**SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU**

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

**Diplomski studij**

**PREDVIĐANJE KVALITETE ČOKOLADE**

**SEMINARSKI RAD**

**RAČUNARSTVO USLUGA I ANALIZA PODATAKA**

**Željko Sabo**

**Osijek, 2019.**

**SADRŽAJ**

[1. UVOD 1](#_Toc17118639)

[2. OPIS 2](#_Toc17118640)

[3. KORIŠTENI PODACI 3](#_Toc17118641)

[4. KORIŠTENI POSTUPCI STROJNOG UČENJA 4](#_Toc17118642)

[5. MODEL STROJNOG UČENJA 5](#_Toc17118643)

[5.1. Planiranje i izrada modela 5](#_Toc17118644)

[5.2. Testiranje i evaluacija dobivenih rezultata 6](#_Toc17118645)

[5.3. Izrada web servisa i korištenje API-a 10](#_Toc17118646)

[5.4. Testiranje izmijenjenog skupa podataka 13](#_Toc17118647)

[6. OPIS PROGRAMSKOG RJEŠENJA 14](#_Toc17118648)

[6.1. Forma za unos parametara i prikaz rezultata 14](#_Toc17118649)

[7. ZAKLJUČAK 16](#_Toc17118650)

[LITERATURA 17](#_Toc17118651)

[POVEZNICE 17](#_Toc17118652)

# UVOD

U ovom projektu obrađuje se problem predviđanja kvalitete čokolade pomoću modela strojnog učenja. Aplikacija treba omogućiti unos parametara na temelju kojih se čokolada, ovisno o parametrima, klasificira u jednu od tri kategorije a to su loša, dobra i izvrsna čokolada.

Pošto je čokoladu potrebno svrstavati u klase ovisno o njenoj kvaliteti, u ovom projektu je korištena metoda klasifikacije pomoću koje se čokolada ovisno o svojoj kvaliteti svrstava određenu klasu.

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

# OPIS

U cijelome svijetu čokolada je najpopularnija slastica koja se konzumira. Najčešće se konzumira u obliku čokoladnih pločica kao „obična“ čokolada, no često se koristi i u kuhinji, prilikom spremanja različitih poslastica kao što su na primjer kolači.

Naravno, postoje različiti brendovi čokolade, te sukladno tome, nisu svi brendovi jednako dobri. Čokolade mogu imati različite postotke kakaa u sebi, mogu isto tako biti i mliječne te samim tim u sebi sadržavati određenu količinu mlijeka, no isto tako mogu i da sadrže mnoge druge dodatke poput karamela, kikirikija, keksa i brojnih drugih. Također ne koriste svi proizvođači ista zrna kakaa i nisu svaka zrna jednako dobra, te se ovaj projekt bazira na prepoznavanju i klasifikaciji čokolade na osnovu tih, i brojnih drugih parametara.



**Slika 2.1. –** Prikaz različitih brendova čokoladnih pločica

O ovom projektu biti će korišteni parametri koje korisnik sam unosi, te pomoću kojih se rezultat može klasificirati u (kao što je prethodno spomenuto) jednu od tri klase (razreda) a to su loše čokolade, dobre čokolade i izvrsne čokolade. Parametri koji će se koristiti za klasifikaciju su:

* Naziv tvrtke
* Mjesto podrijetla zrna (grad ili pokrajina)
* REF (godina kada je čokolada prvi put evaluirana)
* Godina evaluacije (godina kada je zadnji put čokolada evaluirana)
* Postotak kakaa (vrijednost od 0 do 1)
* Država u koju je smještena kompanija
* Ocjena (predstavlja dobivenu ocjenu na zadnjoj evaluaciji)
* Vrsta zrna
* Država podrijetla zrna

# 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 prilikom izrade modela strojnog učenja je bio pronalazak i pred obrada podataka na temelju kojih će se model trenirati. Sve parametre koji su na bilo koji način mogli negativno utjecati na ishod su izbačeni kako bi trenirani model radi što točnije.

