**SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU**

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

**Diplomski studij**

**KLASIFIKACIJA VINA**

**SEMINARSKI RAD**

**RAČUNARSTVO USLUGA I ANALIZA PODATAKA**

**Dominik Sabo  
Ivan Sertić**

Osijek, 2020.

**SADRŽAJ**

[**1.** **UVOD** 1](#_Toc32964576)

[**2.** **OPIS** 2](#_Toc32964577)

[**3.** **KORIŠTENI PODACI** 3](#_Toc32964578)

[**4.** **KORIŠTENI POSTUPCI STROJNOG UČENJA** 3](#_Toc32964579)

[**4.1. Multiclass Decision Forest** 4](#_Toc32964580)

[**4.2. Multiclass Decision Jungle** 4](#_Toc32964581)

[**4.3. Multiclass Neural Network** 4](#_Toc32964582)

[**4.4. Multiclass Logistic Regression** 5](#_Toc32964583)

[**5.** **MODEL STROJNOG UČENJA** 5](#_Toc32964584)

[**5.1. Planiranje i izrada modela** 5](#_Toc32964585)

[**5.2. Testiranje i evaluacija dobivenih rezultata** 6](#_Toc32964586)

[**5.3. Izrada web servisa i korištenje API-a** 11](#_Toc32964587)

[**6.** **OPIS PROGRAMSKOG RJEŠENJA** 13](#_Toc32964588)

[**6.1. Forma za unos parametara i prikaz rezultata** 13](#_Toc32964589)

[**7.** **ZAKLJUČAK** 15](#_Toc32964590)

[**LITERATURA** 16](#_Toc32964591)

[**POVEZNICE** 16](#_Toc32964592)

# **UVOD**

U ovom projektu obrađuje se predviđanje kvalitete vina modelom strojnog učenja. Aplikacija omogućuje unos parametara prema kojima se vino klasificira u lošu, dobru ili vrlo dobru kvalitetu.

Pošto se vino svrstava u klase ovisno o kvaliteti, u ovom projektu korištena je metoda klasifikacije kojom se vino ovisno o kvaliteti svrstava u određenu klasu.

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

# **OPIS**

Vino je jedno od najpopularnijih vrsta alkoholnih pića koje konzumira gotovo cijela populacija svijeta, te postoji vino za svaki ukus. Postoji mnogo proizvođača vina i nitko ne koristi isti recept za proizvodnju, a samim time što postoji puno različitih sorti grožđa koje se može koristiti za proizvodnju vina znači da postoji mnogo faktora koji mogu utjecati na okus i kvalitetu samog vina. Najvažniji parametri prema kojima se određuje kvaliteta vina su vrsta vina (bijelo, crno), količina alkohola i količina šećera u vinu. Time, ovaj projekt se bazira na klasifikaciji vina na osnovu tih parametara.

U ovom projektu bit će korišteni parametri koje korisnik sam unosi u aplikaciju te se prema njima rezultat klasificira u jedan od tri razreda kvalitete po ocjeni koju dobije prema parametrima. Ocjene za vino u ovom projektu su 5, 6 (loše vino), 7, 8 (dobro vino) i 9 (vrlo dobro vino). ocjene 1, 2, 3 ,4 i 10 se u ovome projektu ne koriste jer su u bazi podataka vina bili nepotpuni podaci.

Parametri korišteni za klasifikaciju su:

* Vrsta vina
* Postotak alkohola
* Količina šećera

# **KORIŠTENI PODACI**

Kako bi aplikacija uspješno odrađivala svoj zadatak, potrebno je napraviti model strojnog učenja koji će na temelju predanih podataka moći klasificirati rezultat u jedan od tri razreda u koje ga model može klasificirati (model loših, dobrih i vrlo dobrih vina).

Jako važan dio izrade modela strojnog učenja je pronalazak i obrada podataka kojima će se model trenirati. Svi parametri koji nisu bili potrebni ili su mogli negativno utjecati na ishod rezultata su izbačeni kako bi se dobio model koji se može koristiti u što realnijim uvjetima.

