Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek

Računarstvo usluga i analiza podataka

SEMINARSKI RAD

„Predviđanje broja gledatelja“

Josip Tomaić

Andrej Agatić

Osijek, 2016.

**Sadržaj**

[1. Uvod 1](#_Toc461439990)

[2. Opis problema 1](#_Toc461439991)

[2.1. Korišteni podaci 2](#_Toc461439992)

[2.2. Korišteni postupci strojnog učenja 4](#_Toc461439993)

[3. Opis programskog rješenja 5](#_Toc461439994)

[3.1. Model strojnog učenja 8](#_Toc461439995)

[3.2. Klijentska aplikacija 13](#_Toc461439996)

[4. Zaključak 18](#_Toc461439997)

[5. Poveznice i literatura 18](#_Toc461439998)

# Uvod

U ovom seminarskom radu biti će objašnjen postupak predviđanja broja gledatelja u Njemačkoj, točnije u Bundesligi, na temelju parametara definiranih u podacima koji su pronađeni na različitim web lokacijama. Budući da je skup podataka velik odabrani su podatci posljednje sezone 2015./16. Na osnovu podataka potrebno je napraviti model koji je približno dobro predstavljati podatke i jednako uspješno predviditi traženi podatak te sve popratiti aplikacijom koja će korisniku omogućiti korištenje predviđanja u svrhu povećanja dobiti, poboljšavanja uvjeta i sl. Ova tema je odabrana iz razloga što je pračenje nogometnih utakmica veoma zastupljeno u Njemačkoj, a tako i u ostalim krajevima svijeta, ali glavni razlog je što smo ja i kolega veliki obožavatelji ovog sporta.

# Opis problema

Problem koji se razmatra u ovom seminarskom radu je, kao što je već navedeno, broj gledatelja na utakmici. Na ovaj problem utjeću razni faktori koji često mogu biti i veoma nepredvidivi. Najčešći faktor koji možda i najviše utjeće na rezultat predviđanja i na stvaran podatak o posjećenosti je vrijeme. Vrijeme može biti veoma nepredvidiv i varirajući parametar kao što je često i slučaj. Kao primjer možemo navesti jednu nogometnu utakmicu. Na početku utakmice vrijeme je bilo sunčano (vedro) te je stadion bio gotovo ispunjen do posljednjeg mjesta, međutim, u jednom trenutku se vrijeme jako pogoršalo i počela je padati kiša što je uzrokovalo da dio navijača napusti stadion. Tijekom istraživanja je uočeno da unatoč lošem vremenu, navijači u Njemačkoj dolaze podržati nogometni klub koji podržavaju te često stadioni budu puni i do posljednjeg mjesta. Idući faktor koji može utjecati na posjećenost, a i usko je povezan sa vremenskom prognozom je temperatura. Ukoliko je temperatura trenutno ispod 0 °C postoji mogućnost da neki od navijača odustanu od pračenja utakmice izravno na stadionu. Ostali parametri koji su promatrani tijekom istraživanja ovog problema nisu toliko nepredvidivi i večinom su konstantni.

Pronađena su i dva slična problema koja je moguće vidjeti na sljedećim web lokacijama:

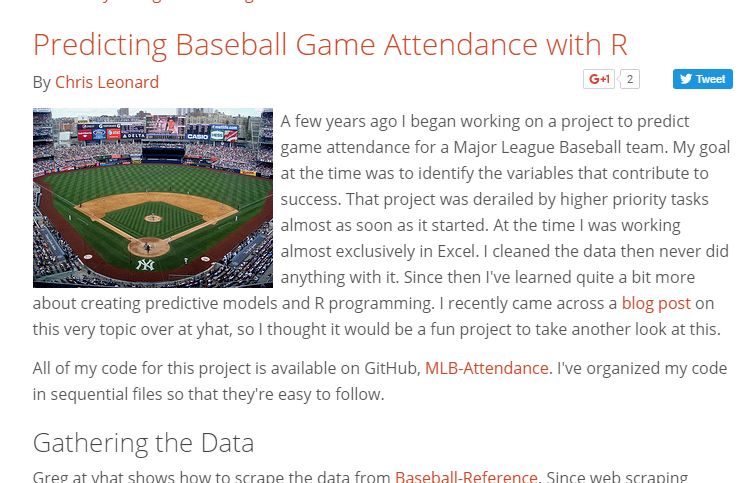
<http://ieeexplore.ieee.org/document/7264306/>

<https://r-dir.com/blog/2015/02/predicting-baseball-game-attendance-with-r.html>

Prva poveznica prikazuje projekt tri osobe sa bliskog istoka i obuhvaća predviđanje posjećenosti bilo kojeg događaja prema sljedećoj ideji: ako korisnik pokaže interes za jednu od kategorija događaja postoji vjerojatnost da će i u budućnosti željeti posjetiti njemu slične događaje. Na taj način se stvara slika o interesima pojedinog korisnika.



Druga poveznica prikazuje članak vezan za projekt predviđanja posječenosti na utakmici baseball klubova u 1. rangu u Sjedinjenim Američkim državama. Prema tekstu koji je napisan može se zaključiti da su podaci za taj projekt lako dostupni. Nakon prikupljanja i uređivanja pomoću kreiranja modela i različitih grafičkih analiza došlo se do konačnog rezultata predviđanja.



## Korišteni podaci

Podaci koji su korišteni za realizaciju ovog problema su:

* Kapacitet stadiona
* Temperatura
* Vrijeme
* Rivalstvo
* Broj stanovnika

**KAPACITET STADIONA**

Ovaj parametar je veoma važan prilikom predviđanja jer ograničava posjećenost iako, kako može biti viđeno u rezultatima predviđanja, ponekad taj broj bude i premašen.

**TEMPERATURA**

Temperatura kao parametar je jedan od najpromjenjivijih parametara i nije ju potrebno previše objašnjavati.

**VRIJEME**

Isto kao i temperatura, vrijeme je parametar koji se konstantno mijenja, a i ponekad uvelike smanjuje ili povečava posječenost nogometne utakmice.