Podatkovni skup koji je korišten je preuzet sa internetske stranice Kaggle, te ga je moguće pronaći na linku koji se nalazi u poveznici [1]. Skup podataka se sastoji od 1795 instanci od kojih su izbačene sve gdje je falio jedan od prethodno navedenih parametara kako bi trenirani model radio što točnije. Korišteni set podataka je objavljen 2017. godine od strane Rachael Tatman koja je znanstvenik za podatke u Kaggle-u.



**Slika 3.1. –** Izgled jedne instance

U dobiven podacima osim navedenih parametara koji će se koristiti za klasifikaciju, postoji jedan dodatan parametar koji se nije koristio, a to je redi broj instance koji je izbačen jer je taj podatak nepotreban, te samim tim suvišan.

# KORIŠTENI POSTUPCI STROJNOG UČENJA

Nakon što su pronađen odgovarajući skup podataka te je odrađena obrada korištenog skupa podataka, za treniranje modela je bilo potrebno je odabrati postupak strojnog učenja koji će biti najpogodniji. Pošto je u ovom projektu bilo potrebno odrediti jeli čokolada loša, dobra ili izvrsna, odnosno postoje samo tri moguća rezultata, kao postupak klasifikacije odabrana je *Multiclass Decision Forest* jer su sa njom postignuti najbolji rezultati prilikom klasifikacije.

*Multiclass Decision Forest* metoda klasifikacije pomoću koje se treniraju modeli ukoliko je potrebno imati više od dvije različite klase. Klasifikacija predstavlja metodu učenja pomoću koje se određuje kategorija, tip ili klasa jednog ili više redova podataka. Najčešća podjela klasifikacija je na to vrše li se klasifikacije na dvije ili više klasa. Ukoliko se klasifikacija vrši na dvije klase tada se radi o binarnoj klasifikaciji. Jednostavan primjer binarne klasifikacije bi bio ukoliko bi se klasifikacija kvalitete čokolade provodila tako da se određuje samo je li čokolada dobra ili loša, no ako je riječ o više od dvije klase kao što je u ovom projektu, onda se radi o više klasnoj klasifikaciji. [2]

# 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 metodu klasifikacije odabrana je *Multiclass Decision Forest* metoda.

Kao prvi korak u kreiranju modela potrebno je dodati skup podataka na osnovu kojeg će model biti treniran. U ovom slučaju skup podataka naziva se „flavors\_of\_cacao.csv“, te je zapisana u CSV (engl. *Comma-Separated Values*) formatu.

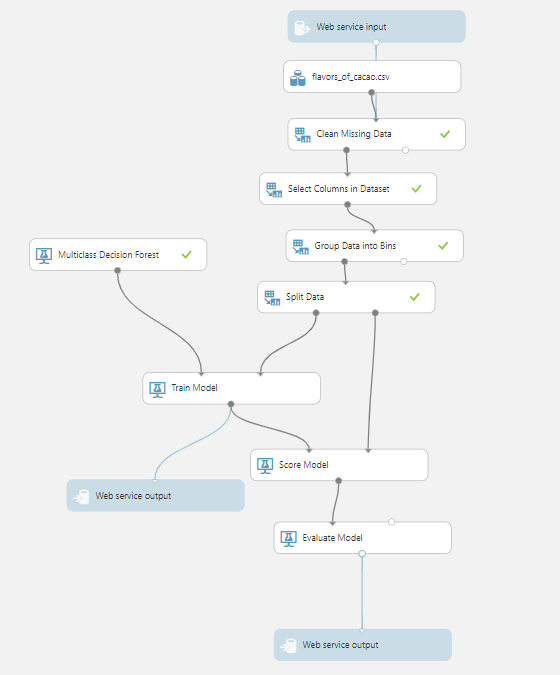
Nakon dodavanja skupa podataka sljedeće što je bilo potrebno napraviti je očistiti i urediti podatke. Kao što je ranije spomenuto uklonjeni su svi redci koji nisu imali sve stupce popunjene. To je napravljeno kako bi se dobio model koji će raditi što točnije.

