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.

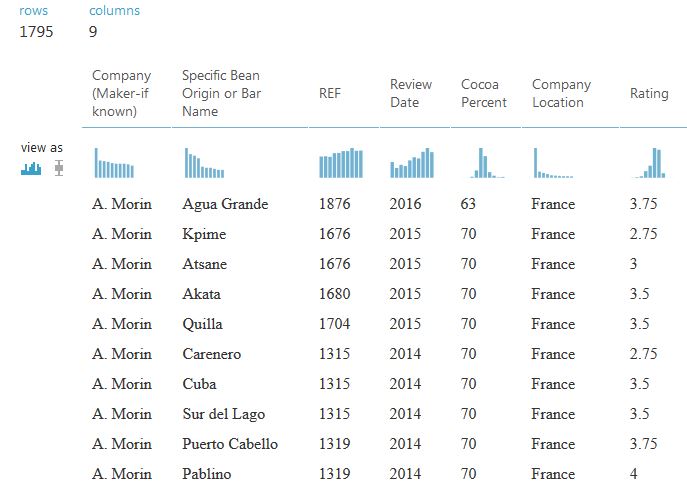
# Opis problema

Podatak koji trebamo pronaći je mogućnost postojanja kakaa s danim postotkom. Podatke koje unosimo u aplikaciju su država u kojoj se nalazi firma koja proizvodi kakao, ocjena kakaa, godina ocjene kakaa i postotak kakaa. Podatke koje unosimo smo odabrali prema tome koji podaci najbolje opisuju neki proizvod. Država u kojem se nalazi firma koja proizvodi kakao sadrži nazive država koje bi uobičajni korisnici najlakše unijeli npr. Francuska, Belgija i sl. Ocjena kakaa i godina ocjenjivanja kakaa su odabrani iz sličnog razloga. Naravno da korisnici žele znati koliki bi postotak kakaa imao kakao koji je ocjenjen s 5 i koji je nedavno dobio tu ocjenu. Postotak kakaa smo stavili kao podatak koji se može unijeti iz razloga što nam treba ipak nekakva referentna vrijednost s kojom možemo usporediti rezultat.

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.

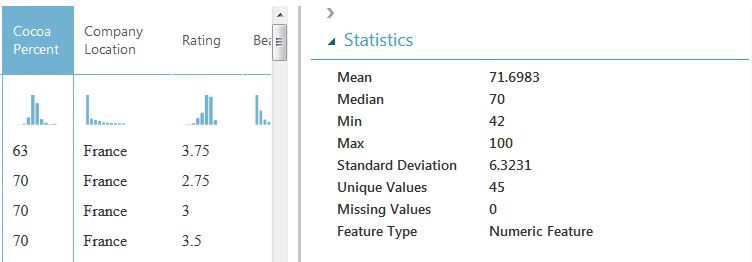
## Korišteni podaci

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 dataset-a sadrže nekoliko desetaka ili nekoliko stotina izgubljenih vrijednosti, onda ćemo koristiti ovih 7 stupaca kako bi istrenirali model. Podatke smo predobradili s modulima Clean Missing Data i Select Columns in Dataset, kasnije se spomene zašto.