**RIVALSTVO**

Ovaj parametar ima direktan utjecaj na posjećenost. Ukoliko se treba odigrati utakmica između dva kluba iz istog grada pri čemu iz nekog povijesnog, nedavnog ili bilo kojeg drugog razloga navijači tih klubova nisu u međusobno dobrim odnosima. Drugi mogući razlog vezan za tu situaciju je da se klubovi bore za naslov, a tako postoji i mogućnost da se klubovi bore za opstanak u elitnom rangu natjecanja u Njemačkom nogometu pa to još više potiće ljude da dođu podržati klub koji podržavaju. Tijekom istraživanja ovog parametra u Bundesligi došlo se do raznih informacija koje nisu bile od prije poznate. Neke od njih su da je jedan klub kupio najbolje igrače iz drugog kluba što se nije svidjelo navijačima odakle su došli igrači što je izazvalo početak rivalstva između dvaju klubova. Još jedna informacija koja je uzrokovala početak rivalstva je čak i promjena trenera tj. da trener koji trenutno vodi klub raskine trenutni ugovor i odluči trenirati momčad koja je gradski rival klubu u kojem je bio. Postoje i razni drugi faktori koji mogu utjecati na rivalstvo ali nisu se pojavljivali u promatranom problemu.

**BROJ STANOVNIKA**

Broj stanovnika je parametar koji utječe na broj gledatelja, ali ne tolikom mjerom kao što utjeću ostali parametri. Razlog ne tolikog utjecaja je što broj stanovnika predstavlja trenutni broj ljudi koji žive u gradu u kojem se održava događaj bez ljudi koji spadaju u protivničke navijače i ljude iz okolnih mjesta koji su odlučili doći na stadion.

Kao što je i navedeno, podaci su prikupljani na različitim web lokacijama. Web lokacije su navedene u poglavlju „Poveznice i literatura“ te im može bilo tko pristupiti ukoliko su nekome potrebne iste informacije tj. podaci. Bitno je napomenuti da su promatrani i prikupljeni podaci za nogometnu sezonu 2015/2016 tj. za prošlu sezonu.

Na početku prikupljanja korištena je Wikipedija na kojoj su pronađeni kapaciteti stadiona za svaki od klubova koji su se prošle sezone natjecali u Bundesligi. Prilikom prikupljanja ovih podataka može se zamijetiti da dva kluba koji su te sezone tek ostvarili promociju u najviši rang natjecanja imaju i najmanji kapacitet stadiona, dok neki klubovi koji ostvaruju lošije uspjehe imaju čak i veće stadione od onih koji se natječu i u inozemnim natjecanjima osim u Njemačkom prvenstvu. Idući podatak koji je bilo i najteže naći je temperatura. Taj podatak je od velike važnosti za eksperiment pa je i od velike važnosti bilo ga pronaći. Nakon što je pronađen web izvor na kojem je moguće pronaći tražene podatke bilo je potrebno samo pratiti odigrana kola i bilježiti temperaturu. Veoma zanimljiva činjenica je da je populacija u Njemačkoj rijetko pod utjecajem temperature jer bilo koja razina temperature bila u trenutku odigravanja utakmice, može se vidjeti da je stadion skoro uvijek pun do posljednjeg mjesta. Idući parametar je pronađen na istoj lokaciji kao i temperatura, a to je vrijeme. Za ovaj parametar se također očekuje da utjeće na posjećenost međutim u Bundesligi to nije slučaj. Dok su prikupljani podaci o vremenu moglo se zamijetiti da su u jednom kolu vidljive veoma male varijacije u gradovima u kojima se održava utakmica ili varijacije uopće nema.

Prikupljene vrijednosti za parametar „vrijeme“ su:

* Oblačno
* Vedro
* Olujna grmljavina
* Maglovito
* Kiša
* Snijeg

Nakon vremena prikupljani su podaci o rivalstvu između klubova koji igraju utakmicu. Za te podatke je odlučeno da će se označavati na sljedeći način:

* 0 označava da dva kluba nisu rivali
* 1 označava da dva kluba jesu rivali

Podaci o rivalstvu su također pronađeni na Wikipediji.

Zadnji parametar je broj stanovnika koji je pronađen na Wikipediji tj. prilikom pretraživanja informacija vezanih za svaki grad koji ima predstavnika u Bundesligi. Za podatke koji su prikupljeni može se reći da nisu točni za trenutni broj stanovnika jer se taj broj iz dana u dan mijenja te je nemoguće pronaći točan broj stanovnika koji trenutno živi u gradu. Podaci su vezani uz datum 31.12.2015. godine tj. datum zadnjeg popisa stanovništva u Njemačkoj.

Nakon što su podaci prikupljeni svi zajedno su povezani u oblik zapisa koji se nalazi unutar CSV datoteke (Comma separated values) tj. svaki podatak u jednom redu je odvojen od druge skupine podataka pomoću zareza. Ostala oblikovanja i prilagodbe zapisa podataka nisu bile potrebne.

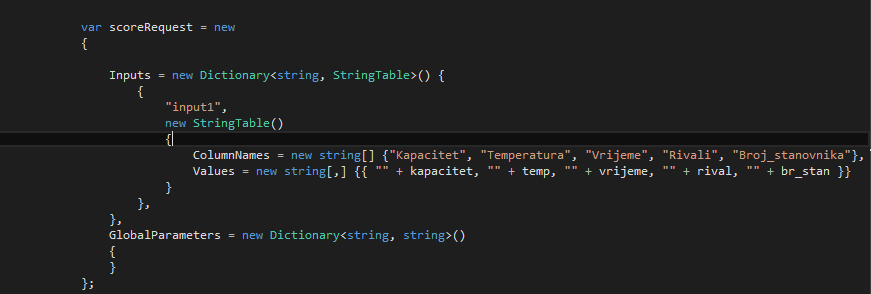
## Korišteni postupci strojnog učenja

Linearna regresija se odnosi na svaki pristup modeliranju relacija između jedne ili više varijabli označene sa *Y*, te jedne ili više varijabli označene sa *X*, na način da takav model linearno ovisi o nepoznatim parametrima [estimiranih](https://hr.wikipedia.org/w/index.php?title=Teorija_estimacije&action=edit&redlink=1) iz [podataka](https://hr.wikipedia.org/wiki/Podatak). Bila je prvi tip [regresijske analize](https://hr.wikipedia.org/w/index.php?title=Regresijska_analiza&action=edit&redlink=1) koja je detaljno proučavana i koja se ekstenzivno koristila u praktičnim primjenama. Razlog za ovo je taj što se modeli koji linerano ovise o svojim nepoznatim parametrima lakše modeliraju nego modeli sa nelinearnom ovisnošću o parametrima. Također, statistička svojstva rezultirajućih estimatora se lakše određuju. Linearna regresija ima mnogo praktičnih primjena.

