih tehnika za dijagnosticiranje ove bolesti. Ovom tehnikom tankom iglom uzima se uzorak iz dojke, te se digitalizira slika obrisa (slika 2.1.)



**Slika 2.1. –** Prikaz digitalizirane slike obrisa, lijevo benigni tumor, desno maligni

O ovom projektu biti će korišteni parametri očitani iz digitaliziranih slika, pomoću kojih se rezultat može klasificirati u dva razreda: benigni i maligni. Parametri koji će se koristiti za klasifikaciju su:

* Radijus (srednja vrijednost udaljenosti od centra do točaka na opsegu)
* Tekstura (standardna devijacija vrijednosti sivih-tonova)
* Opseg
* Područje
* Glatkost (lokalna varijacija u duljini radijusa)
* Zbijenost (opseg2 / područje -1)
* Konkavnost (ozbiljnost konkavnih dijelova na konturi)
* Simetrija
* Fraktalna dimenzija

# KORIŠTENI PODACI

Kako bi aplikacija uspješno obavljala svoje zadatke, potrebo je izraditi model strojnog učenja koji će na temelju podataka koji su mu predani moći predvidjeti, odnosno klasificirati rezultat.

Najvažniji dio kod izrade model strojnog učenja je pronalazak i pred obrada podataka na temelju kojih će model biti treniran. Ukoliko u odabranom skupu podataka postoje parametri koji su nepotrebni ili koji će negativno utjecati na ishod potrebno ih je izbaciti iz podatkovnog skupa.

Podatkovni skup koji će biti korišten uzet je UCI *Machine learning repository-a*. Ime skupa je „*Breast Cancer Wisconsin (Diagnostic) Data Set“* i sastoji se od 569 instanci od kojih su 357 benigni i 212 maligni tumor. Podatci su prikupljeni 1992. godine, a donirani 1995. godine. Izgled instance prikazan je na slici 2.



**Slika 3.1. –** Izgled jedne instance

U dobiven podacima osim navedenih parametara koji će se koristiti za klasifikaciju, postoje 2 dodatna parametra. Prvi je *ID* instance koji će biti potrebno ukloniti, a drugi je klasa koja može imati dvije vrijednosti:

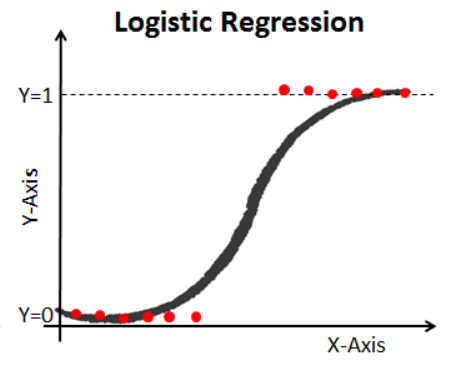
* 2 - benigni
* 4 - maligni

Vrijednosti klase će prilikom obrade podataka biti pretvorene u 0 i 1.

# KORIŠTENI POSTUPCI STROJNOG UČENJA

Nakon što je pronađen odgovarajući skup podataka i odrađena obrada potrebno je odabrati postupak strojnog učenja koji će biti pogodan. Pošto je u ovom zadatku potrebno odrediti jeli tumor benigni ili maligni, odnosno postoje samo dva moguća rezultata potrebno je koristiti binarni klasifikator. Nadalje pošto u dobivenom skupu podataka već imamo rezultat svake instance, odnosno njenu klasu koristiti će se binarni klasifikator *Two-Class Logistic Regression*.

*Two-Class Logistic Regression* je dobro poznata statistička metoda koja služi za predviđanje vjerojatnosti nekog rezultata. Algoritam predviđa događaj umetanjem podataka u logističku funkciju (slika 4.1.). Optimiziran je za predviđanje binarnih varijabli, dok se za klasificiranje višestrukih rezultata koristi *Multi-Class Logistic Regression*. [2]



**Slika 3.1. –** Logistička regresija [3]

# MODEL STROJNOG UČENJA

## Planiranje i izrada modela

Na Azure web platformi napravljen je novi eksperiment u kojem je izrađen model strojnog učenja. Model će na temelju podataka i odabrane klasifikacije predviđati rezultate i njihovu vjerojatnost. Kao što je već ranije napisano odabrana klasifikacija je *Two-Class Logistic Regression*.

Prvi korak u izradi modela je dodavanje skupa podataka koji će biti korišten. U ovom slučaju skup podataka naziva se „breast-cancer.wisconsin.rtf“, te je zapisana u CSV (engl. *Comma-Separated Values*) formatu.