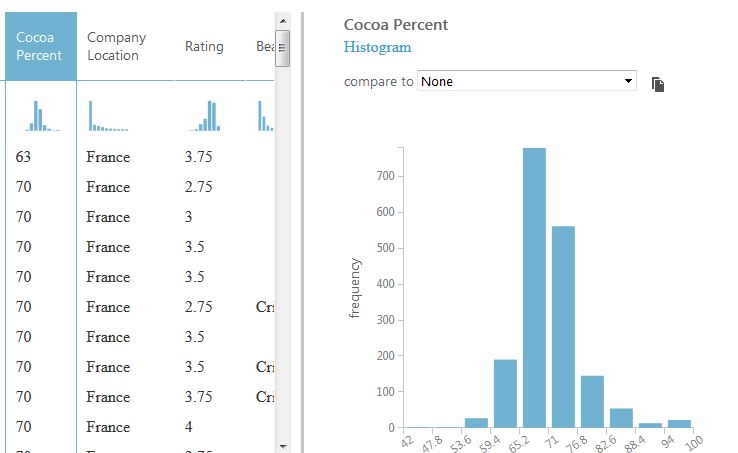


**Slika 1. Primjer dataset-a**

Na slici 1. dan je primjer korištenog dataset-a. Parametri su firma u kojoj se proizvodi kakao, mjesto gdje se uzgaja ta sorta kakaa, referenti broj zrna kakaa (REF), godina kada je kakao ocjenjen, postotak kakaa, država u kojoj se firma koja proizvodi nalazi te ocjena kakaa. Sveukupno je 1795 redova podataka od kojih je 75% korišteno za treniranje modela, a ostalih 25% za evaluaciju.

****

**Slika 2. Statistika postotka kakaa**

****

**Slika 3. Vizualizacija postotka kakaa**

Iz slika 2. i 3. saznajemo da je srednja vrijednost postotka kakaa u skupu podataka oko 71.69, te saznajemo minimalnu i maksimalnu vrijednost postotka kakaa, te na slici 3. vidimo koliko se često pojavljuju određeni postotci kakaa te vidimo one koji su najučestaliji.

## Korišteni postupci strojnog učenja

Strojno učenje je vrsta obrade podataka gdje se računalo može utrenirati da ono 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 korišena regresija. Regresija predstavlja oblik nadziranog učenja. 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.

U Azure ML Studio imamo nekoliko različitih modula za regresiju ponuđenu. Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression, Fast Forest Quantile Regression, Linear Regression, Neural Network Regression, Ordinal Regression i Poisson Regression. U početku smo zanemarili Fast Forest Quantile Regression i Ordinal Regression zato što nećemo unositi brojeve kvantila (fast forest quantile regression) i ne treba nam predviđanje rednih vrijednosti (ordinal regression).

Bayesian Linear Regression trenira i predviđa na temelju Bayesove statistike. Bayesov pristup koristi linearnu regresiju dopunjenu dodatnim informacijama u obliku prethodne raspodjele vjerojatnosti. Prethodne informacije o parametrima kombiniraju se s vjerojatnom funkcijom za generiranje procjena parametara.

Boosted Decision Tree Regression stvara skupinu regresijskih stabala pomoću pojačanja. Povačanje znači da svako stablo ovisi o prethodnim stablima. Algoritam uči prilagodavanjem preostalih stabala koja su mu prethodila. Stoga, povećanje u skupu stabla odlučivanja poboljšava točnost s malim rizikom manjeg pokrića. Boosted Decision Tree koriste učinkovitu implementaciju algoritma za poticanje MART gradijenta. Podizanje gradijenta je tehnika strojnog učenja za probleme regresije. Ona gradi svaki regresijski stablo u korak-mudar način, koristeći unaprijed definiranu funkciju gubitka za mjerenje pogrešaka u svakom koraku i ispravljanje za nju u sljedećem. Prema tome, model predviđanja zapravo je skup slabijih modela predviđanja.

Decision Forest Regression stvara regresijski model koji se temelji na skupu stabala odlučivanja. Stabla odluke su neparametarski modeli koji izvode niz jednostavnih testova za svaku instancu, prelazeći strukturu podataka o binarnom stablu sve dok se ne postigne čvor lišća (odluka). Stabla odluke imaju sljedeće prednosti:  
a) Oni su učinkoviti u računanju i korištenju memorije tijekom treninga i predviđanja.  
b) mogu predstavljati nelinearne granice odluke.  
c) Izvršavaju integrirani odabir i klasifikacija svojstava i otporni su u prisutnosti bučnih značajki.  
Ovaj regresijski model sastoji se od cjeline stabala odlučivanja. Svako stablo u regresijskoj odluci šuma daje Gaussovu distribuciju kao predviđanje. Agregiranje se izvodi preko ansambla stabala kako bi se pronašla Gaussova distribucija koja je najbliža kombiniranoj distribuciji svih stabala u modelu.