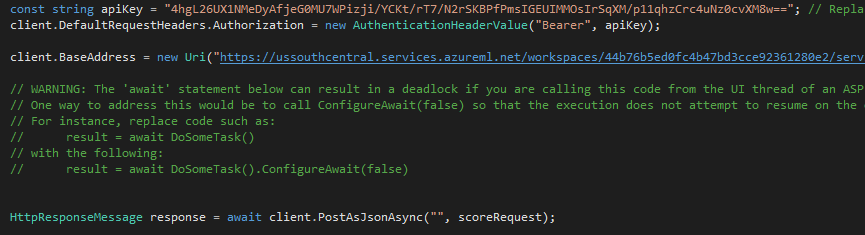
Ako je cilj [predviđanje](https://hr.wikipedia.org/w/index.php?title=Predvi%C4%91anje&action=edit&redlink=1) ili [prognoza](https://hr.wikipedia.org/wiki/Prognoza), linearna regresija se može koristiti za podešavanje preditivnog modela prema promatranom skupu podataka vrijednosti *Y* i *X*. Nakon razvoja ovakvog modela, ako je data vrijednost za *X* bez pripadajuće vrijednosti *Y*, podešeni model se može koristiti za predviđanje vrijednosti *Y*.

# Opis programskog rješenja

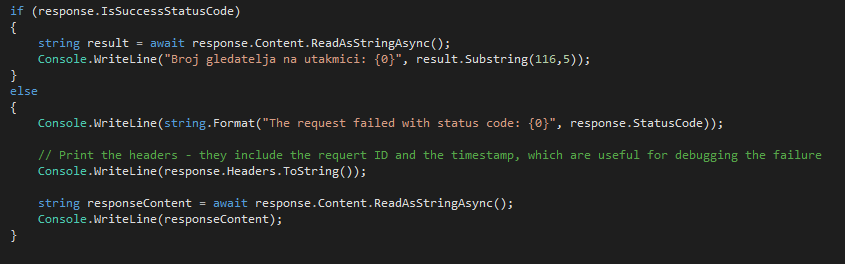
Za realizaciju programskog rješenja korišten je C# programski jezik te razvojno okruženje Visual Studio 2015. Da bi se realiziralo programsko rješenje bilo je potrebno uzeti uzorak koda koji se nalazi unutar Web servisa Azure Machine Learning Studia. Unutar tog koda se nalazi dio za unos parametara vezanih za podatke koji su dodijeljeni modelu, dio koda koji se koristi za slanje unosa natrag u model te dio koda koji se koristi za dohvaćanje i prikaz rezultata predviđanja, a u ovom slučaju je to broj gledatelja na utakmici.



