Usporedba algoritama rudarenja podataka u društvenim znanostima

Tomaš, Stipe

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Stipe Tomaš

USPOREDBA ALGORITAMA RUDARENJA PODATAKA U DRUŠTVENIM ZNANOSTIMA

ZAVRŠNI RAD

Varaždin, 2022.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Stipe Tomaš

JMBAG: 0016142118

Studij: Informacijski sustavi

USPOREDBA ALGORITAMA RUDARENJA PODATAKA U DRUŠTVENIM ZNANOSTIMA

ZARVRŠNI RAD

Mentorica:

Izv. prof. dr. sc. Dijana Oreški

Varaždin, kolovoz 2022.

Stipe Tomaš

Izjava o izvornosti

Izjavljujem da je moj završni/diplomski rad izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autor/Autorica potvrdio/potvrdila prihvaćanjem odredbi u sustavu FOI-radovi

_____

Sažetak

Tema ovog rada je „Usporedba algoritama rudarenja podataka u društvenim znanostima“. Na početku rada napravit će se pregled literature na temu rudarenja podataka, koja će biti temelj za razmatranja u ovom radu. Zatim slijedi teoretsko objašnjenje pojmova, odnosno algoritama vezanih uz rudarenje podataka. Konkretno, riječ je o konceptu CRISP-DM, klaster analizi, stablu odlučivanja i neuronskoj mreži. Navedeni algoritmi primijenit će se na skupu podataka o zločinima počinjenima u Los Angelesu, preuzetih sa platforme Kaggle. Za analizu ovog skupa podataka koristit će se alat BigML te će se primijeniti gore navedeni algoritmi rudarenja podataka. Na kraju rada usporedit će se dobiveni rezultati prediktivnih modela te njihova točnost.

Ključne riječi: rudarenje podataka; CRISP-DM, klaster analiza; stablo odlučivanja; neuronske mreže; BigML

Sadržaj

1. Uvod

Rudarenje podataka predstavlja prekretnicu u načinu kako promatramo podatke i kako ih koristimo. Problemi kao što su predviđanje određenih događaja i ponašanja mogu se savladati određenim tehnikama analize i grupiranja podataka te pronalaženjem poveznica između njih. [7] Kako bi se pokušao savladati problem zločina i manjka policijskih snaga, upotrijebit će se metode rudarenja na skupu podataka koji se odnosni na zločine u Los Angelesu. Cilj rada je predvidjeti područje na kojemu će se odviti neki zločin kako bi se policijske snage mogle bolje rasporediti po gradu i tako efikasnije koristiti.

Prvi korak pri izradi modela za predviđanje je razumijevanje podataka i problema koji se pokušava svladati zbog čega će se prvo analizirati svi podatci koji se nalaze u početnom skupu. Kao drugi korak, potrebno je definirati metode rudarenja koje će se koristiti nad odabranim skupom podataka. [6] Za ovaj rad odabrane su klaster metoda, stablo odlučivanja i neuronske mreže na temelju prijašnjih istraživanja. Na kraju, potrebno je usporediti dobivene rezultate.

2. CRISP-DM

Metoda rudarenja podataka može se koristiti za rješavanje širokog spektra poslovnih problema, uključujući zadržavanje kupaca i prometa, marketing baze podataka, lojalnost kupaca, predviđanje bankrota, segmentaciju tržišta, kao i analizu rizika, afiniteta i portfelja. [7]

Prije nego što krenemo analizirati podatke te izrađivati modele potrebno je objasniti standarde koji se rabe za rudarenje podataka. Tako je primjerice Institut SAS razvio SEMMA kao proces rudarenja podataka. Ima pet koraka (uzorak, istraživanje, modificiranje, modeliranje i procjena) (eng. Sample, Explore, Modify, Model, i Assess), zbog čega je dobio akronim SEMMA. [7] Ovaj rad temeljen je na procesu CRISP-DM (akronim od Cross-Industry Standard Process for Data Mining). [2] Prednost je ovog standarda da se temelji na iskustvu, fokusiran je na poslovne probleme, neovisan je na industriju i primjenu te nije ovisan ni o jednom alatu. [2] Ima šest faza:

1. Razumijevanje domene – Što je potrebno poduzeću?
2. Razumijevanje podataka – Koje podatke imamo/trebamo? Jesu li „čisti“?
3. Priprema podataka – Kako organiziramo podatke za modeliranje?
4. Modeliranje – Koje tehnike modeliranja trebamo primijeniti?
5. Evaluacija – Koji model najbolje ispunjava poslovne ciljeve?
6. Implementacija – Kako dionici pristupaju rezultatima? [6]

Tijekom ovog rada proći ćemo kroz navedene korake te bi na kraju rada trebali imati gotov model koji se može evaluirati i koji bi se na kraju mogao i primijeniti.

3. Opis problema i skupa podataka

U zadnjih nekoliko godina, zločini u SAD-u u stalnom su porastu. Između 2019. i 2020. ta brojka je skočila čak za 30%. Osim toga, broj riješenih slučajeva je pao sa 61% na 54%. Glavni razlog za to još nije poznat, no pretpostavlja se da je ekonomska kriza uzrokovana pandemijom koronavirusa jedan od glavnih razloga. Uz to, nemiri uzrokovani slučajem George Floyda dodatno su povećali tenzije između građana. Još jedan faktor u povećanom broju zločina je vjerojatno i porast prodanog oružja u posljednjih nekoliko godina. Stručnjaci se nadaju da će se ovaj trend smanjiti nakon pandemije, no za to su potrebna uvođenja novih zakonskih mjera prevencije zločina čija implementacija može potrajati [4].

Jedno od rješenja za borbu protiv rasta zločina bilo bi povećanje policajaca u radnoj snazi SAD-a. Naravno, kao i u drugim dijelovima svijeta, SAD se suočava sa nestašicom radne snage, pogotovo u polju policijskih poslova. Problem nije samo u treniranju i kvalifikaciji radnika već u općenitom interesu ljudi da se uopće zaposle kao policajci. Neki od razloga za to su povećani standardi koje je potrebno zadovoljiti za dobiti posao. Osim što je teže zadovoljiti standarde, sve više policajaca je blizu dobi za mirovinu, a mnogi odustaju i mnogo prije te dobi. [8]

Kako zločin raste, a broj policajaca pada, potrebno je bolje upravljati sredstvima koje policija ima. Jedno rješenje mogla bi biti implementacija modela za predviđanje. Kako se svaki zločin dokumentira u policijskim arhivama (koje su sada u digitalnom obliku) s mjestom i vremenom počinjenog zločina, moguće je napraviti model predviđanja sa pouzdanim skupom podataka. Ako se policijske patrole mogu bolje rasporediti po gradu, povećati će se efikasnost i moguće je ublažavanje učinaka nedostatka radnika u policiji. Ako se smanje patrole u određenom dijelu grada u određeno vrijeme, nije potrebno toliko policajaca u patroli što bi moglo pomoći u suzbijanju zločina u SAD-u.

