Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek

Računarstvo usluga i analiza podataka

SEMINARSKI RAD

„Predviđanje parametara kakaa“

Dominik Birtić

Marko Huljak

Osijek, 2018.

Sadržaj

[1. Uvod 1](#_Toc522822927)

[2. Opis problema 2](#_Toc522822928)

[2.1. Korišteni podaci 2](#_Toc522822929)

[2.2. Korišteni postupci strojnog učenja 2](#_Toc522822930)

[3. Opis programskog rješenja 4](#_Toc522822931)

[3.1. Model strojnog učenja 4](#_Toc522822932)

[3.2. Način korištenja API-ja 6](#_Toc522822933)

[3.3. Klijentska aplikacija 7](#_Toc522822934)

[4. Zaključak 8](#_Toc522822935)

[5. Poveznice i literatura 9](#_Toc522822936)

# Uvod

Zadatak ovog projekta bio je napraviti program koji će predvidjeti postotak kakaa u proizvodu na osnovi nekoliko parametara. Potrebno je preuzeti skup podataka, jedan dio iskoristiti za stvaranje regresijskog modela, a drugi dio za treniranje i vrednovanje.

Nakon toga treba napraviti API koji će korisnicima ponuditi funkcionalnost našeg modela i izraditi klijentsku desktop aplikaciju koja će konzumirati API.

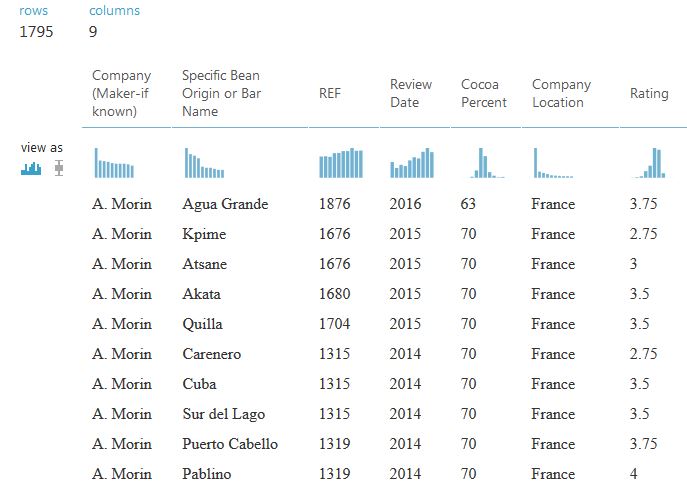
Cilj ovog projekta je omogućiti korisnicima uslugu predviđanja kakaa s određenim postotkom za zadane parametre.

# Opis problema

Podatak koji trebamo pronaći je mogućnost postojanja kakaa s danim postotkom. Podatke koje unosimo u aplikaciju su država u kojoj se nalazi firma u koja proizvodi kakao, ocjena kakaa, godina ocjene kakaa i postotak kakaa. Budući da se string vrijednosti teže predviđaju, onda smo za naš projekt odabrali brojčanu vrijednost (postotak kakaa), te smo pretpostavili da ćemo dobiti preciznije rezultate.

## Korišteni podaci

Korišteni podaci su preuzeti s Kaggle repozitorija [1]. Podaci su bili u .csv formatu i nije bilo potrebno izmijenjivati podatke prije uploadanja na Azure ML Studio. Budući da dva stupca dataseta sadrže nekoliko desetaka ili nekoliko stotina izgubljenih vrijednosti, onda ćemo koristiti ovih 7 stupaca kako bi istrenirali model.



**Slika 2. Primjer dataseta**

Na slici 2. dan je primjer korištenog dataseta. Parametri su firma u kojoj se proizvodi kakao, mjesto gdje se uzgaja ta sorta kakaa, referenti broj (REF), godina ocjene, postotak kakaa, država u kojoj se firma koja proizvodi nalazi i ocjena kakaa. Sveukupno je 1795 redova podataka od kojih je 75% korišteno za treniranje modela, a ostalih 25% za evaluaciju.