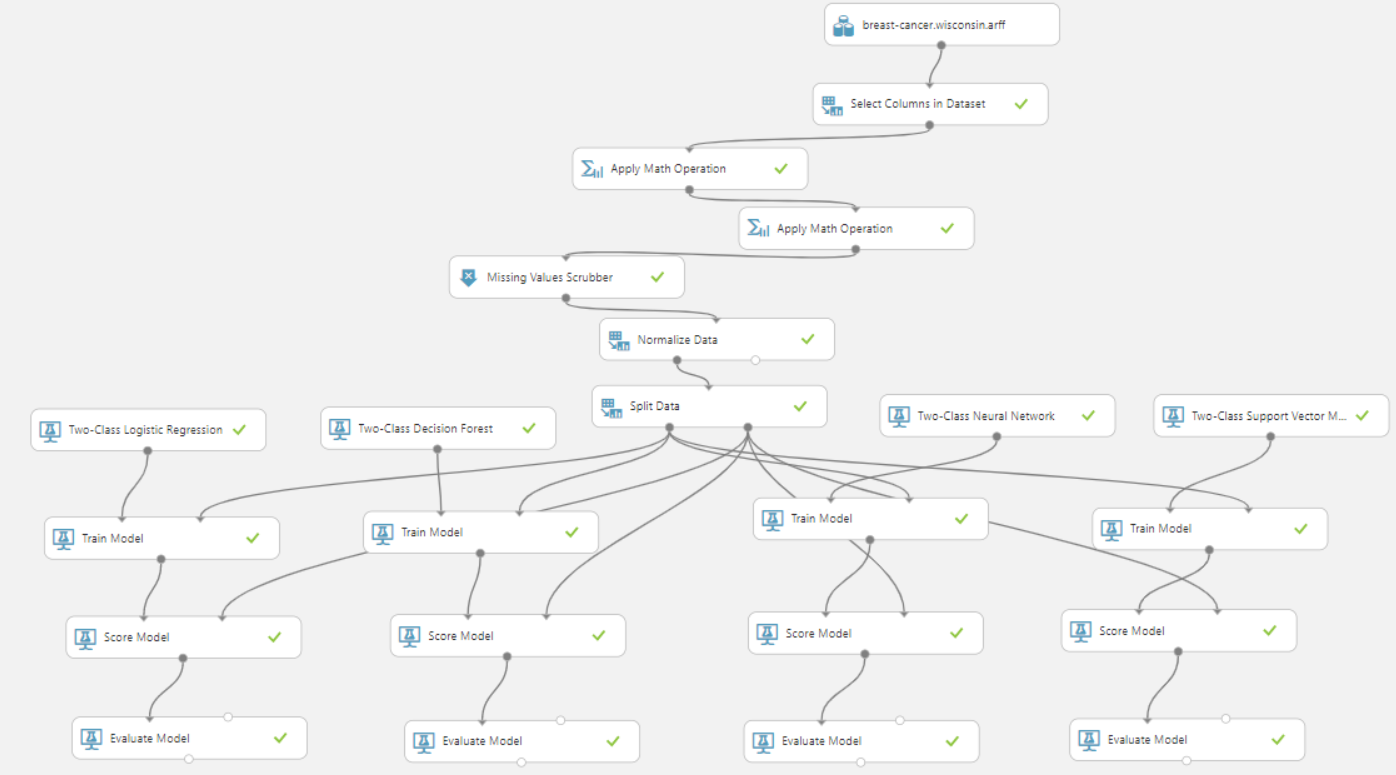
Nakon dodavanja skupa podataka potrebno ga je očistiti i urediti. Kao što je već spomenuto uklonjen će biti stupac koji predstavlja ID instance. Uz to potrebno je napraviti matematičku operaciju dijeljenja na stupcu koji predstavlja klasu instance. U početnom skupu podataka benigni tumor bio je označen sa „2“, dok je maligni bio označen sa „4“. Matematička operacija će vrijednosti podijeliti sa 2, te od dobivenog rezultata oduzeti 1 tako da će krajnje vrijednosti biti „0“ za benigni i „1“ za maligni tumor.

Dalje, podaci su normalizirani i podijeljeni na treniranje i testiranje. 80% podataka biti će odvojeno za treniranje, dok će 20% biti za testiranje kako bi znali koji je model najprecizniji.