3.1. Pregled prethodnih istraživanja

U istraživanju o metodama predviđanja zločina spomenute su razne metode kao što su metoda potpornog vektorskog stroja (SVM), više-varijantne vremenske serije (MTS) i umjetne neuronske mreže (ANN). U predviđanju, metoda SVM se koristila za predviđanje mjesta žarišta zločina jer ta metoda dobro predviđa vremenske serije, odnosno podatke koji su uzeti u različitim, ali sekvencijalnim vremenskim trenucima. Za predviđanje geo-temporalnih promjena u zločinima koristila se ANN metoda zbog svoje sposobnosti povezivanja ulaza i mogućih izlaza u vrlo kompleksnim stvarnim sustavima. MTS se nije pokazala kao najbolja metoda no uspješna je u pronalaženju trendova između različitih zločina koji se mogu koristiti

za predviđanje budućih zločina [5]. U istraživanju [3] koristila se metoda stabla odlučivanja (J48). Metoda stabla odlučivanja je jedan od najboljih algoritama za kategoričku i kontinuiranu analizu podataka. Ispostavilo se da je model dobiven tom metodom pogodan za predviđanje jer je njegova točnost bila nešto viša od 94%. Na kraju, jedna od metoda bila je i metoda GBPA (Graph-Based Progression Analyses) . Koristila se jer pomaže razumjeti kako počinitelji prelaze s jednog zločina na drugi kroz određeno vrijeme. U toj metodi, svaki čvor u grafu prikazuje učestalost zločina i prosječnu dob počinitelja na temelju čega se može izvršiti predviđanje [10].

Podaci u istraživanjima su razdvojeni na dva skupa. Jedan skup podataka se uvijek koristio za treniranje modela, a drugi skup se koristio za testiranje točnosti modela. Kao primjer, u istraživanju [1], 80% instanci se koristilo za treniranje, a 20% za testiranje. U tom istraživanju, podaci su dobiveni „web struganjem“ službenih web stranica policije indijskog grada Indora. Koristeći statističku metodu Extra Trees Clasifier (ETC) skup dobivenih atributa rangirao se prema važnosti. Atributi (po visini važnosti od najveće prema najmanjoj) su: Hour (vrijeme počinjenog zločina), Latitude, Longitude (koordinate na kojemu je počinjen zločin), Year (godina u kojoj je počinjen zločin), Month (mjesec u kojemu je počinjen zločin) i Week (tjedan u kojemu je počinjen zločin). Prema procjeni istraživača, odbačeni su atributi Year, Month, i Week zbog niskog stupnja važnosti za predviđanje. Broj instanci u skupu podataka nije specificiran. Slično kao i u prethodnom istraživanju, istraživanje [12] navodi datum zločina, koordinate i opis zločina kao atribute korištene za kreiranje modela predviđanja. Iz cijelog skupa podataka odabrani su samo ulični zločini. Odabrani zapisi podijeljeni su u tri glavne skupine prema ključnim riječima u opisu: razbojništvo (npr. 'pucnjava na ulici', 'napad') , masakr (npr. 'premlaćivanje u stanu', 'tuča na ulici') i pljačka (npr. 'krađa', 'krađa iz auta'). Kategorija “pljačka” je najbrojnija od odabranih vrsta i sadrži 142452 zapisa. Podaci su dobiveni od Ministarstva unutarnjih poslova Ruske Federacije, a oni sadrže evidenciju zločina u Sankt Peterburgu u razdoblju od 2014. do 2017. Vidimo da su najvažniji podaci za predviđanje mjesta zločina koordinate i vrijeme prijašnjih zločina te njihova klasifikacija. Pronašli smo primjere istraživanja u kojima su navedeni i dodatni atributi. U istraživanju [9] navode se atributi Država SAD-a, stanovništvo, MedIncome (srednji dohodak kućanstva), MedFamInc (srednji prihod obitelji (razlikuje se od prihoda kućanstva za neobiteljska kućanstva)), PerCapInc (dohodak po glavi stanovnika), NumUnderPov (broj ljudi ispod razine siromaštva), PctLess9thGrade (postotak osoba od 25 i više godina s obrazovanjem manje od 9. razreda), PctNotHSGrad (Postotak osoba starijih od 25 godina koji nisu maturanti), PctBSorMore (Postotak osoba starijih od 25 godina s diplomom ili visokim obrazovanjem), PctUnemployed (Postotak osoba starijih od 16 godina, u radnoj snazi i nezaposlenih), PctEmploy (postotak zaposlenih osoba starijih od 16 godina), ViolentCrimesPerPop (ukupni broj nasilnih zločina na 100.000 stanovnika) i Kategorija zločina (kategorizacija zločina u tri kategorije). Kategorija zločina je

atribut dodan u svrhu predviđanja. Novi dodani nominalni atribut ima tri vrijednosti, a to su "Niska", "Srednja" i "Visoka". Podaci su preuzeti iz popisa stanovništva SAD-a iz 1990. godine, provedenih anketa u SAD-u i određenih evidencija FBI-a. U početku je dobiveno 128 atributa, no njihov broj je smanjen na 12 (odnosno 13 s novim nominalnim atributom). Broj atributa smanjen je na temelju iskustva istraživača.

Cilj istraživanja [12] bilo je utvrđivanje koristi različitih statističkih alata za predviđanje broja kaznenih djela u određenom urbanom području. Tijekom istraživanja otkriveni su neki nedostatci modela, poput predviđanja negativnih vrijednosti, koji su napravljeni modelom linearne regresije. Jedna od korištenih metoda bila je metoda povećanja gradijenta koja je dala najbolju točnost predviđanja. Konkretan rezultat istraživanja bili bi proizvedeni modeli predviđanja od kojih je najbolji model nastao povećanjem gradijenta. Istraživanje [13] došlo je do zaključka da modeli dobiveni različitim metodama nisu jednako učinkoviti u predviđanju. Istraživanje se bavilo usporedbom algoritama za predviđanje žarišta zločina. Ostala istraživanja također su dobila razne modele predviđanja kao rezultate, a njihova točnost ovisila je o korištenoj metodi, iako su neka istraživanja koristila iste metode, a dobila vrlo različite rezultate što se može objasniti kao posljedica korištenja različitih skupova podataka za izradu modela.