Slika . Prikaz unosa parametara

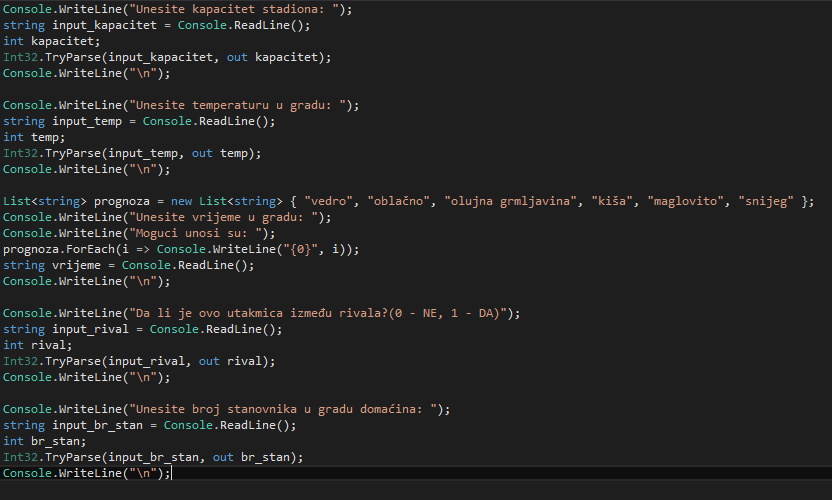


Slika . Prikaz api ključa i čekanje odgovora od strane modela



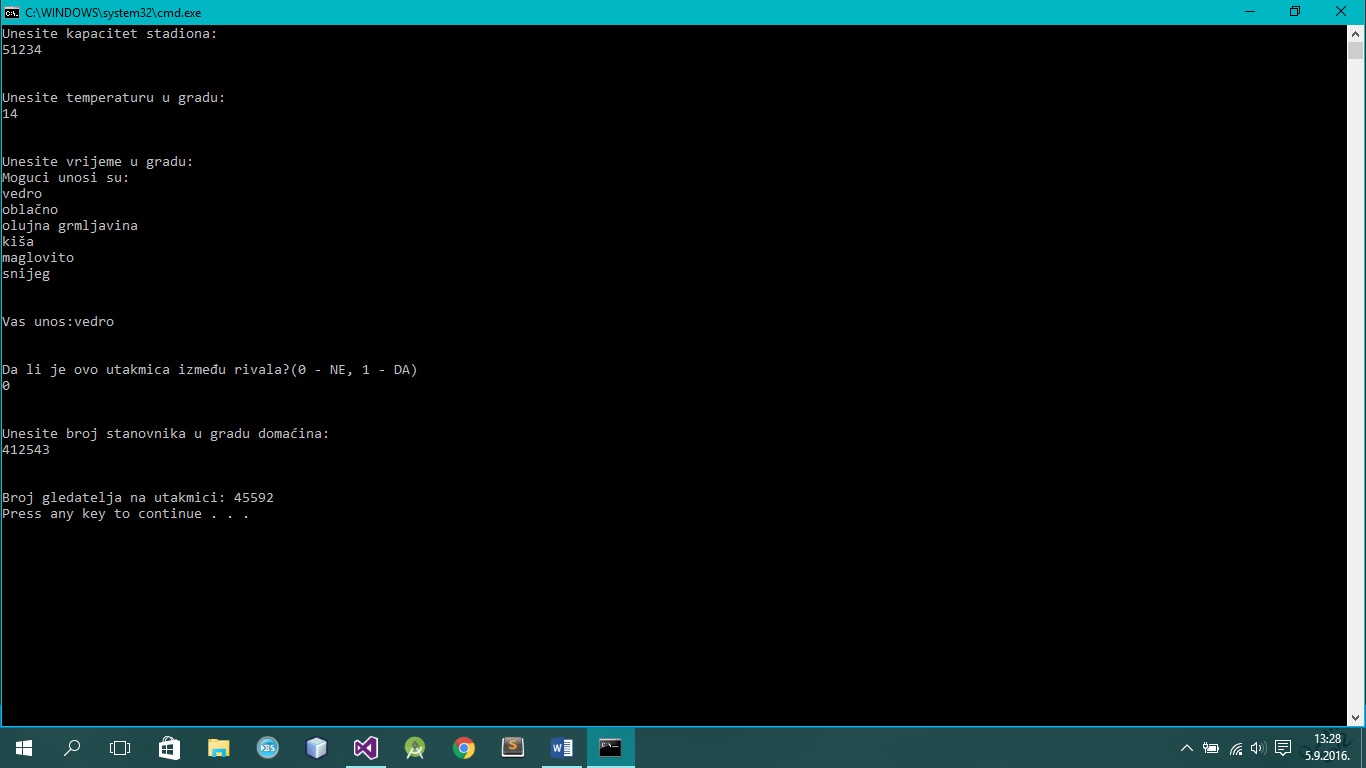
Slika . Ispis rezultata

Sav proces se odvija interakcijom korisnika sa računalom tj. unosom svih parametara putem konzole i nakon toga se unešeni parametri šalju natrag u model koji na temelju primljenih parametara predviđa broj gledatelja na nogometnoj utakmici. Kod za unos se također nalazi u programskom rješenju te ga je moguće vidjeti dolje.



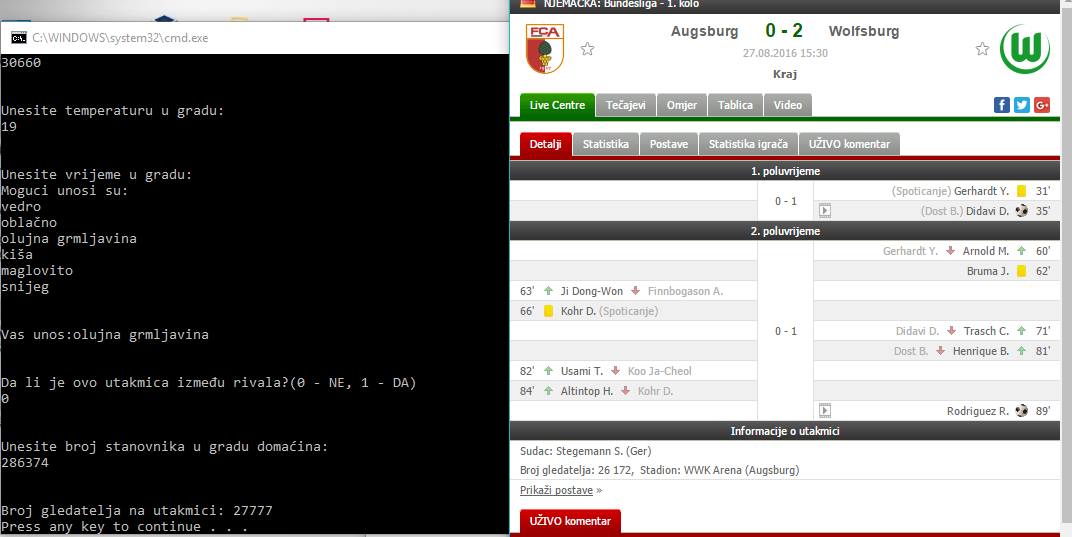
Slika . Prikaz koda za prikupljanje unešenih podataka korisnika

Nakon što su unutar konzole uneseni parametri, model koji se nalazi unutar Azure Machine Learning Studia obrađuje unesene podatke i rezultat predviđanja i unosa u konzoli moguće je vidjeti ispod.



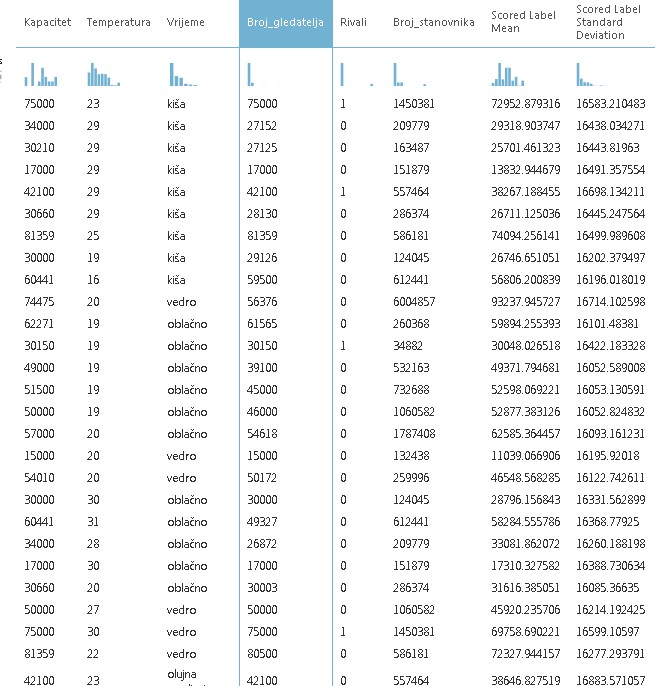
Slika . Prikaz rezultata predviđanja

Prije same izrade web sučelja i aplikacije koja je efikasnija budući da omogućuje lakše snalaženje također je i oku primamljivija potrebno je provjeriti točnost modela te samog predviđanja. Obavljeno je testiranje s podatcima koji su dostupni za sezonu 2016/2017. s kojima se model pokazao ispravan i adekvatan za daljnje korištenje i izradu na temelju njega.