## Korišteni postupci strojnog učenja

Strojno učenje je vrsta obrade podataka gdje se računalo može utrenirati da samo donosi neke rezultate ili odluke bez da je to eksplicitno isprogramirano. Postoji nadzirano, djelomično nadzirano, nenadzirano i učenje s povratnom vezom. Postoje razredi učenja, a to su klasifikacija, regresija, rangiranje, grupiranje i smanjenje dimenzionalnosti.

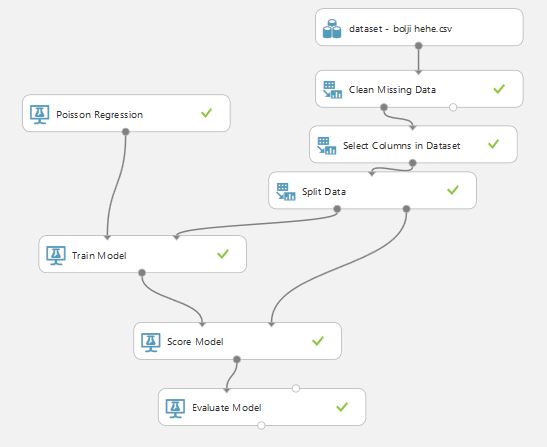
U projektu je korišeno nadzirano učenje. To znači da su računalu dani ulazi i izlazi, te ono na osnovu toga uči. Kao razred učenja korištena je regresija. Cilj je izrada modela za opis ovisnosti izlaza o ulaznim podacima. Uglavnom se koristi za potrebe predviđanja. Modeli mogu biti linearni ili nelinearni. Iako su linearni modeli jednostavniji oni mogu u slučaju manjeg broja podataka i podataka koji su raštrkani davati bolje rezultate u odnosu na nelinearne.

# Opis programskog rješenja

Zadatak je bio napraviti program koji će na osnovu unesenih parametara kakaa kao izlaz vratiti postotak kakaa i predviđenu vrijednost kakaa. Program je napravljen u C#, dok je za strojno učenje korišten Azure Machine Learning Studio. Korisnik unosi parametre u program. Program se spaja na Azure, te šalje te podatke. Nakon što dobije odgovor u obliku stringa, korisniku ispisuje postotak kakaa i predviđenu vrijednost kakaa.

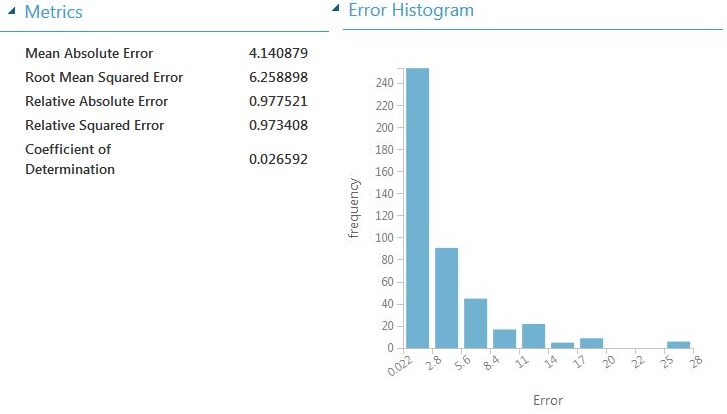
## Model strojnog učenja

Model strojnog učenja vrlo je sličan onome što smo radili na laboratorijskim vježbama. Prvo smo uploadali dataset, a nakon toga sastavili model. Odabrali smo 75% podataka za treniranje i 25% za evaluaciju.



**Slika 3. Azure ML eksperiment**

Za modul regresije odabrali smo Poisson Regression. Poisson Regression je jedan od nekoliko tipova regresije. Temelji se na poissonovoj distribuciji, koristi se za regresijske modele koji predviđaju brojeve, uobičajno prebrojavanja. Koristimo Clean Missing Data kako bi uklonili stupce kojima nedostaju vrijednosti. Koristimo Select Columns in Dataset kako bi rekli Azuru da želimo da uči samo iz određenih stupaca koje ćemo kasnije unositi u aplikaciju.