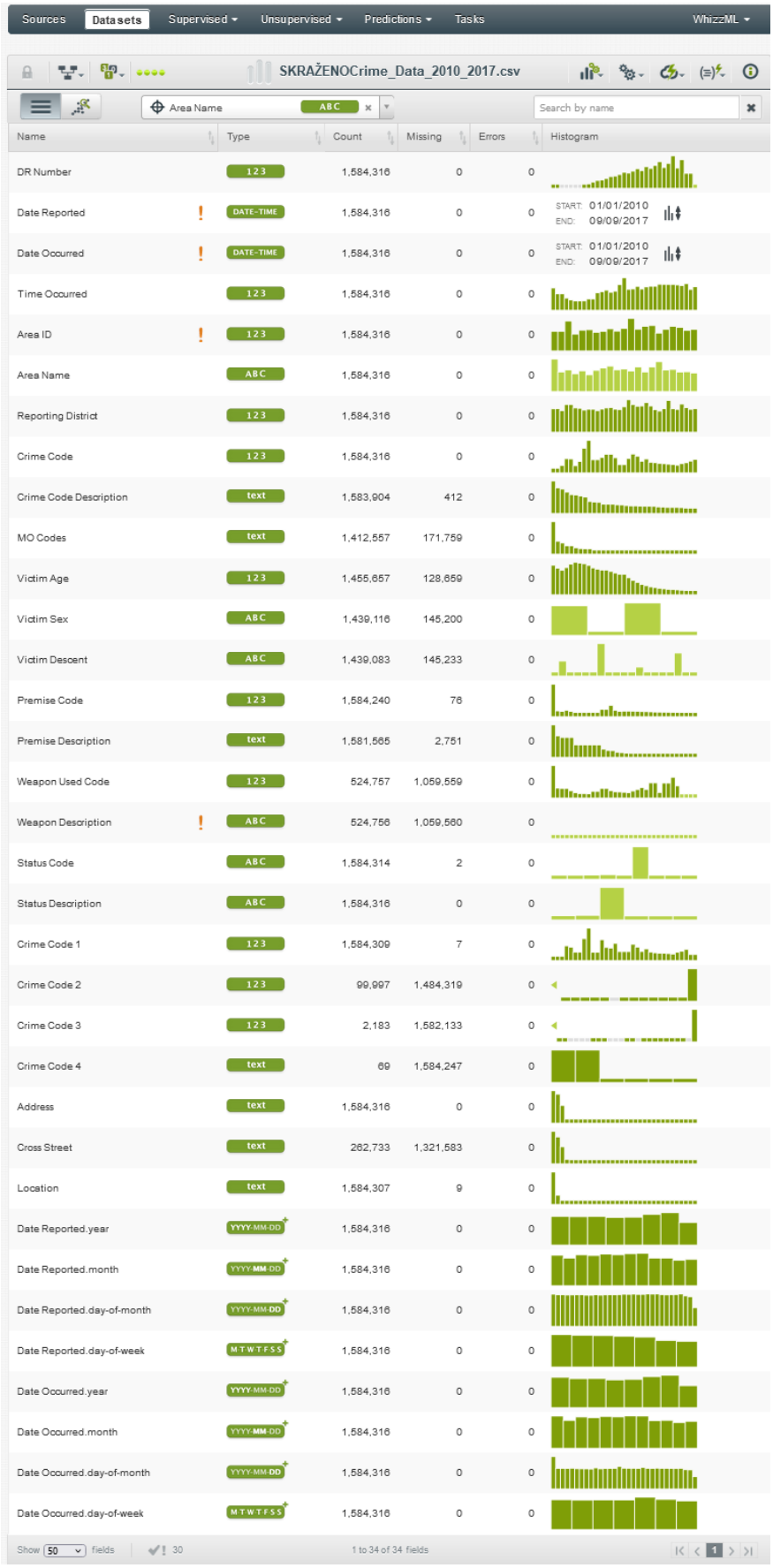
Modeli dobiveni u promatranim istraživanjima variraju u točnosti predviđanja. U istraživanju [5] korištene su metode K-najbliži susjed i Boosted Decision Tree (pojačano stablo odlučivanja) i one su dale modele s točnošću od oko 40%. Za usporedbu, istraživanje [3] koristilo je metodu stabla odlučivanja, a dalo je model s vrlo visokom točnošću od 94%. Kvaliteta tih dvaju modela vjerojatno ovisi o broju instanci koje su korištene u istraživanju, ali i kvaliteti podataka i atributa koji su odabrani. Istraživanje [11] iz tog razloga napominje da ne postoji standardna metoda predviđanja zločina, ali da bi se ubuduće trebalo poraditi na standardizaciji i boljem odabiru metoda za izradu modela. Istraživanje [13] proizvelo je 3 modela. Prvi model nastao algoritmom KNN (K-nearest neighbors) predviđao je s točnošću od oko 32%, drugi model nastao algoritmom SVM (Support Vector Machine) imao je točnost od oko 50% i na kraju posljednji model nastao algoritmom LSTM (Long short-term memory) davao je točnost od oko 58%. Većina autora preporuča isprobavanje drugih metoda izrade modela predviđanja kako bi se našao što točniji.

3.2. Opis i priprema podataka

Izrada ovog modela zasniva se na podacima sakupljenima od 2010. do 2017. na području Los Angelesa. Kako je skup podataka dovoljno velik i proteže se kroz nekoliko godina, moguće je da bi se rezultati modela mogli primijeniti i na drugim područjima SAD-a. Podaci su mogli biti prikupljeni na više načina, u ovom slučaju riječ je o preuzimanju .csv

datoteke sa stranice Kaggle. Podaci na stranici Kaggle su originalno preuzeti sa službenih stranica grada Los Angelesa, a podaci su dobiveni prepisivanjem zapisa iz službenih arhiva policijskih postaja.

Zbog lakše analize, cijeli skup podataka učitan je u online alat BigML. Skup podataka ima ukupno 1,584,316 instanci i 34 atributa od kojih su njih 4 izvedena iz atributa Date Reported i 4 iz atributa Date Occurred (Slika 1).



Slika 1. Skup podataka učitan u alatu BigML (vlastita izrada)

Zbog lakšeg snalaženja, prvo je napravljena tablica sa svim atributima i njihovim opisima (Tablica 1).

Tablica 1. Atributi skupa podataka s opisom (vlastita izrada)

Naziv atributa	Opis atributa
DR Number	Broj službenog zapisa koji se sastoji od dvije znamenke godine, identifikacijskog broja područja u kojemu je počinjen zločin i pet dodatnih znamenki
Date Reported	Datum kada je zločin prijavljen
Date Occurred	Datum kada se zločin zaista dogodio
Time Occurred	Točni sati u kojima je počinjen zločin
Area ID	Identifikacijski broj područja u kojemu je počinjen zločin
Area Name	Naziv područja
Reporting District	Kod od četiri znamenke koji predstavlja pod-područje
Crime Code	Kod zločina
Crime Code Description	Opis koda zločina

MO Codes	Kod postupaka počinjenih za vrijeme istrage zločina
Victim Age	Dob žrtve
Victim Sex	Spol žrtve
Victim Descent	Podrijetlo žrtve
Premise Code	Kod okolice u kojoj je počinjen zločin (vrsta prostora, automobila ili lokacije)
Premise Description	Opis koda okolice
Weapon Used Code	Kod korištenog oružja
Weapon Description	Opis oružja
Status Code	Kod statusa slučaja
Status Description	Opis statusa slučaja
Crime Code 1	Kod koji opisuje počinjeni zločin
Crime Code 2	Kod koji opisuje dodatni počinjeni zločin

Crime Code 3	Kod koji opisuje dodatni počinjeni zločin
Crime Code 4	Kod koji opisuje dodatni počinjeni zločin
Address	Adresa počinjenog zločina
Cross Street	Naziv poprečne ulice
Location	Točna lokacija (koordinate)
Date Reported.year	Godina kada je zločin prijavljen
Date Reported.month	Mjesec kada je zločin prijavljen
Date Reported.day-of-month	Dan u mjesecu kada je zločin prijavljen
Date Reported.day-of-week	Dan u tjednu kada je zločin prijavljen
Date Occurred.year	Godina kada se zločin dogodio
Date Occurred.month	Mjesec kada se zločin dogodio
Date Occurred.day-of-month	Dan u mjesecu kada se zločin dogodio

Date Occurred.day-of-week	Dan u tjednu kada se zločin dogodio

Zbog razumijevanja podataka, atributi su podijeljeni u dva zasebna skupa, kontinuirane (Tablica 2) i kategorijske (Tablica 3) atribute.

