SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU

FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA

Računarstvo usluga i analiza podataka

**Chocolate Prediction**

SEMINARSKI RAD

**Barbara Matijević**

Osijek, 14.02.2020

Sadržaj:

[1. Uvod 3](#_Toc32788260)

[2. Opis problema 4](#_Toc32788261)

[2.1. Korišteni podatci 4](#_Toc32788262)

[2.2. Korišteni postupci strojnog učenja 5](#_Toc32788263)

[3. Opis programskog rješenja 8](#_Toc32788264)

[3.1. Model strojnog učenja 8](#_Toc32788265)

[3.2. Način korištenja API-a 10](#_Toc32788266)

[3.3. Klijentska aplikacija 11](#_Toc32788267)

[4. Zaključak 12](#_Toc32788268)

[5. Poveznice i Literatura 13](#_Toc32788269)

# Uvod

Poznate li nekoga tko ne voli čokoladu?

Čokolada je poslastica koja se dobiva miješanjem **kakaove mase** s većom ili manjom količinom šećera. Najčešće na tržište dolazi u obliku **pločica**. Čokolada je postala univerzalni simbol hedonizma. Slično kao kava, čokolada je postala globalni proizvod zahvaljujući industrijskoj proizvodnji koja je snizila cijene i učinila je masovno dostupnom, navikavajući generacije kupaca na ono što danas smatramo standardom čokolade. Upravo ovo je glavni razlog, zašto je važno razlikovati dobru čokoladu od one koja je loša.

Odrediti kvalitetu čokoladu je zamršen postupak, jer se ocjena čokolade temelji na kombinaciji objektivne kvalitete i subjektivne interpretacije.

# Opis problema

Ovaj zadatak je problem klasifikacije. Klasifikacija je model nadziranog učenja (engl. supervised classification) podataka i ujedno predstavlja najčešći i jedan od temeljnih problem u rudarenju podataka. [1.] Odabrani parametri na temelju kojih će se predviđati kojoj kategoriji pripada koji uzorak su**: naziv tvrtke, postotak kakaa i tip zrna**.

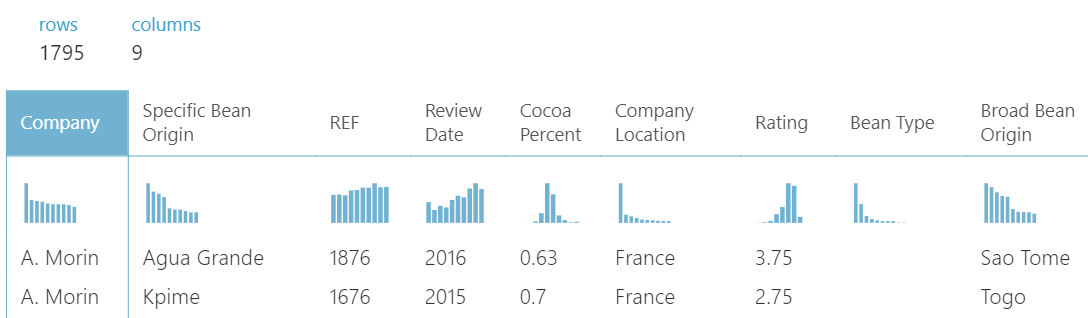
Cilj problema je smjestiti ulazni uzorak u jednu od mogućih klasa : BAD, GOOD i EXCELENTE CHOCOLATE. Svaka klasa ima točno jednu vlastitu oznaku.

## Korišteni podatci

Skup korištenih podataka pronađen je na stranici Kaggle[2.] , koristi se u obliku .csv datoteke. Skup podataka sastoji se od 9 stupaca i 1795 redaka. Svaki redak predstavlja jedan element(jednu čokoladu) skupa podataka.

Svaki stupac predstavlja jedan parametar koji opisuje čokoladu:

* **Company** - Naziv tvrtke koja proizvodi čokoladicu
* **Specific Bean Origin** - Specifična geo-regija s koje potječe pločica
* **REF –** vrijednost koja predstavlja broj unesenih recenzija
* **ReviewDate** – godina objave recenzije
* **CocoaPercent** – postotak kakaa
* **CompanyLocation –** zemlja u kojoj je proizvedena čokolada
* **Rating –** stručna ocjena čokoladice
* **BeanType –** vrsta zrna koje se koristi
* **Broad BeanOrigin –** geo-regija iz koje potječe zrno kakaovca