Skup podataka korišten u ovom projektu preuzet je sa web stranice Kaggle i može se pronaći na poveznici [1]. U skupu postoji 6498 instanci vina od kojih su izbačene sve koje su imale nepotpune podatke kako bi trenirani model radio što bolje. U podacima osim navedenih parametara korištenih za klasifikaciju postoji velik broj parametara koji se nisu koristili, to su: kiselost, kloridi, sumporov dioksid, gustoća, pH vrijednost i sulfati.

Svi ti podaci su izbačeni iz ovog modela jer imaju mali utjecaj na kvalitetu vina i gotovo su potpuno nedostupni korisnicima. Gotovo nikad ne pišu na etiketi vina te su dostupni jedino u vinariji u kojoj je vino proizvedeno.

# **KORIŠTENI POSTUPCI STROJNOG UČENJA**

Nakon pronalaska odgovarajućeg skupa podataka te njegove obrade, sljedeći korak je odabir postupka strojnog učenja koji će biti najpogodniji za model, odnosno koji će davati najveću tolnost i najmanju pogrešku tijekom evaluacije. Pošto je ovdje potrebno odrediti je li vino loše, dobro ili vrlo dobro, korištena je metoda klasifikacije *Multiclass Neural Network* jer su njome postignuti najbolji rezultati.

Klasifikacija je postupak kojim se elementi skupa podataka (vina) dodjeljuju raredu kvalitete u ovisnosti koje kriterije ti elementi ispunjavaju po modelu bitnim parametrima. klasifikacija je metoda strojnog učenja kojom se određuje kategorija, klasa ili tip jednog ili više podataka.

Klasifikacija se najčešće dijeli po tome vrši li se klasifikacija na dvije ili više klasa. Ako se vrši na dvije klase, radi se o binarnoj klasifikaciji. Jednostavan primjer binarne klasifikacije je, recimo, model koji prepoznaje je li osoba muško ili žensko, jer su to dvije klase koje model raspoznaje. Ako ima više od dvije klase, radi se o višeklasnoj klasifikaciji.

## **4.1. Multiclass Decision Forest**

*Multiclass Decision Forest* je skupina metoda učenja namijenjena za klasifikaciju na više klasa, gdje se skupine metoda ne temelje na jednom modelu, već na više njih kako bi dobili bolje rezultate i općenitiji model. Ova metoda funkcionira tako da se izgradi više stabala odlučivanja, te se onda pomoću glasanja određuje izlazna klasa. [2]

## **4.2. Multiclass Decision Jungle**

*Multiclass Decision Jungle* je algoritam koji vraća netrenirani klasifikator, te se model osposobljava korištenjem modela treniranja ili *Tune Model Hyperparameters.* Trenirani model se zatim koristi za klasifikaciju podataka. Džungle odlučivanja su proširene šume odlučivanja, te se sastoje od skupa acikličkih grafova usmjerenih na odluku. [3]

## **4.3. Multiclass Neural Network**

Neuronska mreža je skup međusobno povezanih slojeva, gdje ulazi predstavljaju prvi sloj koji je povezan s izlaznim slojem acikličim grafom koji se sastoji od težinskih rubova i čvorova. Između ulaznog i izlaznog stoja može biti više skrivenih slojeva. Odnos između ulaza i izlaza saznaje se iz treninga neuronske mreže na ulaznim podacima. Smjer grafa odvija se od ulaza preko skrivenog sloja i do izlaznog sloja. Svi čvorovi u sloju su težinski rubovi vezani s čvorovima u sljedećem sloju. Da bi se izračunao izlaz mreže za određeni ulaz, izračunava se vrijednost na svakom čvoru u skrivenim slojevima i u izlaznom sloju. Vrijednost se postavlja izračunavanjem težinske sume vrijednosti čvorova iz prethodnog sloja. Zatim se na taj težinski zbroj primjenjuje funkcija aktivacije. [4]