U svrhu testiranja osim logističke regresije u eksperimentu će biti korištene još tri vrste binarne klasifikacije:

* *Two-Class Decision Forest*
* *Two-Class Neural Network*
* *Two-Class Support Vector Machine* (SVM)

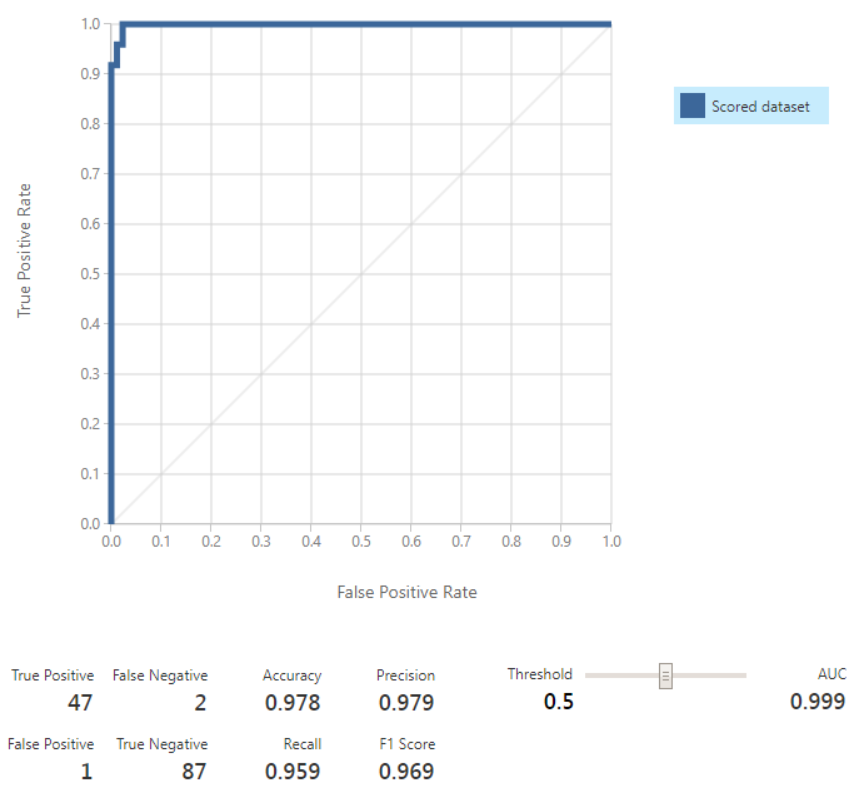
Konačni izgled modela prikazan je na slici 5.1.