Tablica 2. Kontinuirani atributi skupa podataka (vlastita izrada)

Naziv atributa	Broj vrijednosti atributa	Udio nedostajućih vrijednosti (%)	Minimalna vrijednost	Aritmetička sredina	Medijan	Maksimalna vrijednost	Standardna devijacija
DR Number	1,584,316	0	210	135,807,194.91	136,384,923.43	910,220,366.00	22,590,983.75
Time Occurred	1,584,316	0	1.00	1,363.98	1,426.61	2,359.00	646.75
Area ID	1,584,316	0	1.00	11.15	11.51	21.00	5.99
Reporting District	1,584,316	0	100	1,161.59	1,198.13	2,198.00	598.66
Crime code	1,584,316	0	110.00	507.00	451.77	956.00	210.55
Victim Age	1,455,657	8.12	10.00	35.93	33.64	99.00	16.81
Premise Code	1,584,240	0.005	101.00	312.41	211.98	971.00	210.23
Weapon Used Code	524,757	66.88	101.00	370.57	396.77	516.00	114.02
Crime Code 1	1,584,309	0.0004	110.00	506.88	443.23	999.00	210.47
Crime Code 2	99,997	93.69	121.00	954.46	998.00	999.00	123.61
Crime Code 3	2,183	99.86	93.00	970.72	998.00	999.00	88.35

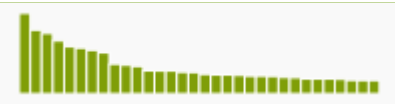
Date Reported.year	1,584,316	0	2,010.00	2,013.47	2,013.51	2,017.00	2.26
Date Reported.month	1,584,316	0	1.00	6.36	6.33	12.00	3.37
Date Reported.day-of-month	1,584,316	0	1.00	15.80	15.85	31.00	8.81
Date Reported.day-of-week	1,584,316	0	1.00	3.85	3.78	7.00	1.97
Date Occurred.year	1,584,316	0	2,010.00	2,013.42	2,013.45	2,017.00	2.26
Date Occurred.month	1,584,316	0	1.00	6.34	6.30	12.00	3.39
Date Occurred.day-of-month	1,584,316	0	1.00	15.50	15.50	31.00	8.93
Date Occurred.day-of-week	1,584,316	0	1.00	4.00	4.04	7.00	1.99

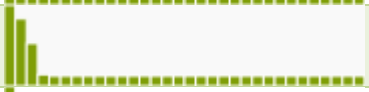
Tablica 3. Kategorijski atributi skupa podataka (vlastita izrada)

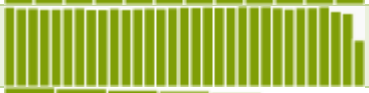
Naziv atributa	Broj vrijednosti atributa	Udio nedostajućih vrijednosti (%)	Mod
Area Name	1,584,316	0	77th Street (110,605)
Crime Code Description	1,583,904	0.026	theft (424,820)
MO Codes	1,412,557	10.84	0344 (526,898)
Victim Sex	1,439,116	9.165	M (739,581)
Victim Descent	1,439,083	9.167	H (549,515)
Premise Description	1,581,565	0.17	dwelling (533,178)
Weapon Description	524,756	66.88	STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE) (319,818)
Status Code	1,584,314	0.00013	IC (1,227,180)
Status Description	1,584,316	0	Invest Cont (1,227,180)
Crime Code 4	69	99.995	998 (56)
Address	1,584,316	0	st (518,734)
Cross Street	262,733	83.42	st (50,967)
Location	1,584,307	0.00057	34 (1,229,507)

Kao završni korak analize podataka grafički je prikazana distribucija podataka svakog atributa.

Tablica 4. Prikaz distribucije podataka određenih atributa (vlastita izrada)

Naziv atributa	Grafički prikaz distribucije	Tip distribucije
DR Number		Normalna pomaknuta ulijevo
Time Occurred		Normalna pomaknuta udesno
Area ID		Uniformna
Area Name		Uniformna
Reporting District		Uniformna
Crime Code		Mulimodalna
Crime Code Description		Normalna pomaknuta ulijevo
MO Codes		Eksponencijalna pomaknuta ulijevo
Victim Age		Normalna pomaknuta ulijevo
Victim Sex		Multimodalna

Victim Descent		Multimodalna
Premise Code		Eksponencijalna
Premise Description		Normalna pomaknuta ulijevo
Weapon Used Code		Multimodalna
Weapon Description		Normalna
Status Code		Normalna pomaknuta udesno
Status Description		Normalna pomaknuta ulijevo
Crime Code 1		Normalna pomaknuta ulijevo
Crime Code 2		Normalna pomaknuta udesno
Crime Code 3		Normalna pomaknuta udesno
Crime Code 4		Normalna pomaknuta ulijevo
Address		Eksponencijalna
Cross Street		Eksponencijalna
Location		Eksponencijalna

Date Reported.year		Uniformna
Date Reported.month		Uniformna
Date Reported.day-of-month		Uniformna
Date Reported.day-of-week		Uniformna
Date Occurred.year		Uniformna
Date Occurred.month		Uniformna
Date Occurred.day-of-month		Uniformna
Date Occurred.day-of-week		Uniformna

Kako je skup atributa u zadanom skupu podataka velik, smanjen je broj promatranih atributa. Atributi koji nas zanimaju su vrijeme i mjesto kada je neki zločin počinjen i eventualno koji je zločin počinjen. Iz tog razloga promatrat će se atributi Date Occurred, Time Occurred, Area Name, Crime Code Description, MO Codes, Premise Description, Crime Code Description i 4 izvedena atributa Date Occurred.year, Date Occurred.month, Date Occurred.day-of-month i Date Occurred.day-of-week. Neki od atributa koje su izbačeni su DR Number, Date Reported, Area ID, Reporting District, Crime Code, Victim Age, Victim Sex, Victim Descent, Premise Code, Weapon Used Code, Weapon Description, Status Code, Status Description, Crime Code 1, Crime Code 2, Crime Code 3, Crime Code 4, Address, Cross Street, Location i 4 izvedena atributa Date Reported.year, Date Reported.month, Date Reported.day-of-month, Date Reported.day-of-week. Atribute DR Number, Reporting District, Address, Cross Street i Location su izbačeni jer sami po sebi sadrže informaciju o mjestu zločina pa bi se samo na temelju jednog od tih atributa moglo odrediti gdje se zločin dogodio. Ako je ijedan od tih atributa poznat, moguće je odrediti u kojem području Los Angelesa se dogodio zločin što bi moglo dovesti do naizgled vrlo pouzdanog modela koji zapravo kao ulaz ima mjesto zločina, a za izlaz također ima mjesto zločina. Atribute Date Reported, Victim Age, Victim Sex, Victim Descent, Weapon Description i Status Description su uklonjeni jer nas oni ne zanimaju i ne vjerujem da oni mogu utjecati na pouzdanost predviđanja modela. Ostali atributi odnose se na kodove koji su pripisani drugim atributima što ih zapravo čini redundantnima ako koristimo atribute koje ti kodovi opisuju.

4. Metode rudarenja podataka

Glavni dio ovog rada je provođenje rudarenja podataka. Metode koje će se koristiti su klaster metoda, metoda stabla odlučivanja i neuronske mreže. Svaka od metoda odabrana je na temelju prethodnih istraživanja prezentiranih ranije. Metoda koja se pokazala kao najefektivnija bila je metoda stabla odlučivanja. Neuronske mreže se koriste pretežito kako bi se grafički prikazalo kolika je šansa da se zločin dogodi na određenom području u određeno vrijeme pomoću alata BigML.

4.1. Klaster metoda

Metoda klaster analize koristi se nad većim skupom podataka koje proučavamo, s obzirom da iste podatke trebamo interpretirati i donijeti zaključak na temelju istih. Ova metoda uzima sva svojstva instanci te ih uredno posloži u skupine kako bi se sve karakteristike jasno prikazale. [1]

Kao rezultat ove metode rudarenja su klasteri koji se dobiju u alatu BigML kad se odabere nenadzirani algoritam učenja te nakon toga klaster. Cilj ove metode jest grupirati slične instance u nekoliko skupina. [2] U nastavku su prikazani brojevi klastera koji su isprobani, koje vrijednosti su dobivene za ratio_ss te kakva je distribucija instanci po klasterima.