## **4.4. Multiclass Logistic Regression**

*Multiclass Logistic Regression* je klasifikacijska metoda koja generalizira logističku regresiju na probleme više klase, odnosno na probleme sa više od dva moguća diskretna ishoda. To je model koji se koristi za predviđanje vjerojatnosti različitih mogućih ishoda kategorički raspodijeljene zavisne varijable, s obzirom na skup neovisnih varijabli. [5]

# **MODEL STROJNOG UČENJA**

## **5.1. 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. Za metodu klasifikacije odabrana je *Multiclass Neural Network* metoda

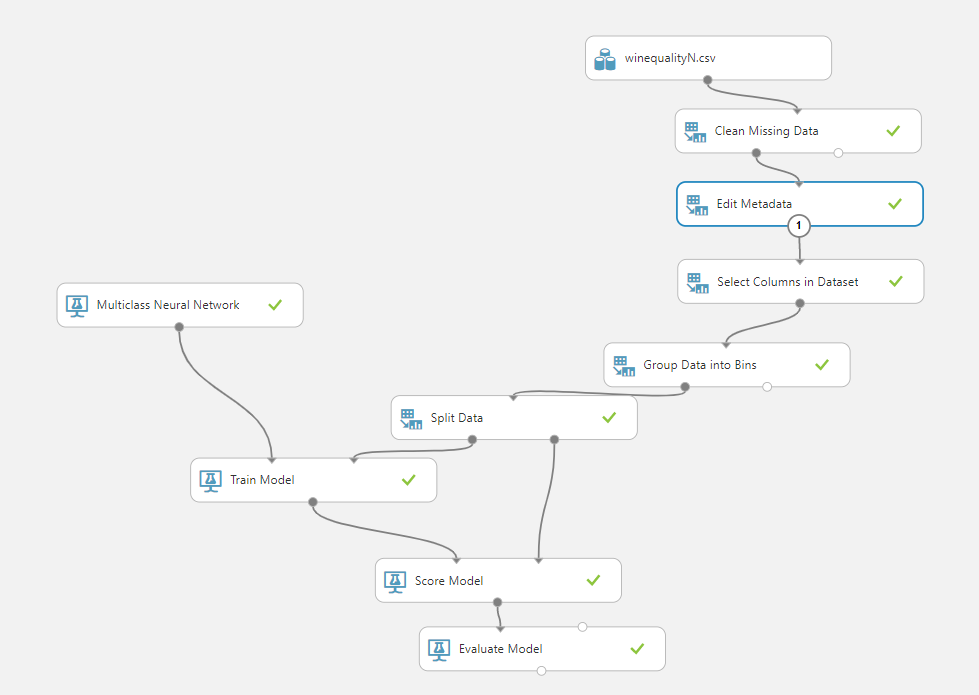
Prvo se u model dodaje skup podataka kojim se trenira model. U ovom slučaju to je datoteka „winequalityN.csv“ formata CSV (engl. *Comma-Seperated Values*).

Nakon dodavanja skupa podataka potrebno je urediti podatke. Uklonjeni su svi redci koji nisu imali sve popunjene stupce kako bi model što točnije vršio klasifikaciju.

Tada se podaci normaliziraju i dijele na podatke koji će se koristiti za treniranje modela i podatke za evaluaciju treniranog modela. 70% podataka je korišteno za treniranje modela, a 30% za evaluaciju.

Prilikom izrade modela korištene su i druge metode klasifikacije kako bi se utvrdilo koja metoda najbolje paše za ovaj model. Tako se došlo do korištenja *Multiclass Neural Network* metode.