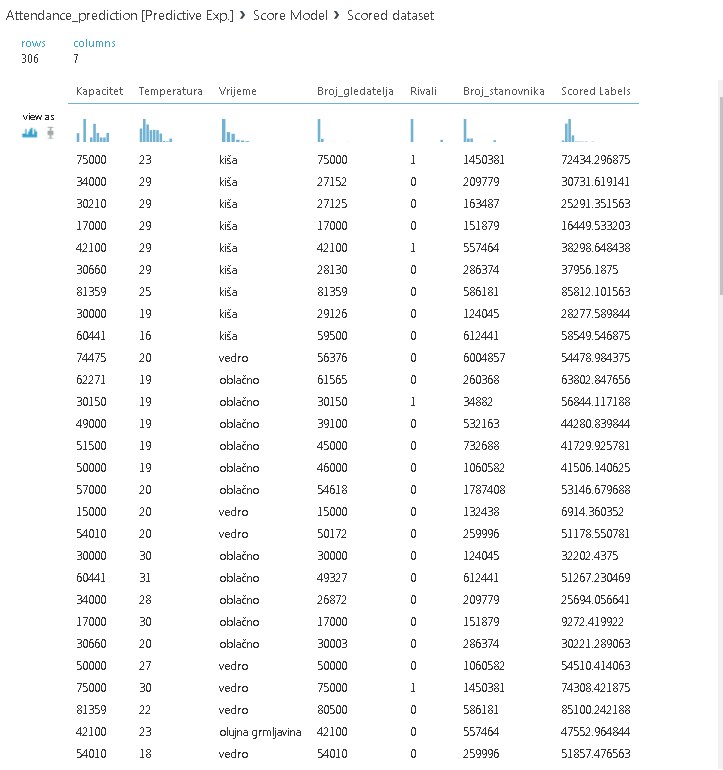
Slika . Prikaz testiranja modela na osnovu aktualnih podataka

## Model strojnog učenja



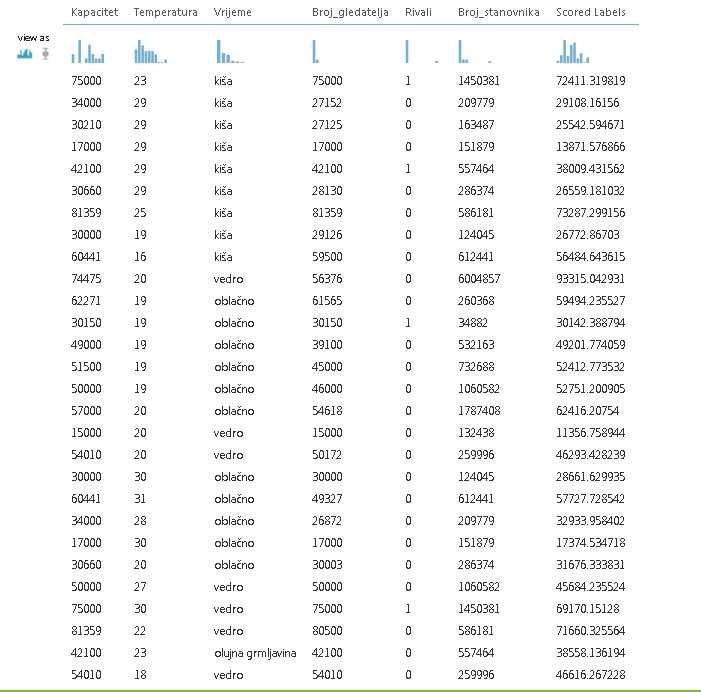
Slika . Prikaz rezultata predviđanja koristeći model s Bayesian Linear Regression

Slikom 6. prikazani su rezultati predviđanja dobivenih korištenjem modela s Bayesian Linear Regression. Vidljivo je da oblik dobivenih rezultata nije odgovarajuć stoga takav model nije moguće koristiti jer rezultati nisu upotrebljivi u obliku koji je potreban.



Slika . Prikaz rezultata dobivenih koristeći model s Boosted Decision Tree Regression

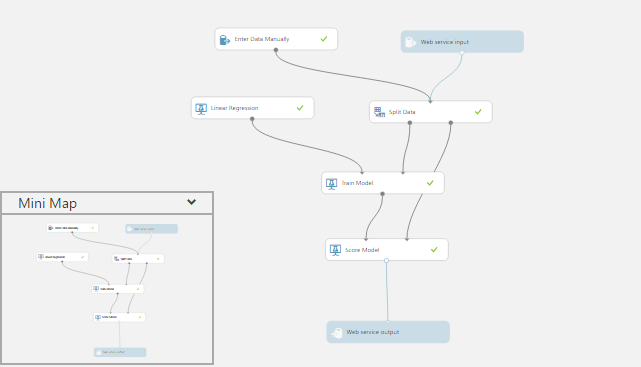
Slika 7. prikazuje rezultate dobivene korištenjem modela s Boosted decision Tree Regression. Promatrajući rezultate predviđanja vidljivo je da rezultati predviđanja broja gledatelja po pojedinom susretu u većem broju znatno odstupaju od kapaciteta stadiona odnosno broja gledatelja koji stadion može primiti.



Slika . Prikaz rezultata dobivenih koristeći model s Lienar regression

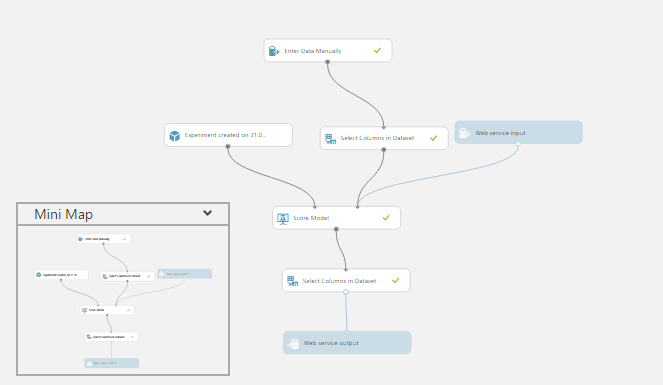
Slikom 8. prikazani su rezultati predviđanja modela koji sadrži Linear Regression komponentu. Dobiveni rezultati su zadovoljavajući stoga je korišten takav model prikazan na Slici 9.