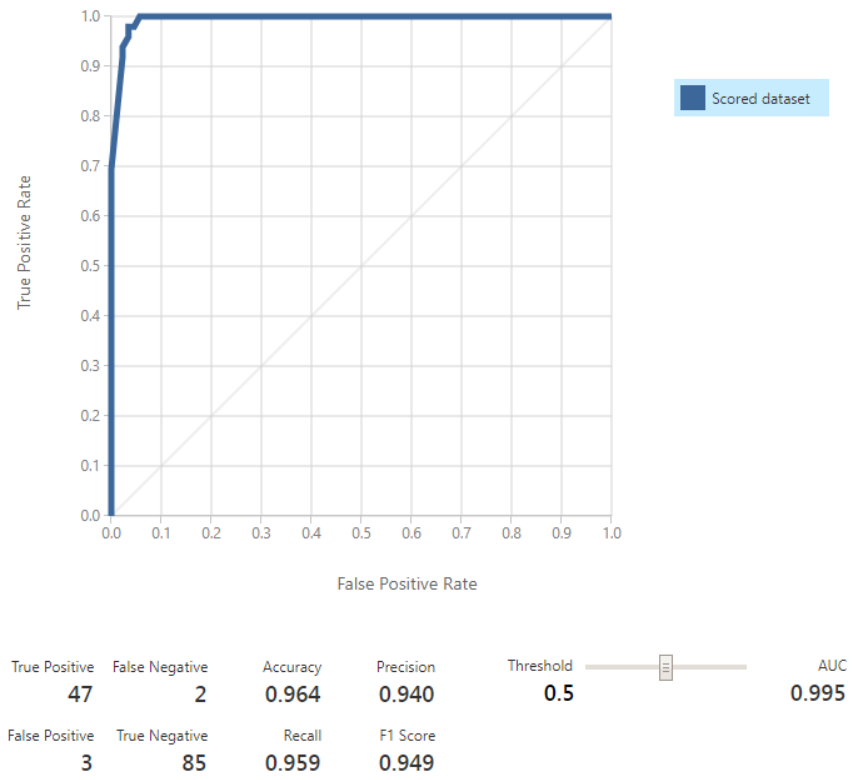
**Slika 5.1. –** Konačni izgled modela

## Testiranje i evaluacija dobivenih rezultata

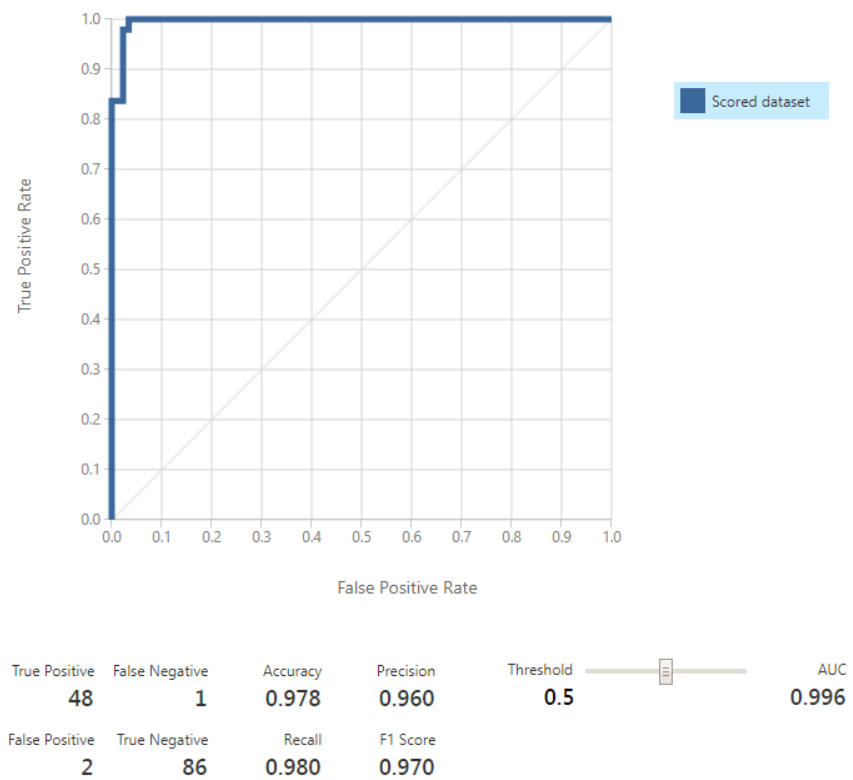
Nakon pokretanja azure servis je obavio testiranje i dobiveni su rezultati. Svi algoritmi klasifikacije koji su korišteni u ovom testu su dali vrlo visoke postotke preciznosti i točnosti. Ipak za vrlo malo iznad svih je logistička regresija, kao što je i pretpostavljeno. Rezultati logističke regresije prikazani su na slici 5.2. dok su rezultati ostalih klasifikacija prikazani na slikama 5.3.-5.5.



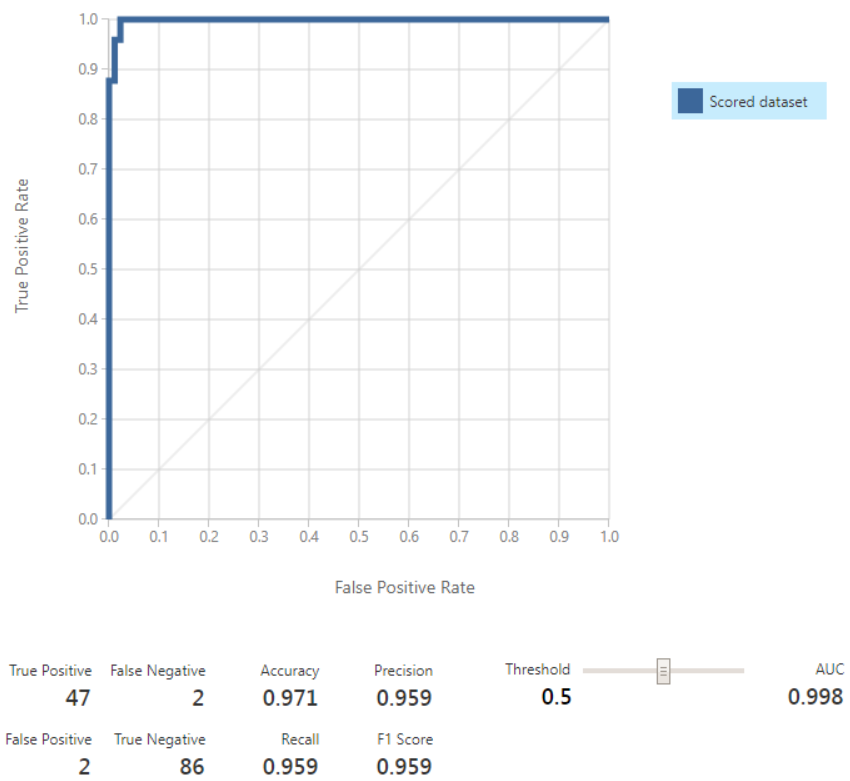
**Slika 5.2. –** Rezultat za *Two-Class Logistic Regression*