Slika 2.1. Vizualizirani skup podataka

## Korišteni postupci strojnog učenja

|  |  |
| --- | --- |
| Postupak Strojnog učenja | Opis metode |
| Multiclass Decision Tree | Stablo odluke je jednsotavan algoritam koji se provodi na sljedeći način:  1. Postavlja se najbolji atribut skupa podataka kao korijen stabla  2. Skup podataka se dijelu u podskupove tako da svi elementi podskupa imaju jednaku  vrijednost odabranog atributa  3. Ponavljanje koraka 1. i 2. na svakom podskupu dok se ne dođe do kraja na svakoj od grana  drveta |
| Multiclass Decision Forest | Šume odluke su skup stabala odluke, a konačni rezultat dobije se kao odabir većinskog  odgovora. |
| Multiclass Decision Jungle | Džungle odluke su poput šuma odluke, samo što umjesto stabala, džungle koriste usmjerene acikličke grafove, gdje čvor može imati više od jednog roditeljskog čvora. |
| Multiclass Logistic Regression | Logistička regresija dobiva izlaz pomoću sigmoidne funkcije za vraćanje vrijednosti vjerojatnosti koja se može mapirati u više diskretnih klasa. Sigmoidn funckija može mapirati bilo koju realnu vrijednost u drugu između 0 i 1 [3.] |

Tablica 2.1. Opis korištenih podataka

|  |  |
| --- | --- |
| **USPOREĐIVANJE REZULTAT POJEDINIH MODELA** | |
| **Multiclass Decision Forest** |  |
| **Multiclass Decision Jungle** |  |
| **Multiclass Logistic Regression** |  |

Tablica 2.2. Uspoređivanje matrica konfuzije

Na Azure web platformi napravljen je novi eksperiment u kojem je izrađen model strojnog učenja, koji će kasnije služiti za klasifikaciju kvalitete čokolade. Prilikom izrade modela potrebno je odlučiti kojom metodom klasifikacije ćemo se koristiti. Kako bi znali koja metoda daje najbolje rezultate, potrebno ih je sve implementirati u isti model, te usporediti dobivene rezultate.

Nakon pokretanja izrađenog modela dobijemo rezultat simulacije. Rezultat simulacije moguće je vizualizirati - rezultat se dobiva u obliku matrice konfuzije (tablica 2.2.).

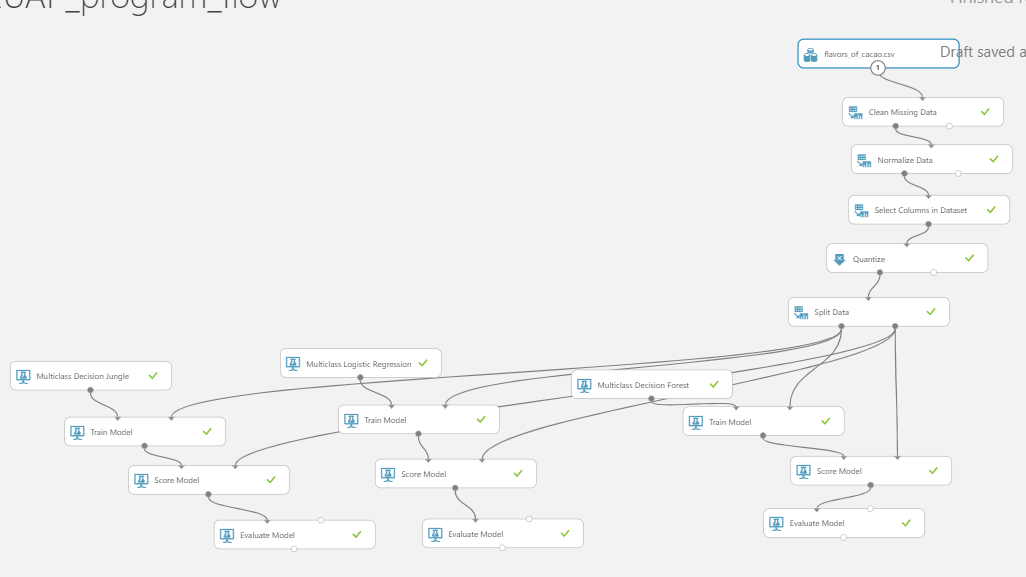
Analiziranjem rezultata, koji su prikazani u tablici 2.2., zaključujemo kako svi klasifikatori imaju visoku točnost klasifikacije, ali Multiclass Decision Forest metoda daje najbolje rezultate te će ona biti korištena u daljnjoj implementaciji modela.