Model strojnog učenja koji je korišten za realizaciju ovog projekta se nalazi unutar Azure Machine Learning Studia i izgleda ovako:



Slika . Prikaz Training modela

Nakon što je izrađen model za treniranje može se izraditi i „Predictive model“ koji je moguće vidjeti na slici ispod. Postoji sitna razlika između dva modela, a to je da je dodan odabir stupaca iz skupa podataka koji se predaju za treniranje, te na kraju modela je dodan još jedan odabir stupaca unutar kojeg se odabire stupac u kojem se nalaze rezultati predviđanja te se samo uzorak iz tog stupca prikazuje kao konačan rezultat u konzoli.

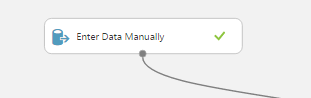


**SEGMENTI MODELA**

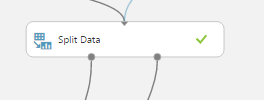
Enter data manually – koristi se za unos podatak koji se ne nalaze unutar skupine podataka koju nudi sam AML Studio. Postoji više oblika podataka i načina unosa, a odabran je oblik CSV kako je već objašnjeno u ovom seminarskom radu.

Mogući oblici unosa podataka su:

* ARFF
* CSV
* SvmLight
* TSV



Split data – segment modela koji se koristi za raspodjelu uzoraka na uzorke za trening i uzorke za vrednovanje. U ovom slučaju 75% uzoraka je dodijeljeno za trening dok je ostalih 25% ostavljeno za vrednovanje.



Train model – segment unutar kojeg se postavlja stupac čije se vrijednosti moraju dobiti treniranjem. Odabran je stupac „broj\_gledatelja“ te nakon završetka procesa rezultate je moguće vidjeti unutar stupca „Scored labels“.



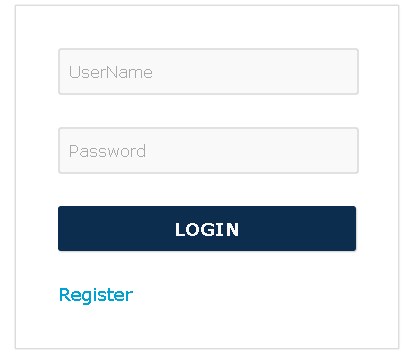
Score model – služi za prikaz rezultata predviđanja. U tablicu sa podacima se dodaje stupac „Scored labels“ u kojem se nalaze rezultati predviđanja za uzorke koji su ostali u skupu podataka za vrednovanje. U ovom slučaju broj preostalih uzoraka je bio 76.



Select Columns in Dataset – koristi se za odabir stupaca koji se uzimaju u obzir za određeni proces unutar modela. U ovom projektu je ovaj segment modela korišten za odabir stupaca koji će biti prikazani u tablici unutar Score Model segmenta te za odabir stupaca koji će biti prikazani kao rezultat u konzoli.

## Klijentska aplikacija

Klijentska aplikacija zasnovana je na web aplikaciji. Web aplikacijom korisnik pristupa stranici gdje se od njega očekuje prijava na sučelje tzv. „Login sistem“ i dopušta registracija novog korisnika ukoliko se korisnik ne nalazi u bazi podataka. Baza podataka koja sadrži korisnike pohranjena je na Azure-u.



Slika . Prikaz Login sistema

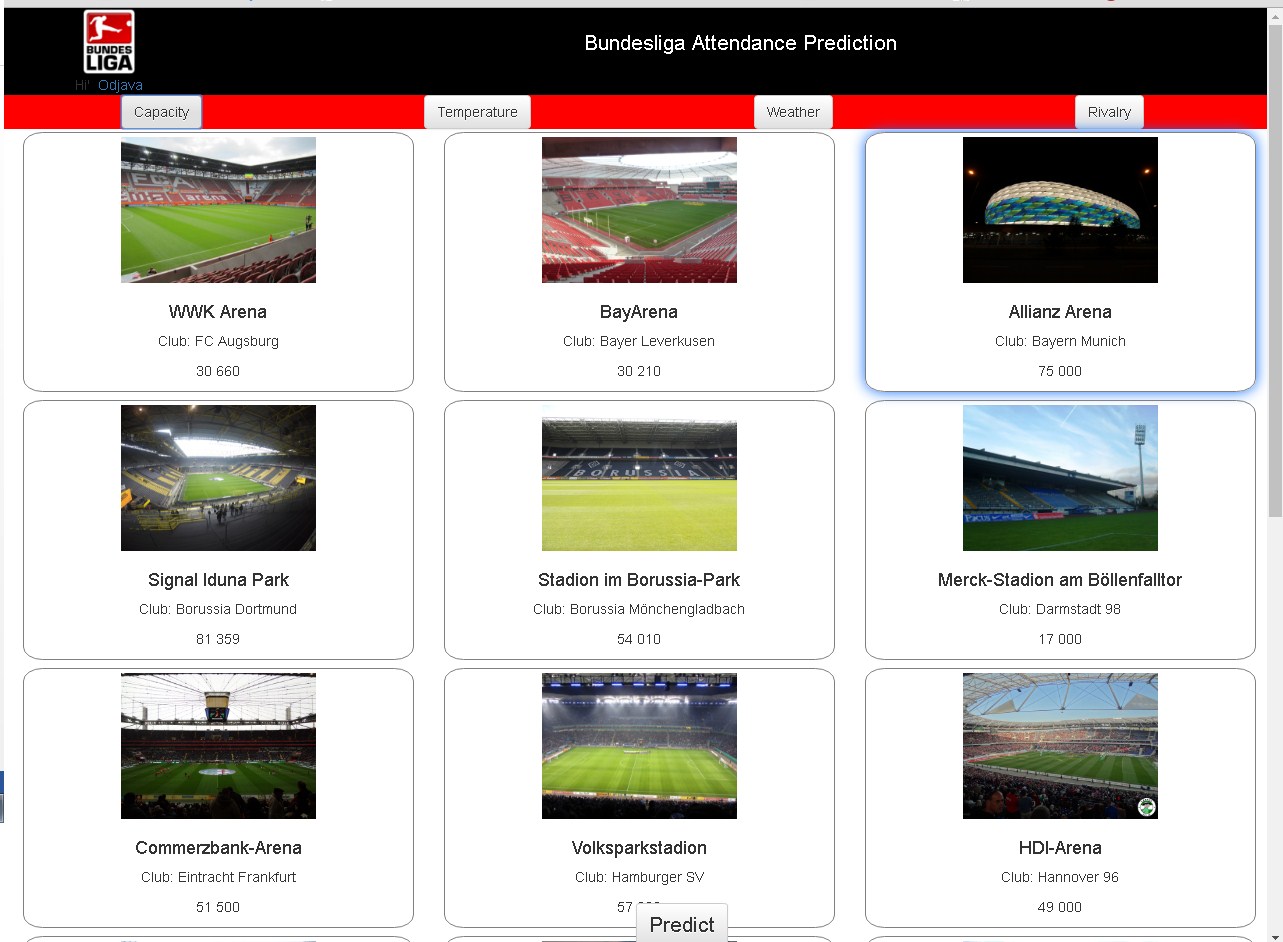
Prilikom prijave korisniku se otvara sučelje sa izbornikom gdje ima mogućnost odabrati vrijednosti parametara koji su neophodni za predviđanje. Klikom na jedan od tabova otvaraju se moguće vrijednosti za tu vrstu parametra.