Tablica 5. Istraživanje optimalnog broja klastera (vlastita izrada)

Broj klastera	Ratio_ss	Distribucija instanci
6	0,264820	Cluster 0: 18.96% Cluster 1: 22.80% Cluster 2: 0.39% Cluster 3: 19.95% Cluster 4: 20.52% Cluster 5: 17.39%
7	0,253110	Cluster 0: 19.41% Cluster 1: 17.25% Cluster 2: 16.16% Cluster 3: 0.39% Cluster 4: 15.61% Cluster 5: 14.73% Cluster 6: 16.45%
8	0,271340	Cluster 0: 18.40% Cluster 1: 13.58% Cluster 2: 14.80% Cluster 3: 0.39% Cluster 4: 14.44% Cluster 5: 14.29% Cluster 6: 14.22% Cluster 7: 9.88%

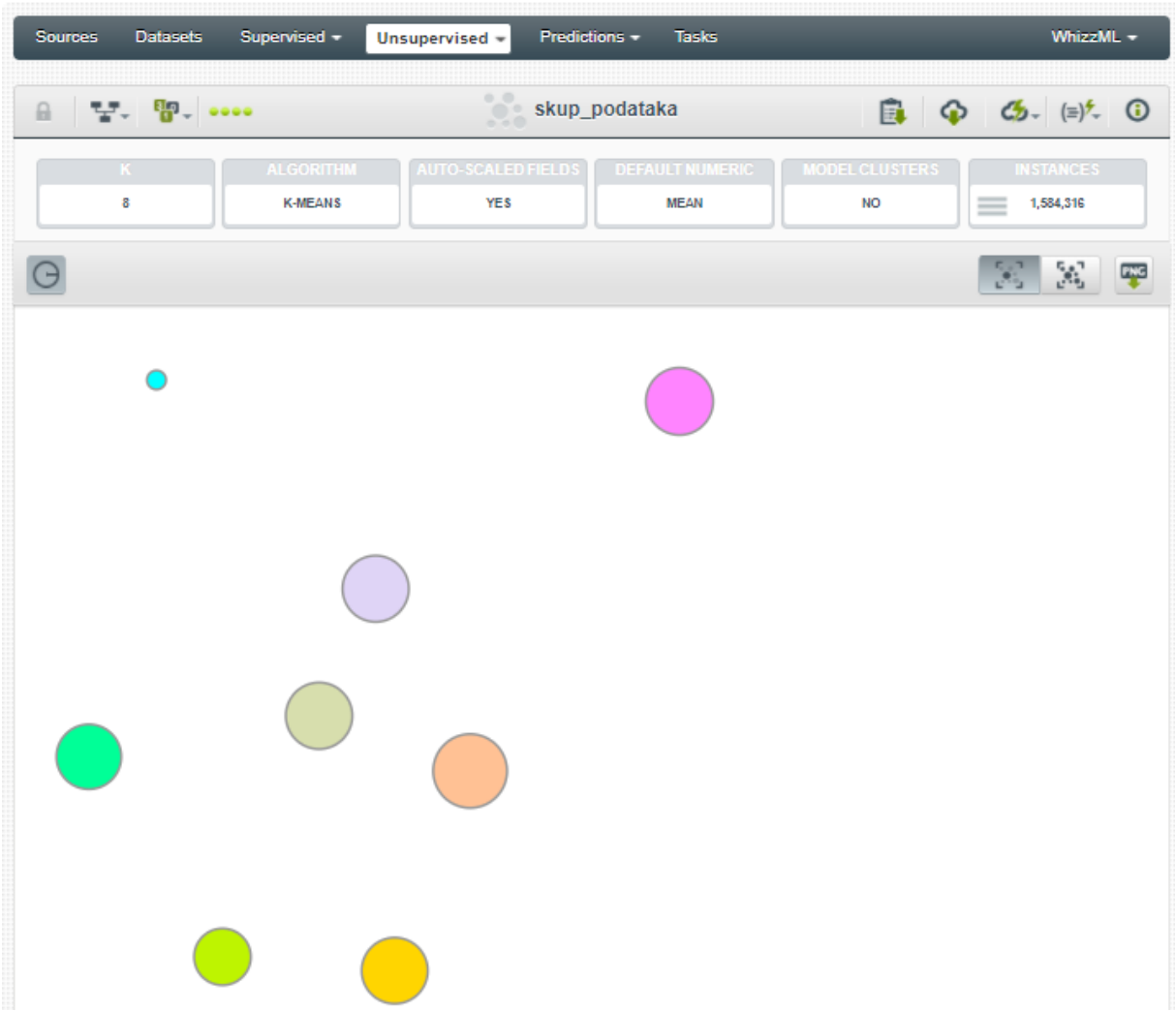
Na temelju dobivenih podataka može se zaključiti da sve distribucije sadrže tzv. stršilo jer distribucija ne smije biti veća od 35% ni manja od 5%. Isto tako na temelju ratio_ss koji mora biti u rasponu od 0 do 1, što je bliže 1 to je bolja distribucija, tako da se najbolje opredijeliti za 8 klastera koji imaju najbolji ratio_ss.

Odabir s 8 klastera:

- Cluster algorithm: K-means
- Number of clusters (K): 8
- Default numeric value: Mean

Broj instanci po klasterima:

- Klaster 0: 291509 instanci
- Klaster 1: 215229 instanci
- Klaster 2: 234452 instanci
- Klaster 3: 6141 instanci
- Klaster 4: 228739 instanci
- Klaster 5: 226411 instanci
- Klaster 6: 225235 instanci
- Klaster 7: 156600 instanci



Slika 2. Vizualizacija klastera (vlastita izrada)

Klasteri se razlikuju te se njihove instance razvrstavaju u različite skupine obzirom na ono što prikazuju i što se u njima nalazi.

- Klaster 0: 291509 instanci, nalazi se na intervalu od 0.17 do 1.65, srednja vrijednost mu je 0.39. Standardna devijacija mu iznosi 0.07.
- Klaster 1: 215229 instanci, nalazi se na intervalu od 0.08 do 1.65, srednja vrijednost iznosi 0.38, dok je standardna devijacija 0.06.
- Klaster 2: 234452 instanci, skup je na intervalu 0.11 – 1.66, srednja vrijednost određenog klastera je 0.36, a standardna devijacija jest 0.07.
- Klaster 4: 228739 instanci, klaster se nalazi u intervalu od 0.14 do 1.66 sa srednjom vrijednošću 0.37 te sa standardnom devijacijom od 0.07.
- Klaster 5: 226411 instanci, nalazi se na intervalu od 0.10 do 1.68. Sadržana srednja vrijednost je 0.38, a standardna devijacija 0.07.
- Klaster 6: 225235 instanci, sadržana je na intervalu 0.17 do 1.68. Srednja vrijednost iznosi 0.40, dok je standardna devijacija 0.07.
- Klaster 7: 156600 instanci, nalazi se na intervalu od 0.17 do 1.68. Srednja vrijednost iznosi 0.37, a standardna devijacija 0.06.

Pojedine klastere interpretiramo na temelju srednjih vrijednosti koje imaju atributi u tom klasteru [2] Klasteri se razlikuju prema tome gdje se i kada dogodio pojedini zločin. Tako su primjerice u klasterima 0 i 2 zločini grupirani oko 2015. godine, dok su u klasterima 4, 5, 6 i 7 grupirane oko 2011. godine. Istance unutar klastera povezane su i prema vrsti zločina. U klasteru 2 najviše je istanci povezanih s provalama i krađama, dok se u klasteru 1 nalaze kaznena djela vezana uz napade i vandalizam. Klasteri se razlikuju i prema vrijednosti atributa „Area Name“. Tako se u klasteru 1 nalaze istance kojima je uglavnom vrijednost ovog atributa „Pacific“, u klasteru 5 „Mission“, u klasteru 6 „Devonshire“ i tako dalje. Za svaki klaster različita je vrijednost ovog atributa.