**Slika 5.3. –** Rezultat za *Two-Class Decision Forest*



**Slika 5.4. –** Rezultat za *Two-Class Neural Network*

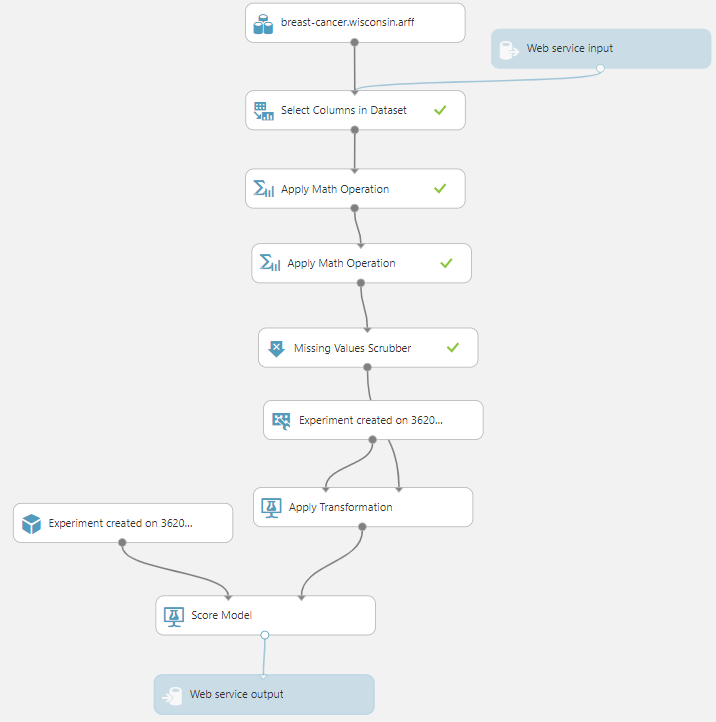


**Slika 5.5. –** Rezultati za SVM

Na primjeru logističke regresije je vidljivo da je od 48 testiranih *positive label* podataka čak njih 47 točno klasificirano, dok je za 89 testiranih *negative label* podataka njih 87 točno klasificirano. Pošto je u ulaznim vrijednostima već bila definirana klasa svake instance model se može samo ocjenjivati na temelju tih vrijednosti. Što se tiče preciznosti i točnosti vidljivo je da su vrijednosti vrlo visoke iz čega se zaključuje da je skup podataka veoma dobar.

## Izrada web servisa i korištenje API-a

Nakon što je model istreniran, potrebno je napraviti web servis koji će preko API-a primati parametre, klasificirati rezultat i vratiti ga kao odgovor. Za izradu web servisa potrebno je odabrati opciju „*Train web service“* pomoću koje se stvara prediktivni eksperiment. (slika 5.6.)



**Slika 5.6. –** Izgled prediktivnog eksperimenta

Nakon što je izrađen prediktivni eksperiment, potrebno ga je pokrenuti i zatim odabrati opciju *„Deploy web service“*. Web servisu se pristupa pomoću dobivenog Request URI te API ključa. Uz to je potreban odgovarajuća JSON forma za zahtjev (programski kod 5.1.) na koji se dobiva JSON odgovor (programski kod 5.2.) JSON je transportni format podatka koji omogućuje jednostavno slanje i primanje podataka putem API-a.

**Programski kôd 5.1. –** Primjer zahtjeva

{

"Inputs": {

"input1": {

"ColumnNames": [

"id",

"clumpThickness",

"uniformityCellSize",

"uniformityCellShape",

"marginalAdhesion",

"singleEpithelialCellsize",

"bareNuclei",

"blandChromatin",

"normalNucleoli",

"mitoses",

"class"

],

"Values": [

[

"0",

"0",

"0",

"0",

"0",

"0",

"0",

"0",

"0",

"0",

"0"

],

[

"0",

"0",

"0",

"0",

"0",

"0",

"0",

"0",

"0",

"0",

"0"

]

]

}

},

"GlobalParameters": {}

}

**Programski kôd 5.2. –** Primjer odgovora

{

"Results": {

"output1": {

"type": "DataTable",

"value": {

"ColumnNames": [],

"ColumnTypes": [],

"Values": [

[],

[]

]

}

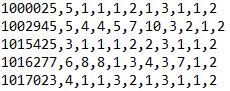
}

}

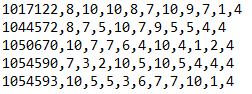
}

## Testiranje izmijenjenog skupa podataka

U svrhu testiranja napravljen je novi eksperiment, u kojemu je iz korištenog skupa podataka izbačeno 5 instanci za benigni i 5 instanci za maligni tumor. Ponovno je izrađen web servis na kojemu će se testirati izbačene vrijednosti i dobiveni rezultati će se usporediti sa stvarnima. Izbačene instance vidljive su na slikama 5.3. i 5.4.