Linear Regression se koristi za izradu linearnog regresijskog modela za upotrebu u eksperimentu. Linearna regresija pokušava uspostaviti linearni odnos između jedne ili više nezavisnih varijabli i numeričkog ishoda, ili ovisne varijable. Linearna regresija je uobičajena statistička metoda, koja je usvojena u strojnom učenju i poboljšana je s mnogim novim metodama za ugradnju linije i pogreške u mjerenju. U najosnovnijem smislu, regresija se odnosi na predviđanje numeričkog cilja. Linearna regresija je i dalje dobar izbor kada želite vrlo jednostavan model za osnovni prediktivni zadatak. Linearna regresija također ima tendenciju dobro funkcionirati na visokim dimenzionalnim, rijetkim skupovima podataka koji nemaju kompleksnost.

Neural Network Regression stvara regresijski model korištenjem prilagodljivog algoritma neuronske mreže. Iako su neuronske mreže široko poznate za uporabu u dubokom učenju i modeliranju složenih problema kao što su prepoznavanje slike, lako se prilagođavaju problemima regresije. Bilo koja klasa statističkih modela može se nazvati neuronskom mrežom ako koriste prilagodljive težine i može približiti nelinearne funkcije njihovih ulaza. Stoga je regresija neuronske mreže pogodna za probleme u kojima tradicionalni regresijski model ne može stati u rješenje. Regresija neuronske mreže je nadzirana metoda učenja, pa stoga zahtijeva označeni skup podataka koji uključuje stupac s oznakom. Budući da model regresije predviđa numeričku vrijednost, stupac oznake mora biti numerička vrsta podataka.

Poisson Regression stvara model poissonove regresije. Poissonova regresija namijenjena je korištenju u regresijskim modelima koji se koriste za predviđanje numeričkih vrijednosti, obično se broji. Stoga biste trebali koristiti ovaj modul da biste izradili regresijski model samo ako vrijednosti koje pokušavate predvidjeti odgovara sljedećim uvjetima:  
a) Varijabla odgovora ima Poissonovu distribuciju.  
b)  Brojevi ne mogu biti negativni. Metoda neće uspjeti ako pokušate upotrijebiti ga s negativnim oznakama.  
c)  Poissonova distribucija je diskretna distribucija; stoga, nema smisla upotrijebiti ovu metodu s ne-cjelovitim brojevima.

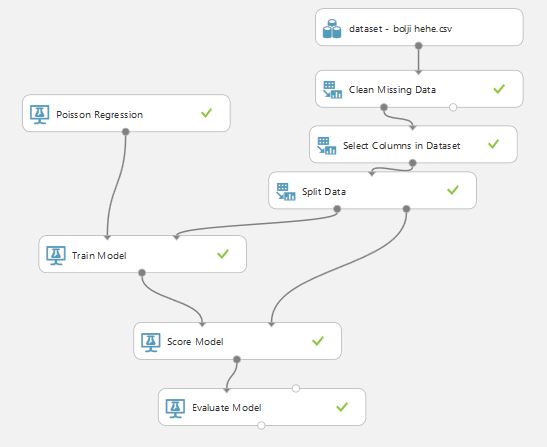
Poissonova regresija je posebna vrsta regresijske analize koja se obično koristi za modeliranje brojeva. Na primjer, Poissonova regresija bila bi korisna u ovim scenarijima:  
a) Modeliranje broja prehlada povezanih s avionskim letovima  
b) Procjena broja poziva hitnih službi tijekom događaja  
c) Projiciranje broja upita kupaca nakon promocije  
d) Stvaranje tablica za slučaj opasnosti  
Budući da varijabla odgovora ima Poissonovu distribuciju, model daje različite pretpostavke o podacima i njegovoj raspodjeli vjerojatnosti od, recimo, regresije najmanjeg kvadrata. Zbog toga se Poissonovi modeli trebaju interpretirati drugačije od ostalih regresijskih modela.

# 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.

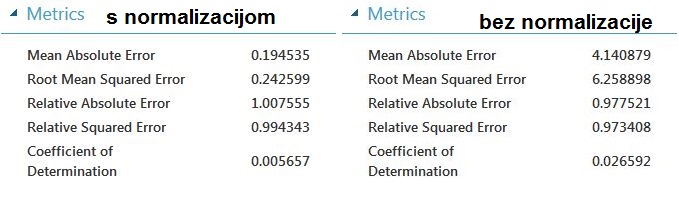
## Model strojnog učenja

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.