Slika . Prikaz izbornika parametara

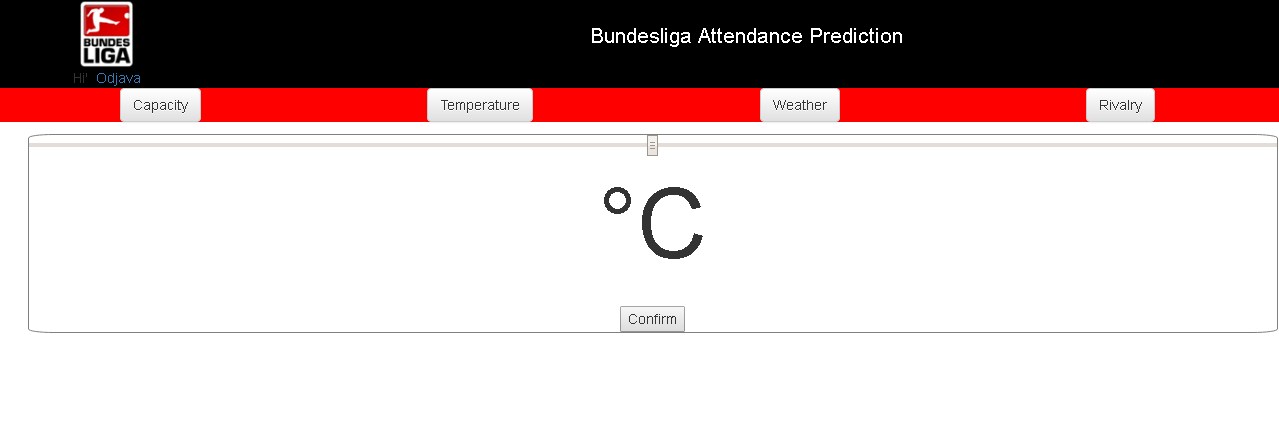
Klikom na tab „Capacity“ otvaraju se svi stadioni Bundeslige koji su realizirani pomoću gumbova sa odgovarajućim sličicama nazivom stadiona, klubom kojem stadion pripada te kapacitetom gledatelja koji stadion prima. Najbitni podataka je kapacitet budući da je neophodan za predviđanje. Odabirom bilo kojeg stadiona kapacitet se sprema u u varijablu „kapacitet“ te se kasnije pri odabiru svih nužnih parametara šalje serveru kako bi server obradio podatke i vratio predviđenu vrijednost.

Budući da se svaki stadion nalazi u određenom gradu, a svaki grad ima svoj broj stanovnika, parametar „broj stanovnika“ poprima vrijednost odmah pri odabiru stadiona. Takav sistem je korišten kako bi se izbjeglo dodatno kodiranje s obzirom da bi bilo moguće da korisnik odabere stadion te sasvim drugi grad neovisno gdje se stadion nalazi.



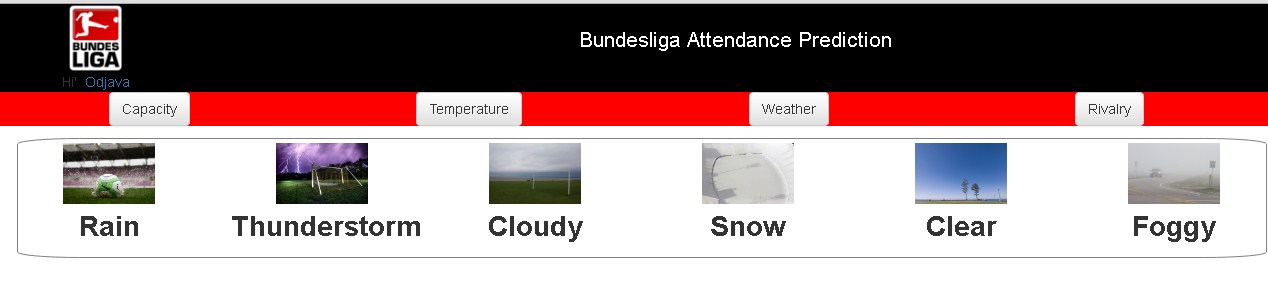
Slika . Prikaz odabira vrijednosti parametra kapaciteta

Za predviđanje osim kapaciteta također se koristi i temperatura, odabir tog podatka realiziran je klizačem koji omogućuje odabir vrijednosti temperature te klikom na gumb „Confirm“ odabrana vrijednost se sprema u varijablu te je spremna za slanje.