**Slika 5.3. –** Izbačene instance benignog tumora



**Slika 5.4. –** Izbačene instance malignog tumora

Nakon što su sve pripreme obavljene, ručno su testirane izbačene instance. Rezultati su prikazani u matrici konfuzije koja je vidljiva u tablici 5.1.

|  |  |  |
| --- | --- | --- |
| n = 10 | Predviđeni: Benigni | Predviđeni: Maligni |
| Stvarni: Benigni | 3 | 2 |
| Stvarni: Maligni | 0 | 5 |

**Tablica 5.1. –** Matrica konfuzije

Iz matrice konfuzije je vidljivo da je od 5 benignih instanci točno predviđeno 3, dok su dvije instance krivo predviđene. Sa druge strane, od 5 instanci koje su maligne sve su točno klasificirane.

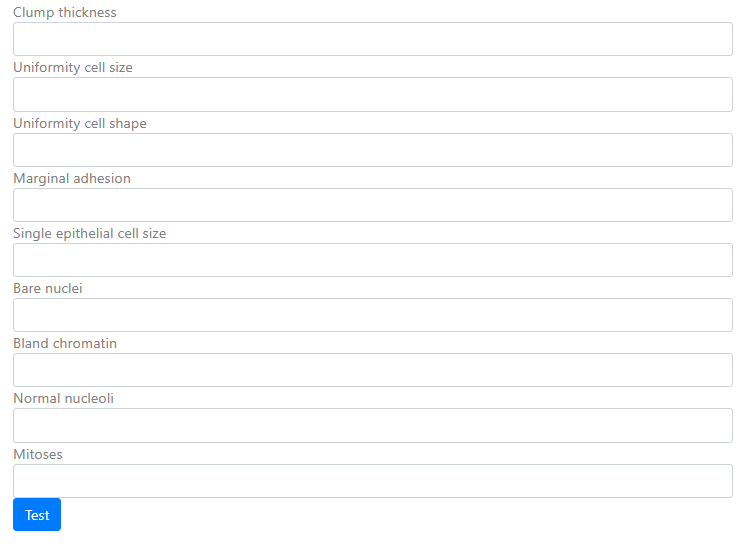
Iz dobivenih rezultata možemo zaključiti da su za benigne instance uzete vrijednosti koje se dosta razlikuju od vrijednosti nad kojima je model treniran i zbog toga su samo 3 od 5 točno klasificirane. Za maligne su uzete vrijednosti slične onima koje se već nalaze u skupu podataka pa su rezultati puno bolji.

# OPIS PROGRAMSKOG RJEŠENJA

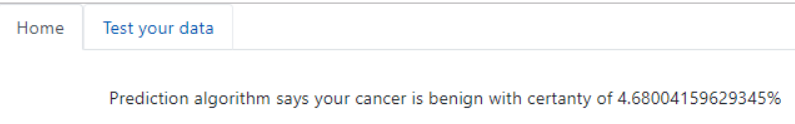
Nakon izrade modela strojnog učenja i objave web servisa potrebno je izraditi aplikaciju za testiranje. Aplikacija treba omogućiti unos parametara koji će biti testirani, te ispis rezultata u obliku određene klase i vjerojatnosti. Uz to biti će napravljen grafički prikaz svih podataka iz podatkovnog skupa, te podataka koje će korisnici unositi. Podaci će biti spremljeni u bazi podataka, a prikazivat će se pomoću grafa s paralelnim koordinatama. Paralelne koordinate su odličan način za vizualizaciju podataka sa više varijabli. U našem slučaju biti će idealne za prikaz dobivenih podataka iz kojih će se za svaku instancu moći prikazati i očitati parametri na efikasan način.

## Forma za unos parametara i prikaz rezultata

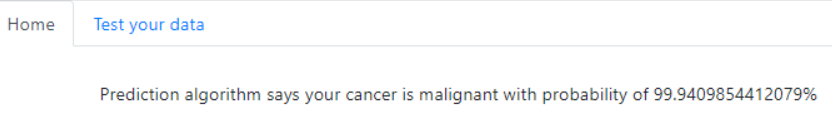
Za unos parametara napravljena je forma (slika 6.1.) u kojoj se unose potrebne vrijednosti, te se pritiskom da tipku „Test“ podaci šalju web servisu na obradu. Dobiveni rezultat ispisuje dali je tumor benigni ili maligni, te vjerojatnost odgovora. Ispis pozitivnog i negativnog rezultata vidljiv je na slikama 6.2 i 6.3.