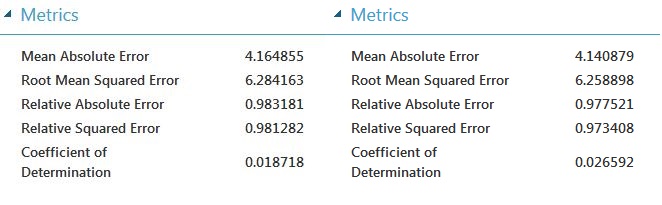


**Slika 4. Azure ML eksperiment**

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, uobičajeno prebrojavanja. Koristimo Clean Missing Data kako bi uklonili stupce kojima nedostaju vrijednosti. Koristimo Select Columns in Dataset kako bi rekli Azure-u da želimo da uči samo iz određenih stupaca koje ćemo kasnije unositi u aplikaciju. Nad podacima nismo vršili postupke normalizacije ili druge metode skaliranja podataka jer smo dobili slabije rezultate kada smo normalizirali podatke i zato što Poisson Regression najbolje radi s cijelim brojevima. Usporedbu vidimo na slici 5.



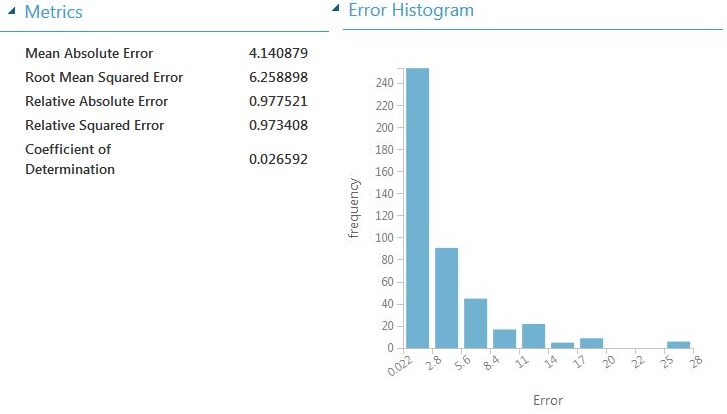
**Slika 5. Usporedba modela s normalizacijom (lijevo) i bez normalizacije (desno)**



**Slika 6. Usporedba modela koji koristi sve stupce (lijevo) i modela koji koristi odabrane stupce (desno)**

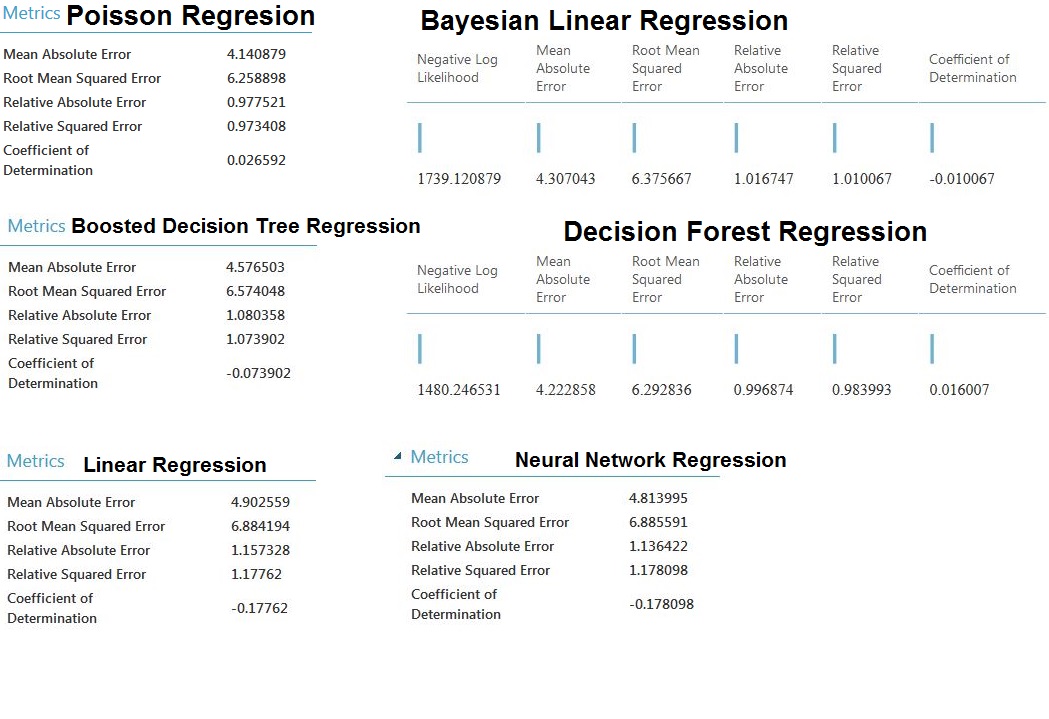
Možemo zaključiti iz slike 6. da nam se više odgovara koristiti određene stupce nego sve stupce jer time biramo točno podatke koje želimo da model strojnog učenja uči i brže uči i trenira jer koristi manje podataka.