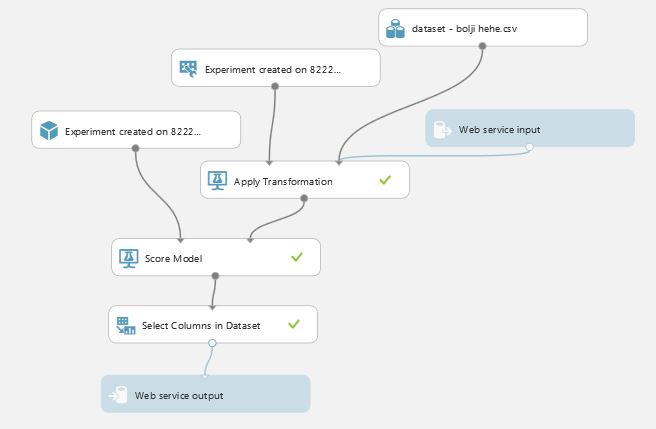
****

**Slika 4. Evaluacija modela**

Lijeva slika predstavlja evaluaciju Poisson Regresion modula za podatke postotka kakaa. Taj se modul pokazao daleko najboljim za takvu vrstu podataka. Desna slika predstavlja histogram poissonove regresije postotak kakaa u proizvodu. Drugi regresijski modeli nisu dobro odgovarali tipu podataka koji je bio postotak kakaa te je poissonova regresija imala najveći koeficijent odlučnosti uz neuronsku mrežu, neuronska mreža je imala nešto manji koefiicjent odlučnosti, ostali moduli su imali koeficijent odlučnosti manji od 0. Uspoređivali smo po koeficijentu odlučnosti jer se on može tumačiti kao omjer varijacija s objašnjenim modelom, veći broj znači da je model točniji, a ako je on 1 onda znači da model u potpunosti točno pogađa traženu vrijednost.

## Način korištenja API-ja

Nakon što je model istreniran i pruža zadovoljavajuće rezultate, potrebno ga je staviti na Azure Web Service.



**Slika 5. Model za postavljanje na Web Service**

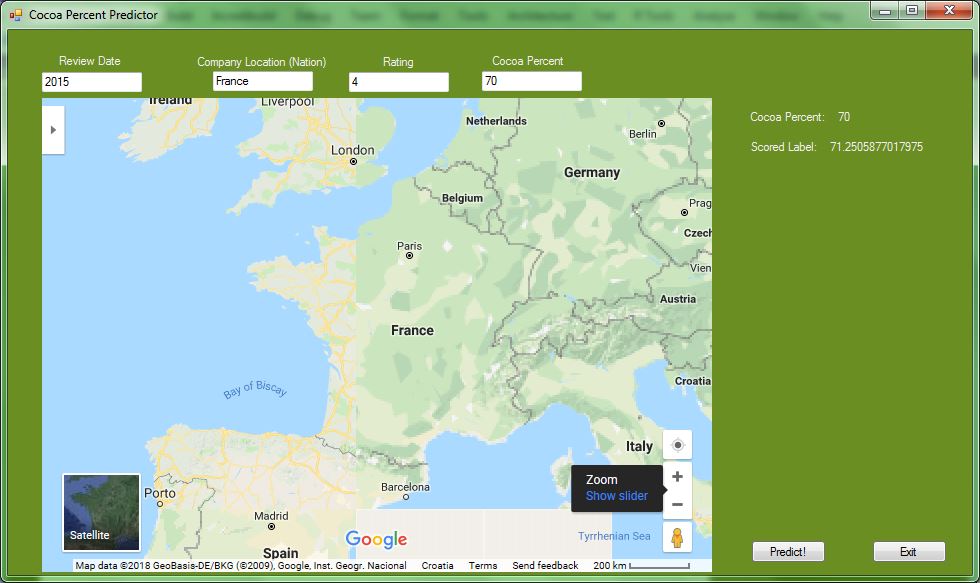
Time je napravljen API preko kojega se može putem Weba koristiti istrenirani model. Također je moguće preuzeti programski kod za klijentsku aplikaciju koji koristi taj API.

Uz taj programski kod dokumentirani su input i output. Input se sastoji od naziva stupaca i jednog ili više seta podataka. Kao output se vraća se postotak kakaa, te scored labels. Scored labels nam pokazuje predviđenu vrijednost kakaa s tim postotkom.