Dalje, podaci su normalizirani i podijeljeni na podatke koji će se koristiti prilikom treniranja modela i na podatke koji će se koristiti za evaluaciju treniranog modela. 70% podataka je korišteno za treniranje modela, dok je 30% korišteno za njegovu evaluaciju.

U svrhu testiranja osim iznad navedene klasifikacije korištene su tri vrste više klasne klasifikacije a to su:

* *Multiclass Neural Network* [5]
* *Multiclass Decision Jungle* [3]
* *Multiclass Logistic Regression* [4]

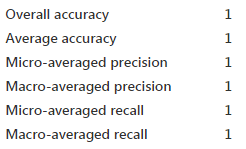
Konačni izgled modela moguće je vidjeti na slici 5.1.



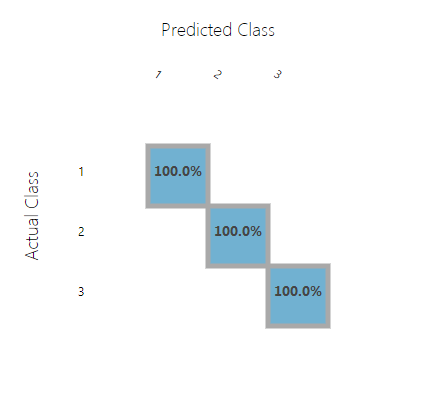
**Slika 5.1. –** Konačni izgled modela

## Testiranje i evaluacija dobivenih rezultata

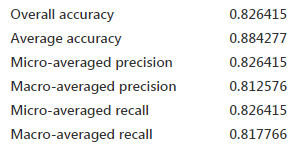
Nakon pokretanja azure servis je obavio testiranje i dobiveni su rezultati. Svi algoritmi klasifikacije koji su korišteni u ovom testu su dali vrlo visoke postotke preciznosti i točnosti, no najtočnija metoda je ipak bila *Multiclass Decision Forest*. Rezultati *Multiclass Decision Forest* prikazani su na slici 5.2. dok su rezultati ostalih klasifikacija prikazani na slikama 5.3.-5.9.



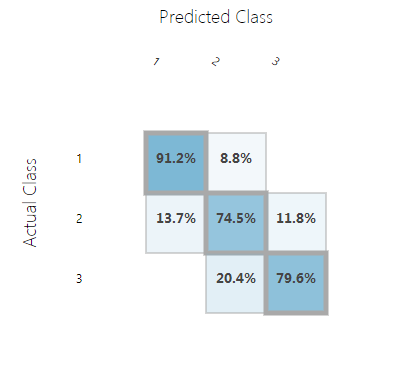
**Slika 5.2. –** Rezultat za *Multiclass Decision Forest*



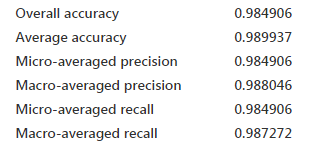
**Slika 5.3. –** Matrica konfuzije za *Multiclass Decision Forest-a*



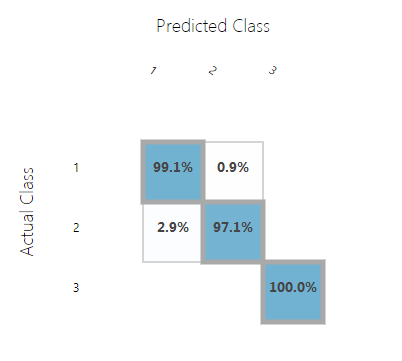
**Slika 5.4. –** Rezultat za *Multiclass Neural Network*



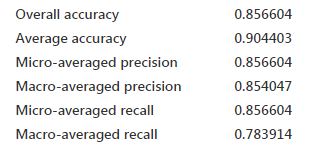
**Slika 5.5. –** Matrica konfuzije za *Multiclass Neural Network-a*