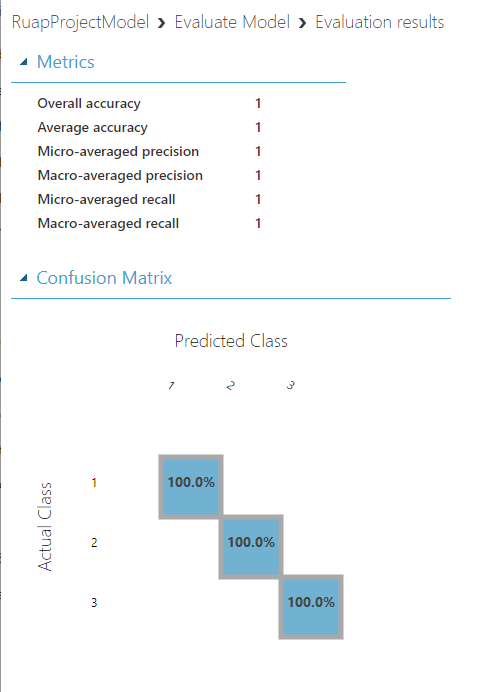
Konačni izgled modela vidljiv je na slici 5.1.



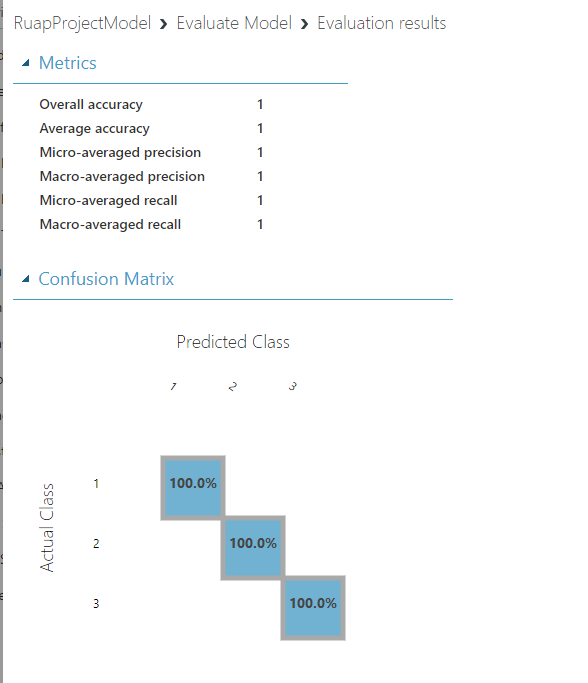
**Slika 5.1.** – Konačni izgled modela

## **5.2. Testiranje i evaluacija dobivenih rezultata**

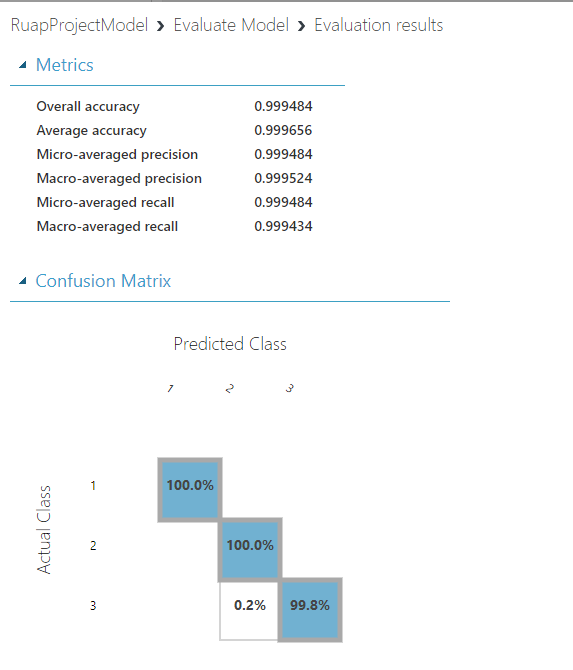
Nakon izrade modela i pokretanjem, Azure servis je obavio testiranje i dobiveni su rezultati. Algoritmi *Multiclass Decision Jungle* i *Multiclass Decision Forest*, iako najtočniji, u njihovim rezultatima nije se vidio utjecaj promjene dva od tri bitna parametra za klasifikaciju u ovom projektu. Zbog toga je odabran *Multiclass Neural Network* algoritam koji je bio točniji od *Multiclass Logistic Regression* algoritma. Retultati klasifikacije svih algoritama vidljivi su na slikama 5.2. – 5.5.