**Slika 6.1. –** Izgled forme za unos podataka



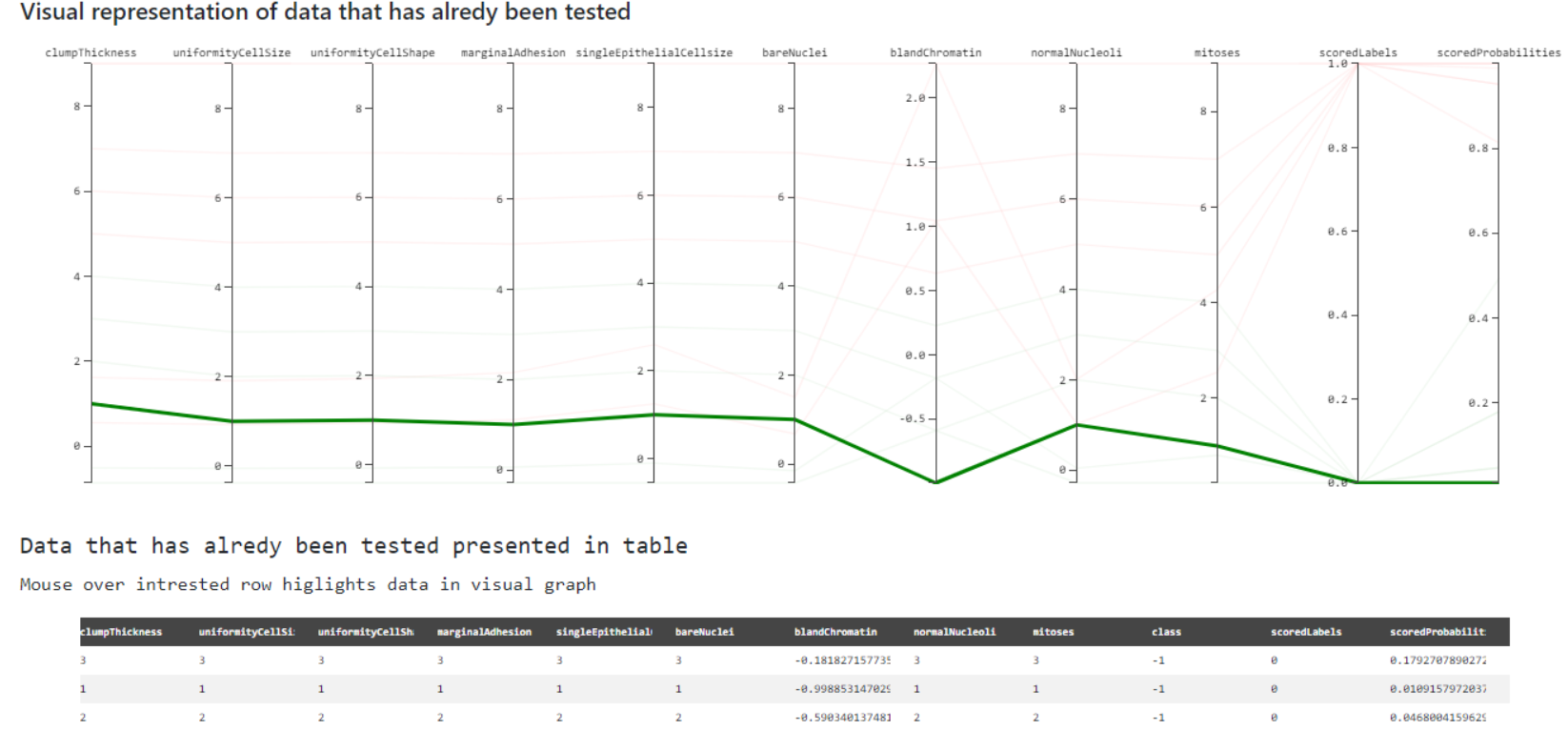
**Slika 6.2. –** Izgled dobivenog rezultata za benigni tumor