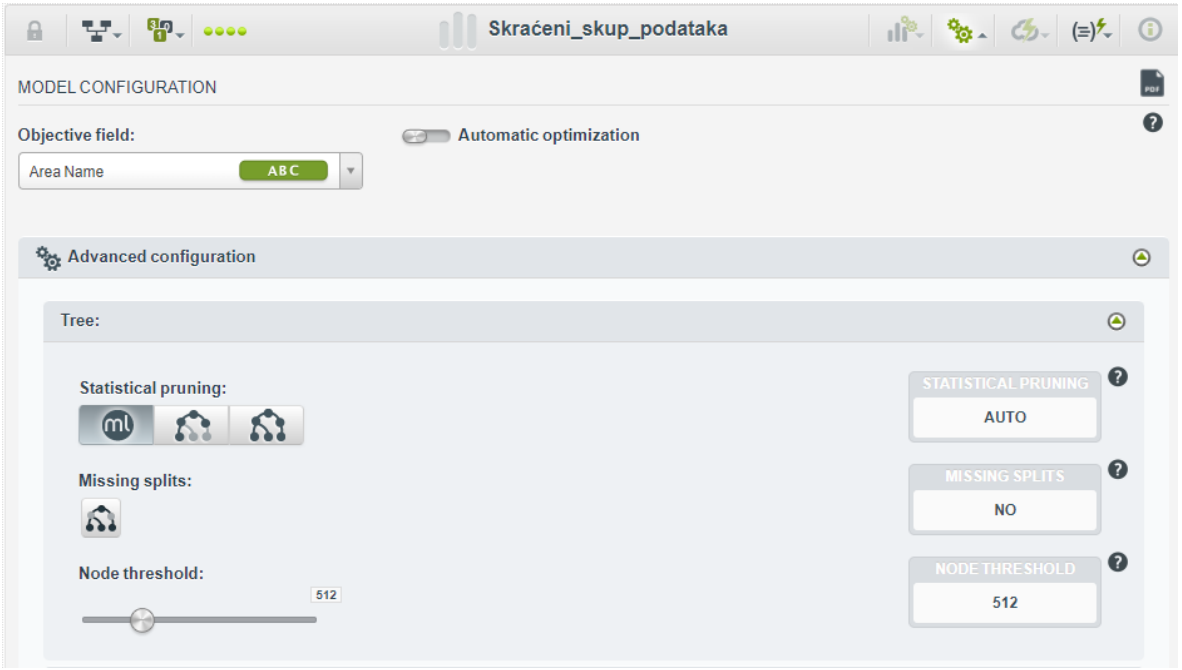
Tablica 6. Prikaz klastera i nekih atributa (vlastita izrada)

	Klaster 0	Klaster 1	Klaster 2	Klaster 4	Klaster 5	Klaster 6	Klaster 7
Time Occurred	1392.60	1212.51	1425.55	1425.65	1548.07	1094.86	1898.07
Area Name	77th Street	Pacific	West Valley	West LA	Mission	Devonshire	Van Nuys
Crime Code 1	590.58	561.12	423.42	506.58	511.41	490.57	382.72
Crime Code 2	955.96	955.12	955.14	955.43	955.38	955.36	960.82
Date Occurred.year	2015.44	2012.65	2015.34	2011.43	2011.46	2011.15	2011.54

4.2. Stablo odlučivanja

U ovom odlomku bit će objašnjen prediktivni model koji smo izradili algoritmom stabla odlučivanja u alatu „BigML“ koristeći već spomenuti reducirani skup podataka o počinjenim zločinima u Los Angelesu. Za zavisnu varijablu, odnosno varijablu za koju će se vršiti predviđanje određena je varijabla „Area Name“, a to je zapravo naziv područja. Iza odabira zavisnog atributa slijedi izgradnja prediktivnog modela koji predviđa u kojem se području zločin dogodio.

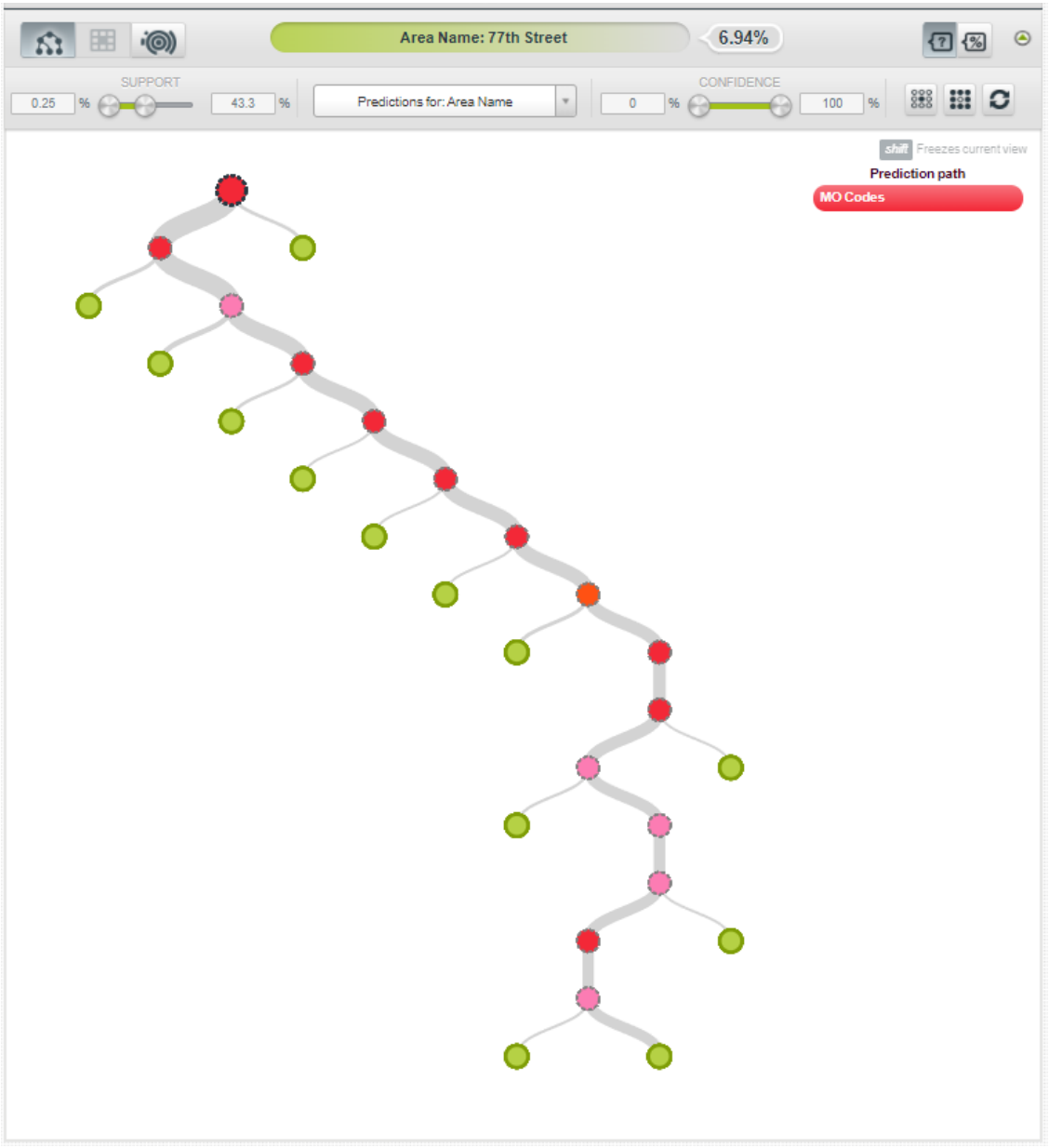
Kod izrade modela stabla odlučivanja može se odabrati jedna od vrsta obrezivanja (engl. pruning). Ti pristupi obrezivanju nam omogućavaju da se optimizira veličina stabla kako bi se zadovoljila glavna prednost stabla odlučivanja. [2] Ta prednost predstavlja to da se modeli stabla odlučivanja mogu lako interpretirati ako nisu previše razgranati i naravno ako su točni i pouzdani. Tri su vrste obrezivanja, a to su nestatističko obrezivanje, pametno obrezivanje i aktivno statističko obrezivanje.