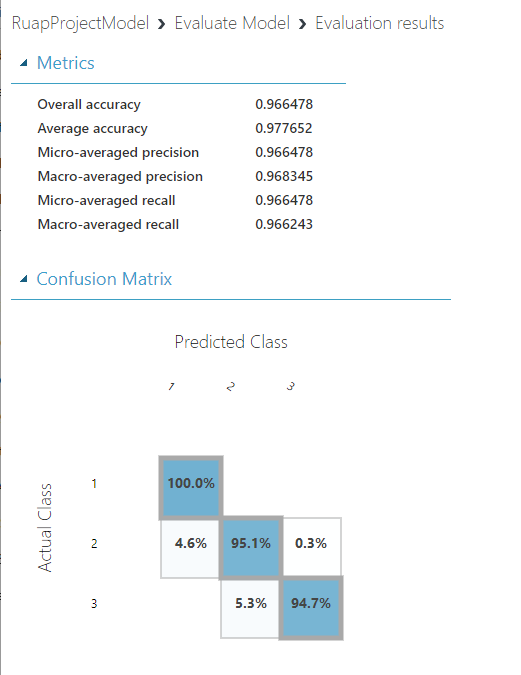
**Slika 5.2.** – Rezultat za *Multiclass Decision Forest*



**Slika 5.3.** – Rezultat za *Multiclass Decision Jungle*



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

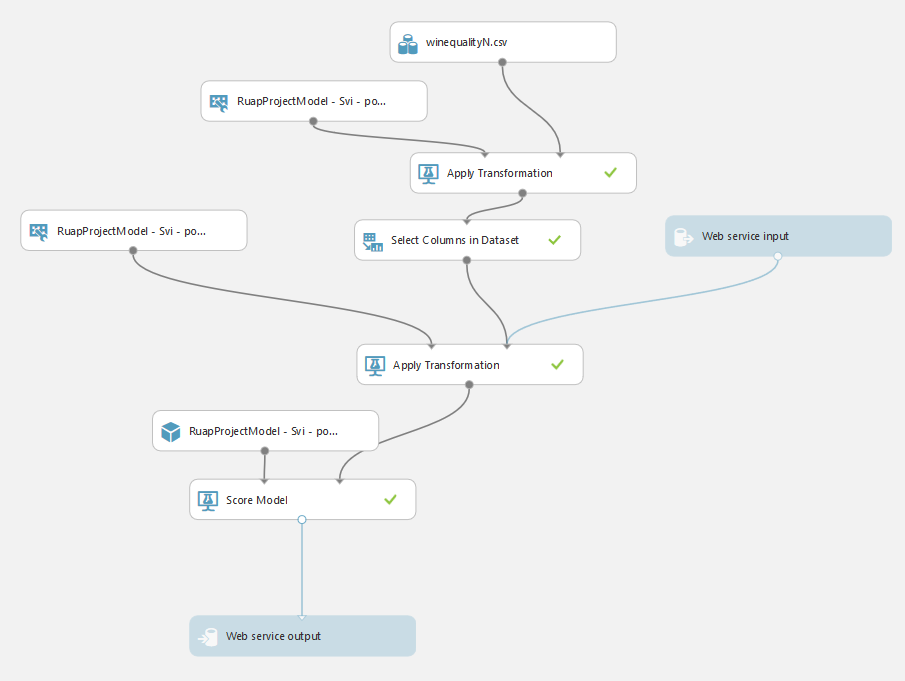


**Slika 5.5.** – Rezultat za *Multiclass Logistic Regression*

Iz dobivenih matrica konfuzije vidi se da su svi algoritmi točni, *Multiclass Logistic Regression* ističe se kao najmanje točan algoritam. Zbog prije spomenutih razloga, iako su Jungle i Forest metode najtočnije, koristimo *Multiclass Neural Network* algoritam.

## **5.3. Izrada web servisa i korištenje API-a**

Nakon što je završeno s treniranjem modela, izrađujemo web servis koji će preko API-a primati parametre iz aplikacije, klasificirati rezultat i vratiti ga korisniku. Za izradu web servisa potrebno je izabrati opciju „*Train web service“* kojom se stvara prediktivni eksperiment (slika 5.6.).