Izraz "error" ovdje predstavlja razliku između predviđene vrijednosti i stvarne vrijednosti. Apsolutna vrijednost ili kvadrat ove razlike se obično izračunavaju kako bi se prikupila ukupna veličina pogreške u svim slučajevima, jer bi razlika između predviđene i stvarne vrijednosti u nekim slučajevima mogla biti negativna. Mjerenje pogrešaka mjeri predviđajuće performanse regresijskog modela u smislu srednjeg odstupanja svojih predviđanja od stvarnih vrijednosti. Niže vrijednosti pogrešaka znače da je model točniji pri izradi predviđanja. Ukupni mjerni podatak o pogrešci od 0 znači da model savršeno odgovara podacima.  
  
Koeficijent određivanja, koji je također poznat kao R kvadrat, također je standardni način mjerenja koliko dobro model odgovara podacima. Može se tumačiti kao omjer varijacija objašnjenim modelom. Veći udio je bolji u ovom slučaju, gdje 1 označava savršeno mjesto.

****

**Slika 7. 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 za 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 koeficijent 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.

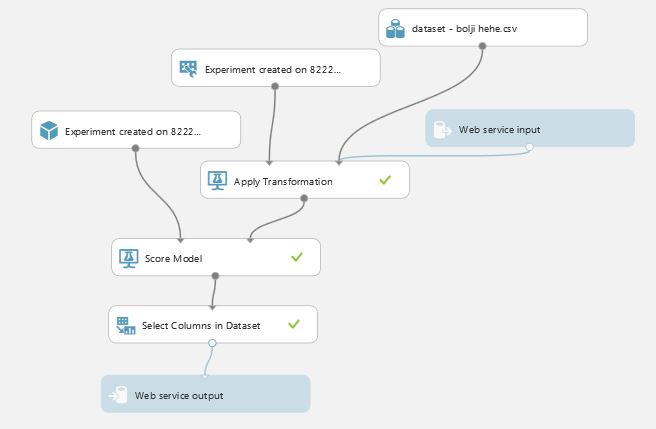


**Slika 8. Usporedba svih modula regresije**

Na slici 8. vidimo ponašanje, tj. evaluaciju modula regresije koje smo mogli koristiti u Azure ML Studio. Budući da smo se koristili Coefficient of Determination kao vodilju koja nam govori koji je modul najtočniji onda smo prvo odbacili module koji su imali negativnu vrijednost jer ako je koeficijent negativan onda nešto ne valja u korištenju tog modula (skup podataka i sl.). Nakon što smo zanemarili module čiji je koeficijent bio manji od 0 onda smo dobili poissonovu regresiju i decision forest regression. Budući da poissonova regresija ima veći koeficijent odlučnosti, onda smo odabrali poissonovu regresiju.