Slika 3. Postavke kod kreiranja modela stabla odlučivanja (vlastita izrada)

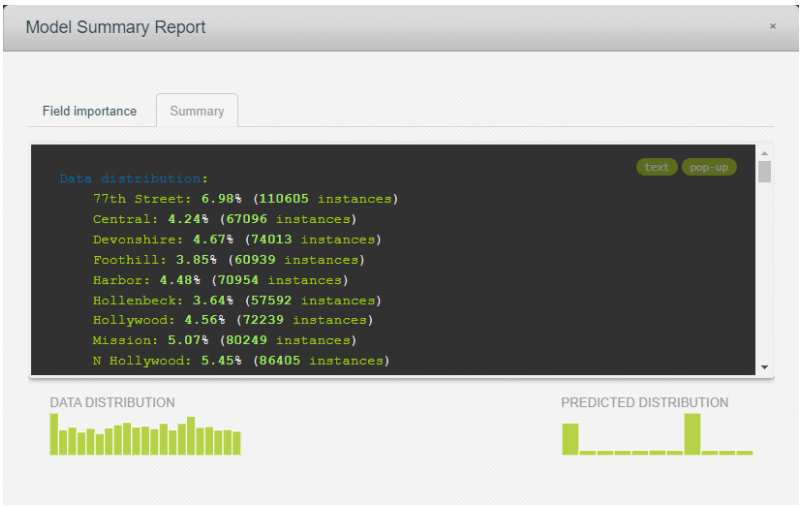
U nastavku slijede modeli za svaku vrstu obrezivanja te će biti objašnjeno koji model daje najbolje rezultate.

4.2.1. Pametno obrezivanje



Slika 4. Prvi model stabla odlučivanja (vlastita izrada)

Na slici 4 prikazana je jedna od grana stabla koje je dobiveno metodom pametnog obrezivanja (eng. *Smart pruning*). Varijable koje su uključene u put (listovi) su MO Codes, Crime Code Description i Premise Description. Zeleni listovi na krajevima predstavljaju pojedina mjesta zločina. Pouzdanost dobivenog modela iznosi 6,94 %, a to je razumljivo zato što se zločin nikako ne može predvidjeti.



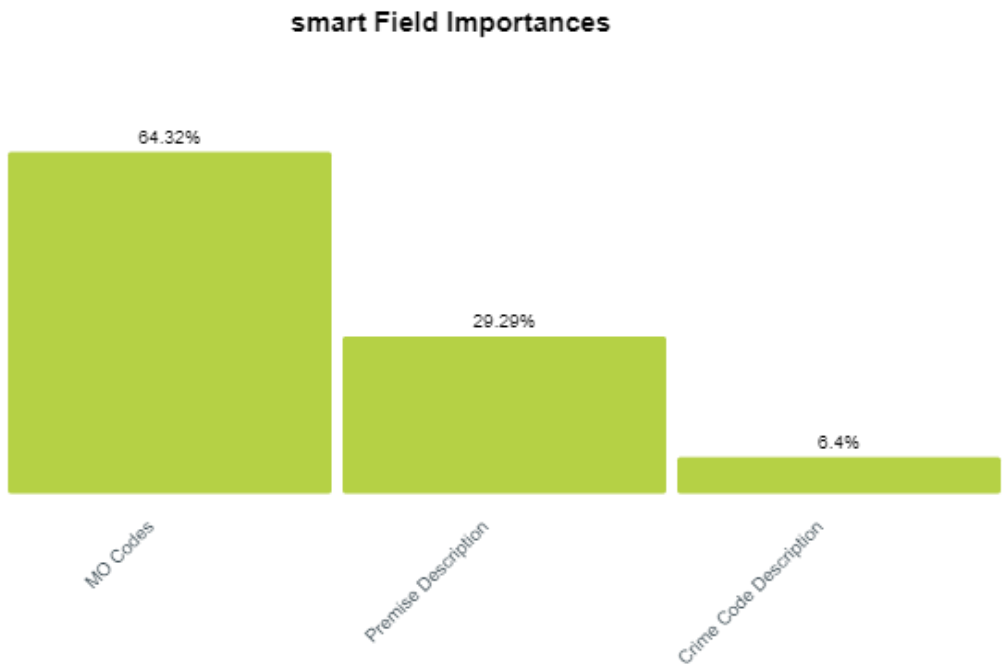
Slika 5. Točnost prvog modela (vlastita izrada)

```
Data distribution:
77th Street: 6.98% (110605 instances)
Central: 4.24% (67096 instances)
Devonshire: 4.67% (74013 instances)
Foothill: 3.85% (60939 instances)
Harbor: 4.48% (70954 instances)
Hollenbeck: 3.64% (57592 instances)
Hollywood: 4.56% (72239 instances)
Mission: 5.07% (80249 instances)
N Hollywood: 5.45% (86405 instances)
Newton: 4.70% (74385 instances)
Northeast: 4.84% (76627 instances)
Olympic: 4.43% (70133 instances)
Pacific: 5.29% (83763 instances)
Rampart: 4.23% (67010 instances)
Southeast: 5.27% (83517 instances)
Southwest: 6.45% (102259 instances)
Topanga: 4.63% (73291 instances)
Van Nuys: 4.76% (75456 instances)
West LA: 4.19% (66388 instances)
West Valley: 4.28% (67797 instances)
Wilshire: 4.01% (63598 instances)

Predicted distribution:
77th Street: 33.42% (529555 instances)
Central: 3.22% (51068 instances)
Harbor: 4.88% (77374 instances)
Hollenbeck: 0.46% (7348 instances)
N Hollywood: 2.27% (35925 instances)
Newton: 5.79% (91700 instances)
Olympic: 3.19% (50484 instances)
Pacific: 43.54% (689879 instances)
Southwest: 1.20% (18999 instances)
Topanga: 1.32% (20859 instances)
Wilshire: 0.70% (11125 instances)
```

Slika 6. Model Summary Report (vlastita izrada)