**Slika 5.6.** – Izgled prediktivnog eksperimenta

Nakon što je izrađen prediktivni eksperiment, potrebno ga je pokrenuti i odabrati opciju „*Deploy web service*“ čime dobivamo *Request URI* i API ključ. Web servisu pristupamo pomoću toga, te je uz to potrebna odgovarajuća JSON forma za zahtjev koja se generira iz C# riječnika na koji se dobiva JSON odgovor. JSON je transportni format podatka koji omogućuje jednostavno slanje i primanje podataka putem API-a.

var scoreRequest = new

{

Inputs = new Dictionary<string, List<Dictionary<string, string>>> () {

{

"input1",

new List<Dictionary<string, string>>(){new Dictionary<string, string>(){

{

"type", ""

},

{

"residual sugar", "1"

},

{

"alcohol", "1"

},

}

}

},

},

GlobalParameters = new Dictionary<string, string>() {

}

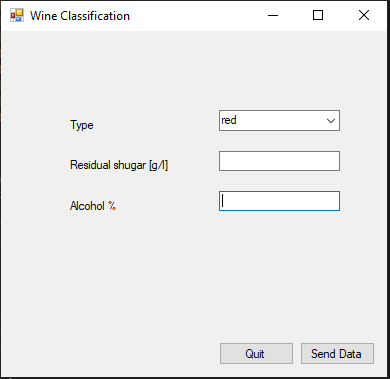
**Programski kod 5.1.** – Izgled C# riječnika kojime se generira JSON zahtjev

# **OPIS PROGRAMSKOG RJEŠENJA**

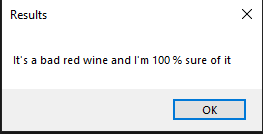
Nakon objave web servisa potrebno je izraditi aplikaciju za testiranje. Aplikacija omogućuje unos parametara koji će biti testirani, te ispis rezultata u obliku klase koju je model odredio i postotak točnosti. Aplikacija je izrađena u C# kao *Windows Form* u *Visual Studio* programskom okruženju.

## **6.1. Forma za unos parametara i prikaz rezultata**

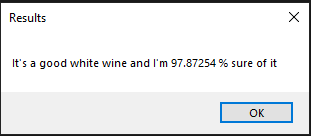
Vrijednosti parametra unose se u formu (slika 6.1.), te se pritiskom na tipku „Send Data“ podaci šalju web servisu na obradu. Dobiveni rezultat ispiše dobivenu vrijednost za razred vina. Ispisi za loše, dobro i vrlo dobro vino vide se na slikama 6.2. – 6.4.



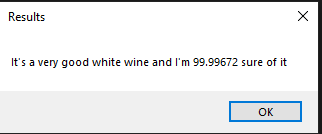
**Slika 6.1.** – Forma za unos podataka



**Slika 6.2.** – Rezultat za loše vino



**Slika 6.3.** – Rezultat za dobro vino



**Slika 6.4.** – Rezultat za vrlo dobro vino

# **ZAKLJUČAK**

Strojno učenje postaje sve popularniji postupak obrade informacija u modernom vremenu. Postoje mnoge mogućnosti i primjene za strojno učenje, a jedna od njih je opisana u ovom radu. Korištenjem skupa podataka s repozitorija Kaggle, izrađen 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. Servis na temelju podataka unesenih u aplikaciju određuje rezultat i vjerojatnost tog rezultata velikom točnošću.

# **LITERATURA**

[1] [https://www.kaggle.com/rajyellow46/wine-quality](https://www.kaggle.com/rajyellow46/wine-quality?fbclid=IwAR1BYPvzK-gR_LIAtJnK-zJLzkxrPNZg2nDKClglaXTyk-VDioc7RtMtRaw)

[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-neural-network>

[5] <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-logistic-regression>

# **POVEZNICE**

[1] <https://github.com/ivansertic/wineClassification>