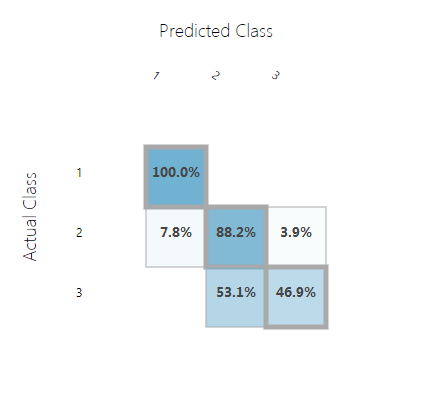
**Slika 5.6. –** Rezultat za *Multiclass Decision Jungle-a*



**Slika 5.7. –** Matrica konfuzije za *Multiclass Decision Jungle-a*



**Slika 5.8. –** Rezultati za *Multiclass Logistic Regression*

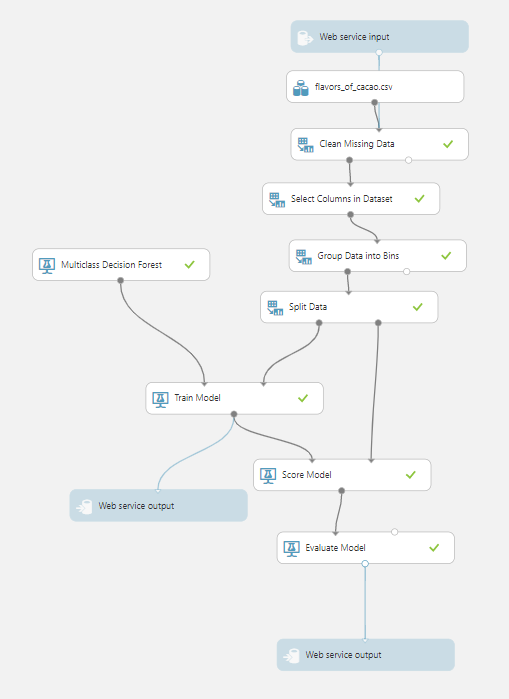


**Slika 5.9. –** Matrica konfuzije za *Multiclass Logistic Regression*

Na primjeru Na primjeru *Multiclass Decision Forest-a* je moguće vidjeti da su svi podatci točno klasificirani dok kod ostalih vrsta klasifikacije postoji pogreška koja iznosi skoro 20% kao što je slučaj kod *Multiclass Neural Network-a*. 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 završeno sa treniranjem modela sljedeći korak je izrada web servisa 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 5.10.)



**Slika 5.10. –** 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 koja se generira iz C# rječnika na koji se dobiva JSON odgovor koji je moguće vidjeti na slici (5.2). JSON je transportni format podatka koji omogućuje jednostavno slanje i primanje podataka putem API-a.

**Programski kôd 5.1. –** Izgled C# rječnika pomoću kojeg se generira JSON zahtjev

{

Inputs = {

System.Collections.Generic.Dictionary<string, System.Collections.Generic.List<System.Collections.Generic.Dictionary<string, string>>>},

GlobalParameters = {System.Collections.Generic.Dictionary<string, string>

}

}

**Programski kôd 5.2. –** Primjer JSON odgovora

{

"Results": {

"output1": [

{

"Company":"Guittard",

"Specific Bean Origin":"Complexite",

"Cocoa Percent":"0.7",

"Company Location":"Guittard",

"Rating":"35",

"Bean Type":"Bland",

"Broad Bean Origin":"U.S.A.",

"Rating\_quantized":"3",

"Scored Probabilities for Class "1"":"0.0419902912621359",

"Scored Probabilities for Class "2"":"0.193446601941748",

"Scored Probabilities for Class "3"":"0.764563106796117",

"Scored Labels":"3"