# Opis programskog rješenja

## Model strojnog učenja

Na Microsoft Azure Machine Learning studia kreiran je novi **workspace**. Unutar workspace-a kreiran je novi eksperiment, a unutar eksperimenta pravimo model. Prilikom izrade modela moramo izabrati klasifikator koji daje najbolje rezultate. Prilikom izrade modela korištena su tri, već ranije opisana, klasifikatora – a to su:

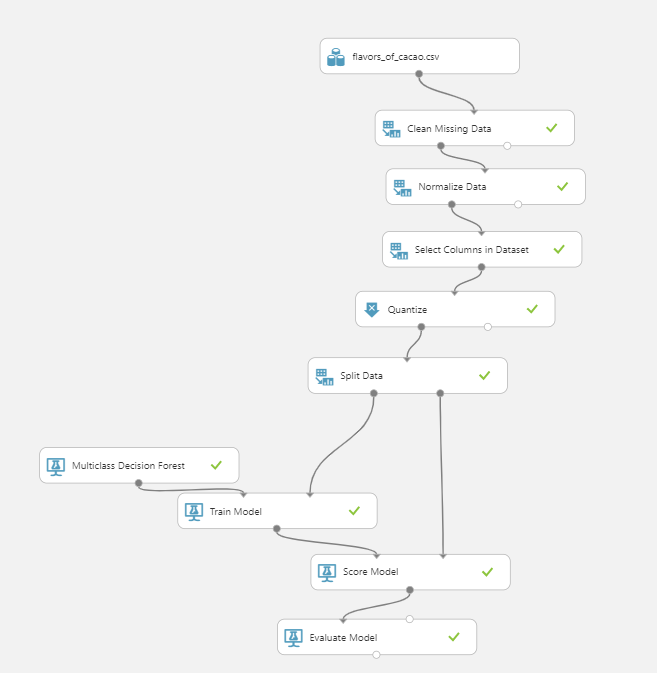
* Multiclass Logistic Regression
* Multiclass Decision Jungle
* Multiclass Decision Forest



Slika 3.1. Model prilikom donošenja odluke

Slika 3.1. prikazuje kako je izgledao model prilikom donošenja odluke o optimalnom klasifikatoru. Nakon što je model pokrenut i nakon što su uspoređeni rezultati, donosena je odluka da „**Multiclass Decision Forest**“ daje najbolje rezultate s prosječnom točnošću od **99.62%,** ostala dva klasifikatora su uklonjena sa modela.

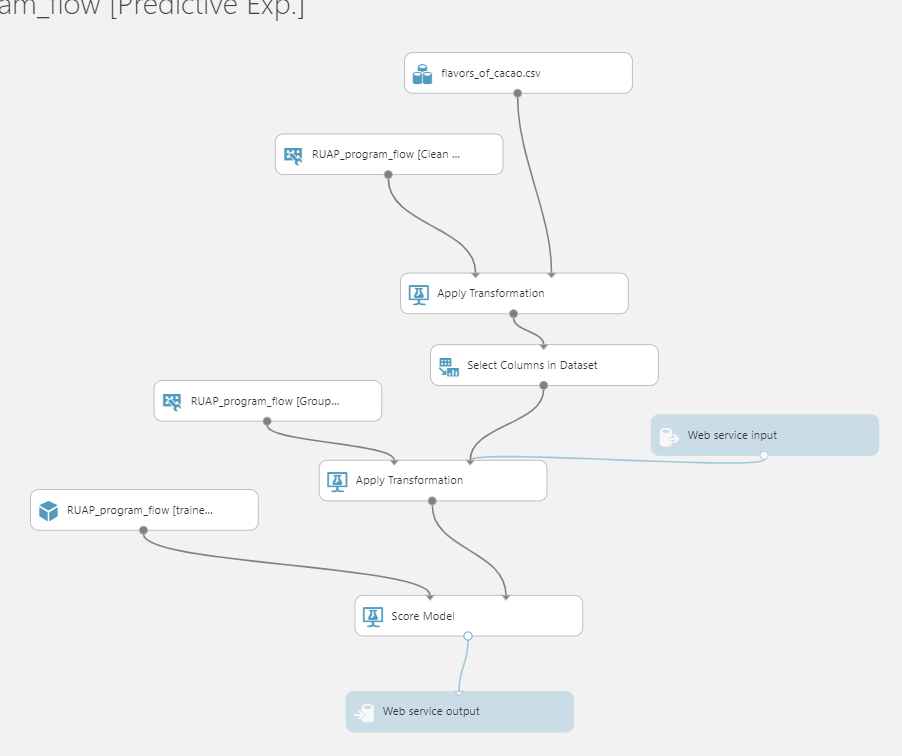
Konačni model, prije Deploy web servisa sada izgleda ovako:



Slika 3.2. Konačni model prije Deploy Web Service

## Način korištenja API-a

Nakon odabira najboljeg modela, model je postavljen kao web usluga na Microsoft Azureu pomoću opcije „Deploy Web Service[Classic]“. Naš model sada izgleda ovako:



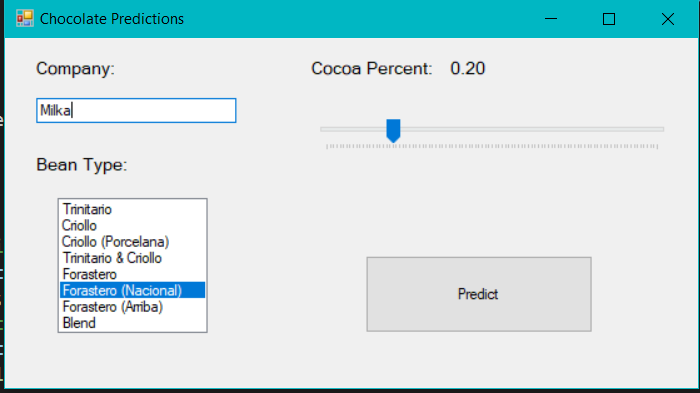
Slika 3.3. Web service

Web servis generira API ključ i Request/Response link – upravo ovi elementi omogučavaju korištenje modela preko različitih aplikacija.

## Klijentska aplikacija

Klijentska aplikacija izrađena je pomoću c# programskog jezika, unutar Visual Studio platforme. Aplikacija omogućava korisniku slanje http Request-a, te obradu podataka na Azure sa povratnim odgovorom o klasi unutar koje pripada novi uzorak.

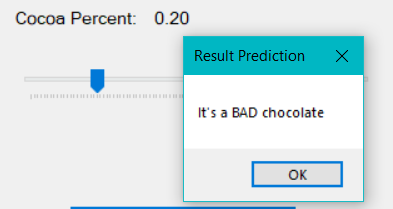
|  |  |  |
| --- | --- | --- |
| Aplikacija se sastoji od | | |
| *Komponenta* | *Naziv* | Opis |
| *Text Box* | *Company* | Unutar ovog polja korisnik upisuje naziv tvtke koja proizvodi čokoladu |
| *List Box* | *Bean Type* | Lista u kojoj su navedeni svi mogući tipovi zrna kakaovca |
| *Track bar* | *Cocoa Percent* | Korisnik bira postotak kakaa – vrijednost je u intervalu [0,1] |
| *Gumb* | *Predict* | Tipkom na gumb predict, uneseni podatci se parsiraju te se šalju na azure gdje se obrađuju. Kao response Dobijemo MessageBox koji nam daje poruku o novodoseljenoj klasi |



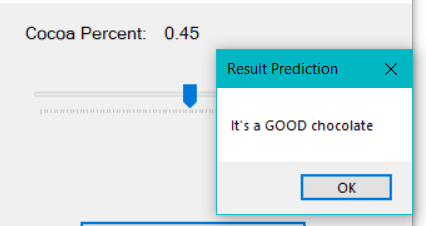
Slika 3.4. Prikaz Klijenteske aplikacije

Mogući odgovori (Klase):

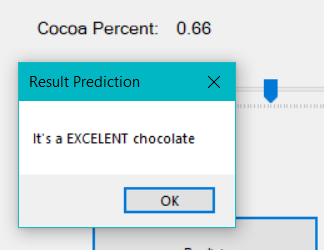
1. BAD CLASS



1. GOOD CLASS



1. EXCELENTE CLASS



# Zaključak

Klasifikacija je postupak dohvaćanja uzorka i smještanja tog uzorka u odgovarajuću kategoriju, odnosno klasu. Svaka kategorija ima točno jednu vlastitu oznaku. Klasifikacija podataka predstavlja nadzirano klasificiranje/učenje (engl. supervised classification).

Izrada klasifikacijskog model odrađena je u Azure Machine Learning Studio-u, pomoću najtočnijeg klasifikatora, u ovom slučaju - Multiclass Decision Forest. (Važno za napomenuti je da točnost algoritma uvelike ovisi o korištenim podatcima tj. o njihovoj podijeli na trening podatke i testne podatke )

Azure nam, također nudi i opciju postavljanja modela kao web usluge, što nam kasnije omogućava pristupanje rezultatima predikcije preko različitih aplikacija

# Poveznice i Literatura

[1.] RUAP, Predolaž za laboratorijske vježbe, LV5, Analiza podataka

[2.] <https://www.kaggle.com/rtatman/chocolate-bar-ratings>

[3.] <https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html>

[4.] <https://github.com/bmatijevic65/RUAP_Chocolate_predictions.git>

<https://europewest.studio.azureml.net/Home/ViewWorkspaceCached/e4a7547db8f74b0e87e7382f6d2ace05#Workspace>