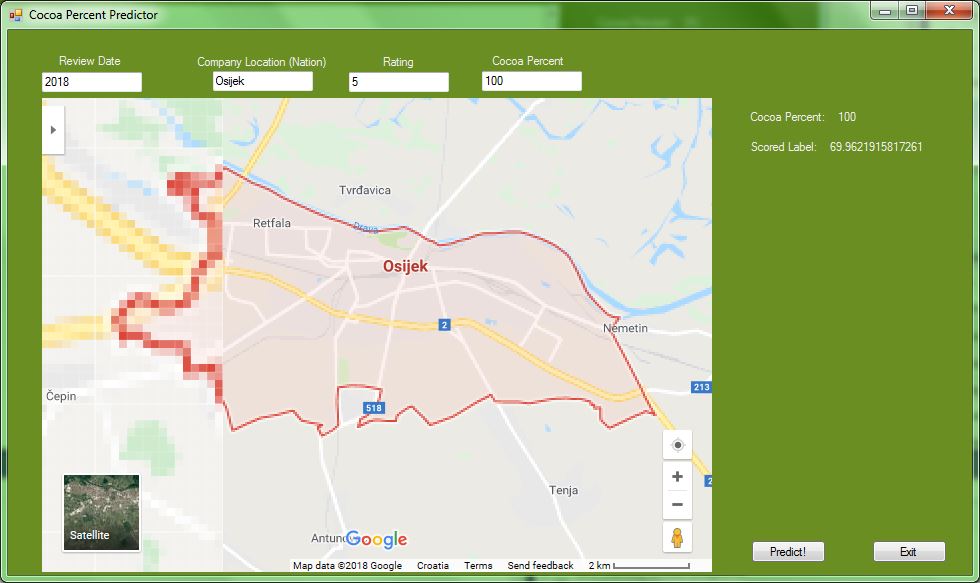
## Klijentska aplikacija

Kostur aplikacije je bio dan od strane Azure-a u C#. Aplikaciju smo prilagodili tako da prima unos od korisnika, šalje zahtjev na Azure i dobija JSON kao odgovor.

Iz JSON-a smo izvukli podatke koje smo na Azuru stavili kao izlaz (output) te smo ih stavili da se prikažu na mjestima gdje smo željeli da se prikažu. Također smo napravili mogućnost da korisnik može vidjeti koju je lokaciju unio kao jedan od parametara (company location). Na slikama 6 i 7 su prikazani primjeri rada aplikacije. Budući da alat WebBrowser u Visual Studio 2017 koristi trenutnu inačicu Internet Explorera koja je instalirana na računalu onda se dogode neke poteškoće kao što je sporije učitavanje karte i sporo ažuriranje karte ako smo nestrpljivi.



**Slika 6. Klijentska aplikacija**



**Slika 7. Klijentska aplikacija**

# Zaključak

Prilikom izrade projektnog zadatka koristili smo C# i Microsoft Azure ML Studio. Nakon umetanja podataka u Microsoft Azure i kreiranja ML modela, dobili smo kostur API-a kojeg smo oblikovali na svoj način tako da korisniku omogućimo konzolnu aplikaciju i lakše rukovanje. Korisnik unosi željene podatake, program ih predaje Azure-u, te kao povratnu informaciju dobiva predikciju u obliku stringa. Kako bi poboljšali predikciju, potrebno je koristiti bolji dataset iz kojeg će model učiti ili je potrebno koristiti drugačiju kombinaciju podataka i odabrati modul koji bi najbolje radio s tim podacima. Komunikaciju s Azure-om se uspostavlja tek nakon što pritisnemo Predict! gumb i istovremeno se učitava i karta na WebBrowser alatu. Kao povratnu informaciju dobijemo postotak kakaa i njegovu predviđenu vrijednost.

# Poveznice i literatura

Programskom je rješenju moguće pristupiti preko:

|  |
| --- |
| https://github.com/Dbirtic/Projekt-Birtic\_Huljak |

1. <https://www.kaggle.com/spectoro/chocolate/data>
2. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/poisson-regression>
3. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/data-input-and-output>
4. Laboratorijske vježbe
5. <https://docs.microsoft.com/en-us/azure/machine-learning/studio/evaluate-model-performance>
6. <https://developer.telerik.com/topics/machine-learning/consuming-azure-machine-learning-asp-net-core/>