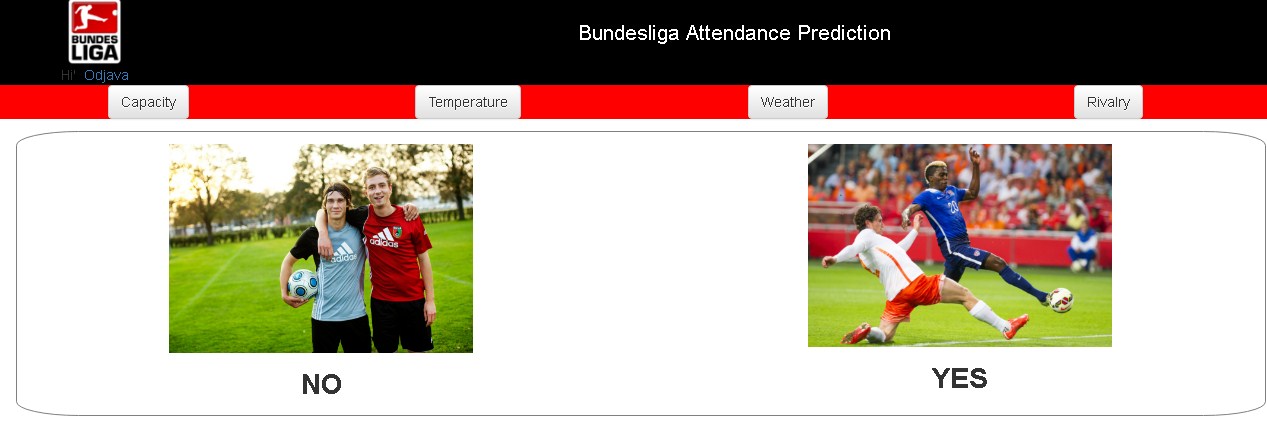
Slika .Prikaz odabira vrijednosti parametra temperature

Klikom na gumb „Weather“ u izborniku korisniku se otvaraju vrijednosti parametra vremena koji predstavlja vremenske prilike.



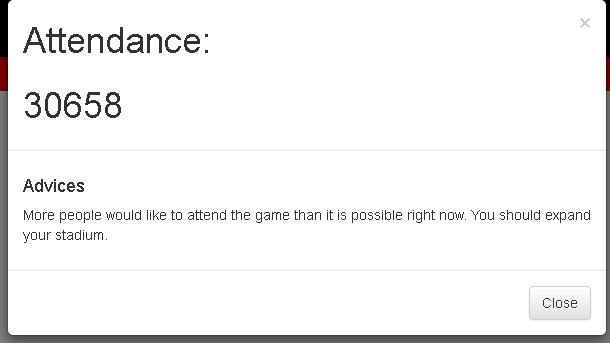
Slika .Prikaz vrijednosti parametra vremena

S obzirom da je za predviđanje potrebno pet podataka zadnji podatak je rivalitet, korisnik klikom na gumb „Rivalry“ dobije ponuđene dvije slike koje označavaju jesu li klubovi rivali ili ne. Odabirom jedne od opcija odabrani su svi podatci te je kompletan niz podataka koji se šalje serveru sada spreman.



Slika .Prikaz odabira vrijednosti parametra rivalstva

Kako bi dobili željeno predviđanje potrebno je da korisnik klikom na gumb „Predict“ koji se nalazi na dnu web sučelja da informaciju da su svi podatci odabrani te da se započne proces slanja svih parametara serveru kako bi od servera nakon obrade dobio informaciju. U nekoliko slučajeva prilikom prikupljanja podataka dogodilo se da broj gledatelja premašuje kapacitet stadiona za određenu vrijednost. Tu činjenicu smo prilikom izrade web aplikacije uzeli u obzir te prilikom ispisa rezultata ukoliko je korisnik odabrao stadion a rezultat predviđanja premašuje kapacitet stadiona za određeni iznos, korisniku se ispisuje opaska kako više ljudi želi prisustvovati utakmicu i da bi trebao proširiti stadion.



Slika . Prikaz rezultata predviđanja

# Zaključak

Svrha ovog projekta je naučiti kako iz velikog skupa podataka (Big Data) izdvojiti podatke u ovom slučaju kapacitet stadiona, temperatura, vrijeme, rivalstvo i broj stanovnika te promatrati parametre koje ti podatci mogu poprimiti. Na osnovu njih načiniti model koji odgovara podatcima te dobiti što bolje predviđanje i dobiti što bolju vrijednost za podatak koji nas zanima. U ovom slučaju rađeno je predviđanje broja gledatelja na nogometnim utakmicama njemačke Bundeslige. Takvo predviđanje može biti jako korisno, budući da svaki klub na osnovu praćenih podataka te njihovih parametara koje poprime uz određene uvjete može pokušati ispraviti određene nedostatke kako bi u što lošijim uvjetima postigao što bolje rezultate, odnosno bolju posjećenost i time povećao i imao približno jednaku i konstantnu dobit u svim uvjetima. Uz to korisno je upoznavanje sa Microsoft Azure-om budući da omogućuje razvoj, udomljivanje usluga, pohranu podataka i okolinu upravljanja uslugama.

# Poveznice i literatura

STRANICE NA KOJIMA SE NALAZE PODACI:

<http://www.rezultati.com/nogomet/njemacka/bundesliga-2015-2016/rezultati/>

<http://info.nowgoal.com/en/League.aspx?SclassID=8&MatchSeason=2015-2016>

<https://en.wikipedia.org/wiki/2015%E2%80%9316_Bundesliga>

TUTORIAL ZA RESPONSIVE DESIGN (BOOTSTRAP)

<https://www.coursera.org/learn/responsivedesign/home/week/1>

Programskom je rješenju moguće pristupiti preko:

|  |
| --- |
| [Programsko rješenje na GitHubu](https://github.com/aagatic/RuapProjekt) |
| [ML model](https://gallery.cortanaintelligence.com/Experiment/Attendance-prediction-1) |
| [Web rješenje](http://attendanceprediction.ml/) |