**Slika 6.3. –** Izgled dobivenog rezultata za maligni tumor

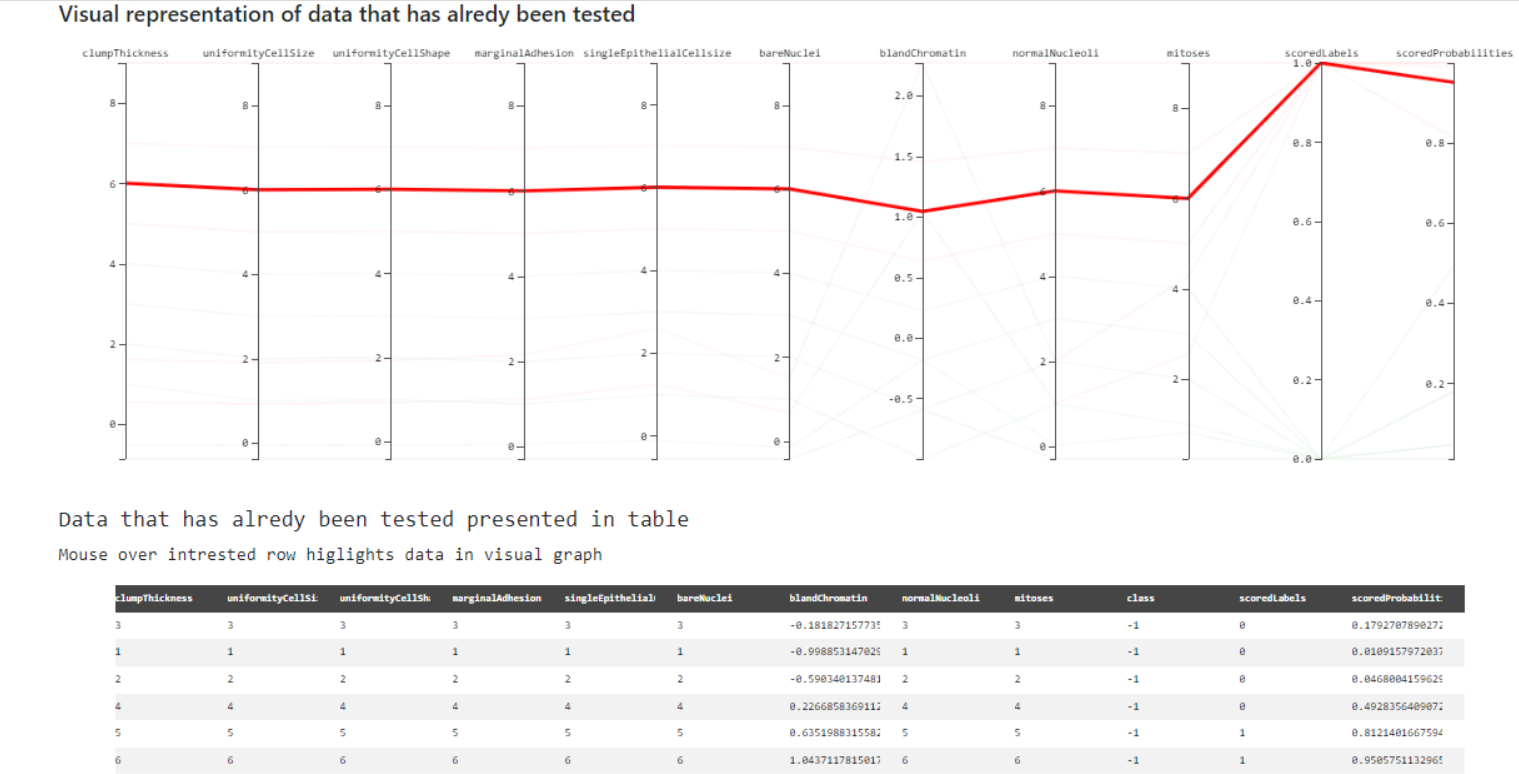
## Prikaz podataka

Kao što je već navedeno za prikaz se koristi graf sa paralelnim koordinatama. Na grafu su vidljive vertikalne koordinatne osi koje predstavljaju parametre iz podatkovnog skupa. Zadnja vertikalna os predstavlja klasu. Na grafu je moguće označiti svaku instancu iz podatkovnog skupa i vidjeti njene parametre te u koji je razred klasificirana. Izgled grafa za instancu koja je klasificirana kao benigni tumor vidljiv je na slici 6.4.



**Slika 6.4. –** Prikaz grafa sa paralelnim koordinatama za benigni tumor

Vidljiva je vrijednost za svaki parametar instance i klasa koja može biti 0 ili 1. Primjer grafa za instancu koja je klasificirana kao maligni tumor prikazan je na slici 6.5.



**Slika 6.5. –** Prikaz grafa sa paralelnim koordinatama za maligni tumor

# ZAKLJUČAK

Strojno učenje, iako veoma star postupak obrade informacija sve je popularniji u modernom vremenu. Postoje mnoge mogućnosti i primjene, a jedna od njih je opisana u ovom radu. Koristeći skup podataka sa popularnog repozitorija UCI, napravljen je i istreniran model koji na temelju ulaznih parametara predviđa dali je tumor dojke benigni ili maligni.

Na temelju istreniranog modela izrađeni su web servis i aplikacija koja ga koristi. Aplikacija u vrlo brzom roku može odrediti rezultata za testirane parametri i vjerojatnost tog rezultata, te uz to prikazuje već testirane podatke koristeći graf sa paralelnim koordinatama koji se pokazao kao idealan prikaz za ovaj problem.

Aplikacija bi mogla imati nekoliko primjena, a najočitija je da doktorima pomogne u preliminarnom dijagnozama pacijenata, kako bi se potrebne mjere liječenja pacijenata poduzele u što bržem roku.

# LITERATURA

1. <http://hlpr.hr/rak/vijest/rak-dojke>
2. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-logistic-regression>
3. <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>

# POVEZNICE

[1] <https://github.com/tomi-cro/RUAP_Predvidanje_tipa_raka_dojke>