## Način korištenja API-ja

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



**Slika 9. Model za postavljanje na Web Service**

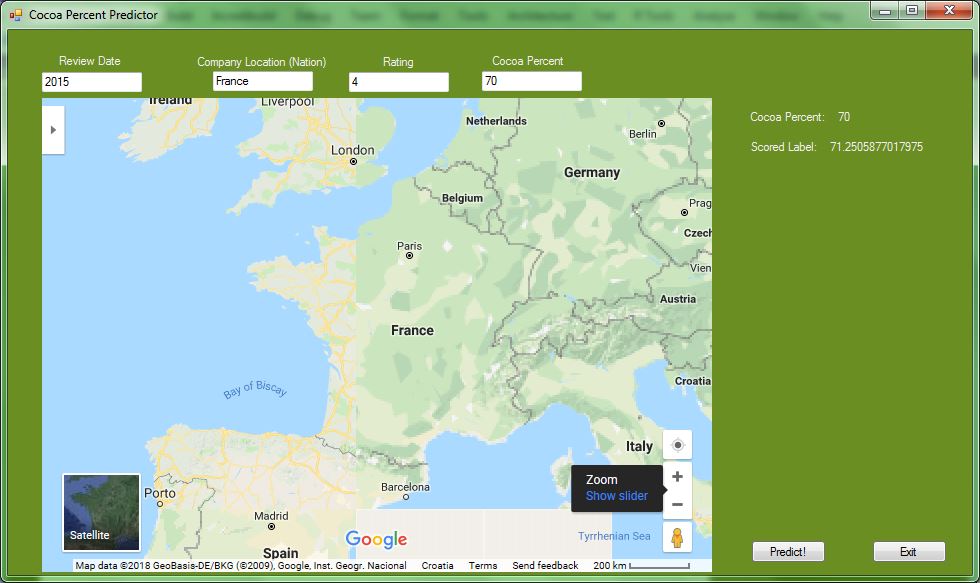
Time je napravljen API preko kojega se može putem Weba koristiti istrenirani model. Također je moguće preuzeti programski kod za klijentsku aplikaciju koji koristi taj API.

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 sepostotak kakaa, tescored labels. Scored labels nam pokazuje predviđenu vrijednost kakaa s tim postotkom.

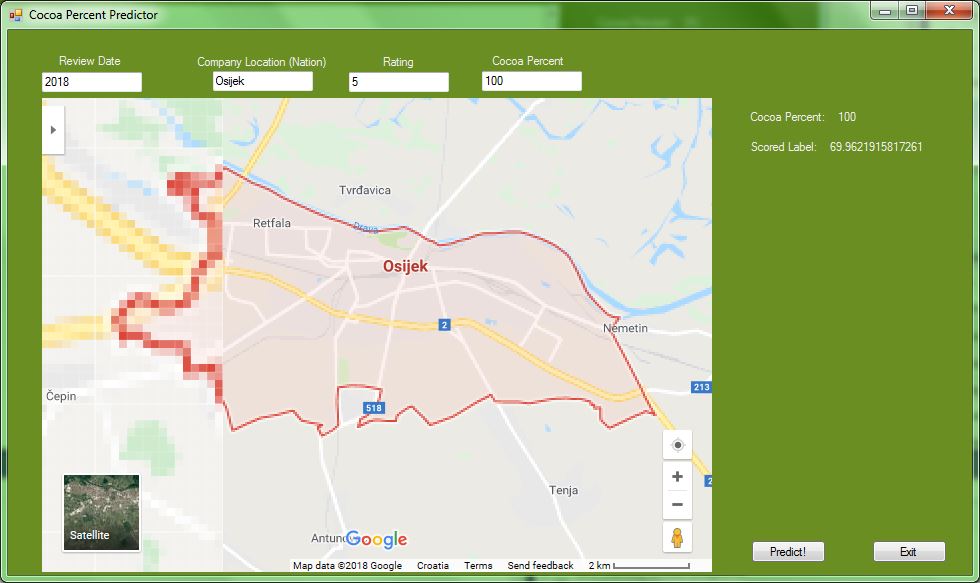
## Klijentska aplikacija

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 Azure-u 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 10 i 11 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 10. Klijentska aplikacija**



**Slika 11. 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 desktop 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/>