}

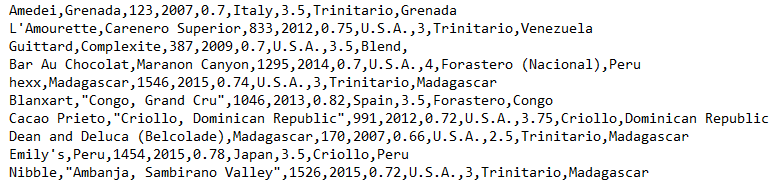
]

}

}

## Testiranje izmijenjenog skupa podataka

U svrhu testiranja aplikacije iz skupa podataka izdvojeno je 10 podataka koji će služiti za potrebe testiranja. Podatci su uneseni u aplikaciju te se je vršila evaluacija aplikacije. Podatci su odabrani nasumično te se mogu vidjeti na slici 5.3.



**Slika 5.3. –** Instance nasumično odabrane za testiranje

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:  Loša | Predviđeni:  Dobra | Predviđeni: Izvrsna |
| Stvarni:  Loša | 4 | 0 | 0 |
| Stvarni:  Dobra | 0 | 5 | 0 |
| Stvarni: Izvrsna | 0 | 0 | 1 |

**Tablica 5.1. –** Matrica konfuzije

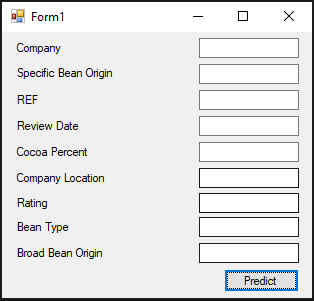
Iz matrice konfuzije je vidljivo da je od 10 čokolada samo 1 je klasificirana kao izvrsna, 5 kao dobre i samo 1 kao loša.

# OPIS PROGRAMSKOG RJEŠENJA

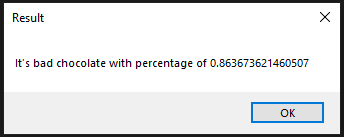
Nakon izrade modela strojnog učenja i objave web servisa potrebno je izraditi aplikaciju za testiranje. Aplikacija treba omogućiti unos parametara koji će biti testirani, te ispis rezultata u obliku određene klase i vjerojatnosti.

## Forma za unos parametara i prikaz rezultata

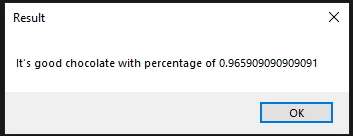
Za unos parametara napravljena je forma (slika 6.1.) u kojoj se unose potrebne vrijednosti, te se pritiskom da tipku „Predict“ podaci šalju web servisu na obradu. Dobiveni rezultat ispisuje o kojoj klasi čokolade se radi, te vjerojatnost odgovora. Ispis loše, dobre i izvrsne kvalitete čokolade vidljiv je na slikama 6.2 - 6.4.



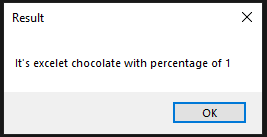
**Slika 6.1. –** Izgled forme za unos podataka



**Slika 6.2. –** Izgled dobivenog rezultata kada je u pitanju loša čokolada



**Slika 6.3. –** Izgled dobivenog rezultata kada je dobra čokolada u pitanju



**Slika 6.4. –** Izgled dobivenog rezultata kada je u pitanju loša čokolada

# 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 Kaggle, napravljen je i istreniran model koji na temelju ulaznih parametara vrši klasifikaciju čokolade.

Na temelju istreniranog modela izrađeni su web servis i aplikacija koja ga koristi. Aplikacija u vrlo brzom roku može odrediti rezultata za testirane parametre i vjerojatnost tog rezultata.

Aplikacija bi mogla imati nekoliko primjena, a najočitija je da gurmanima čokolade pomogne u pri klasifikaciji čokolade.

# LITERATURA

1. [https://www.kaggle.com/rtatman/chocolate-bar-ratings](https://www.kaggle.com/rtatman/chocolate-bar-ratings%20)
2. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-forest>
3. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-jungle>
4. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-logistic-regression>
5. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-neural-network>

# POVEZNICE

1. <https://github.com/szabbo/ruapProject>