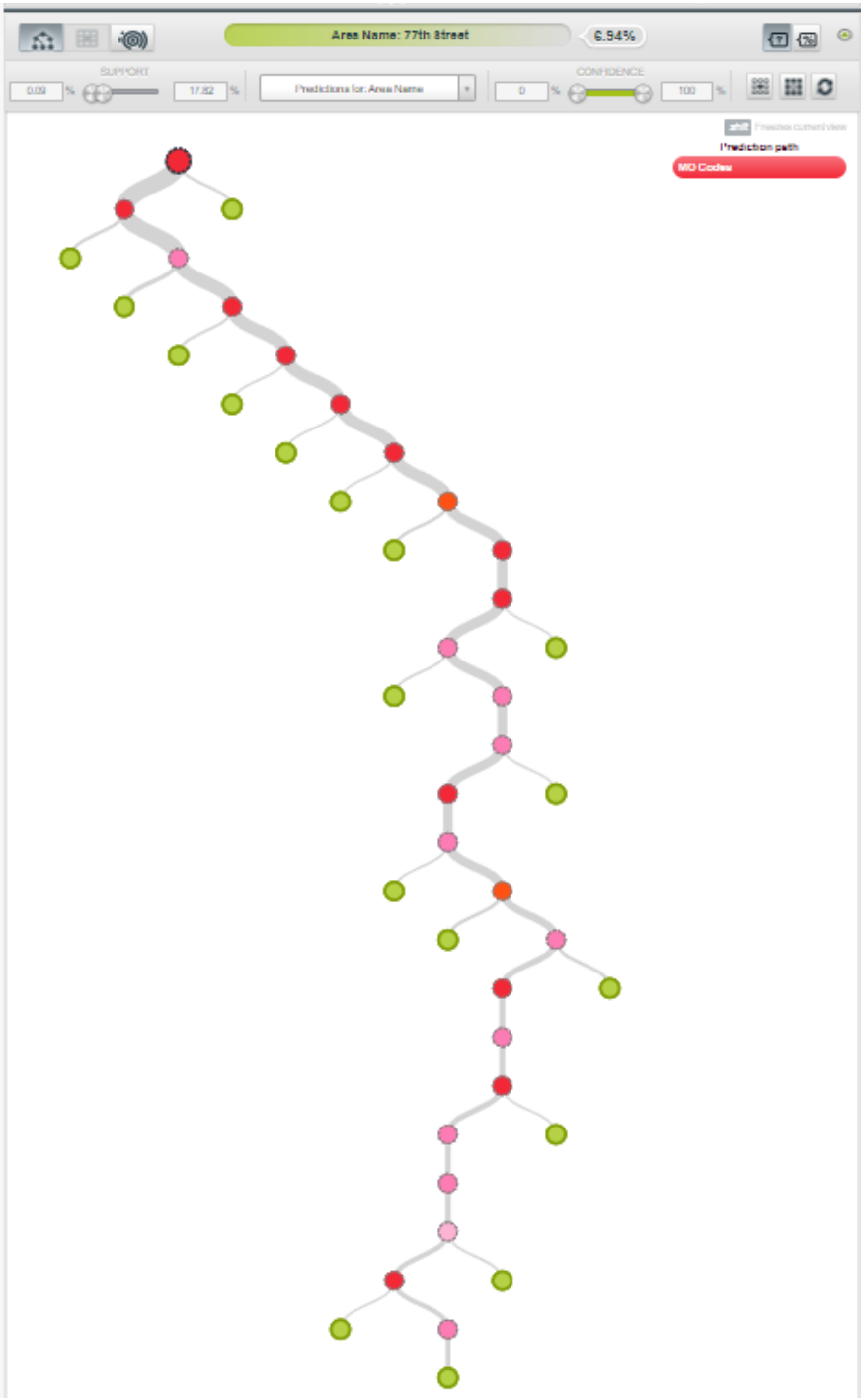
Iz priloženog Model Summary Report-a na prvu se čini da taj model uopće nije točan. To je zato što ako usporedimo prvo područje „77th Street“ vidimo da je „Data distribution“ 6,98 %, dok je kod „Predicted distribution“ za isto to područje 33,42%. Prema tome ispada da je razlika točnosti, odnosno pogreška modela 26,44%. S druge strane npr. za područje „Harbor“ pogreška u točnosti iznosi samo 0,4%. Opet treba napomenuti, ne postoji savršeni model koji može predvidjeti područje u kojem će se i kada dogoditi neki zločin te je pouzdanost i točnost modela takva kakva jest. Važno je da policija otprilike može raspodijeliti vlastite resurse po područjima unutar grada.



Slika 7. Prikaz važnosti atributa koji utječu na predikciju (vlastita izrada)

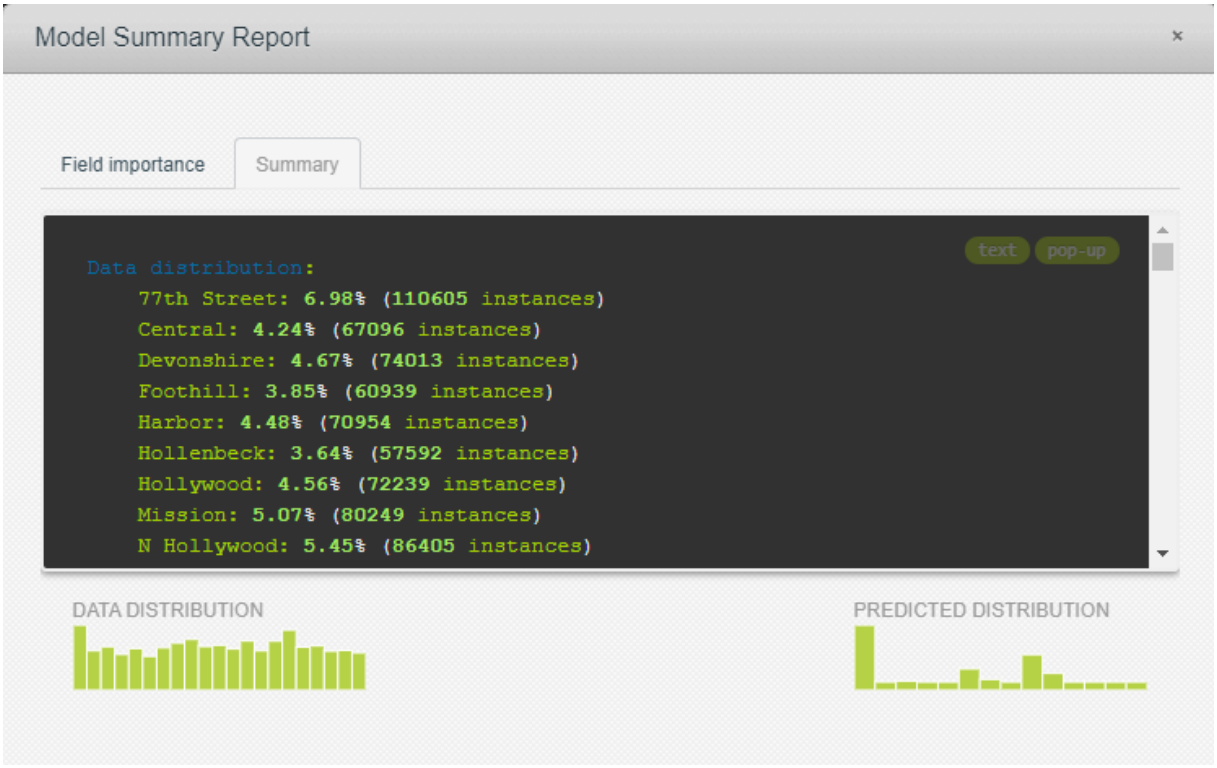
Na slici 7 može se uočiti da su samo tri varijable koje utječu na predikciju ovoga modela. Prva i najvažnija varijabla je „MO Codes“ (kod postupaka počinjenih za vrijeme istrage zločina) i ta varijabla ima 64,32% utjecaja na model. Sljedeća manje važnija varijabla je „Premise Description“ (opis koda okolice) i ona ima 29,29% utjecaja na model. Posljednja i najmanje utjecajna varijabla, je varijabla „Crime Code Description“ (opis koda zločina) i ona utječe samo 6,4% na model.

4.2.2. Aktivno statističko obrezivanje



Slika 8. Drugi model stabla odlučivanja (vlastita izrada)

Na slici 8 prikazana je jedna od grana stabla dobivenog metodom aktivnog statističkog obrezivanja (eng. *Active statistical pruning*). Varijable uključene u predviđanje iste su kao i na prethodnom modelu. Pouzdanost drugog modela, isto kao i kod prvog modela, iznosi 6,94 %.



Slika 9. Točnost drugog modela (vlastita izrada)

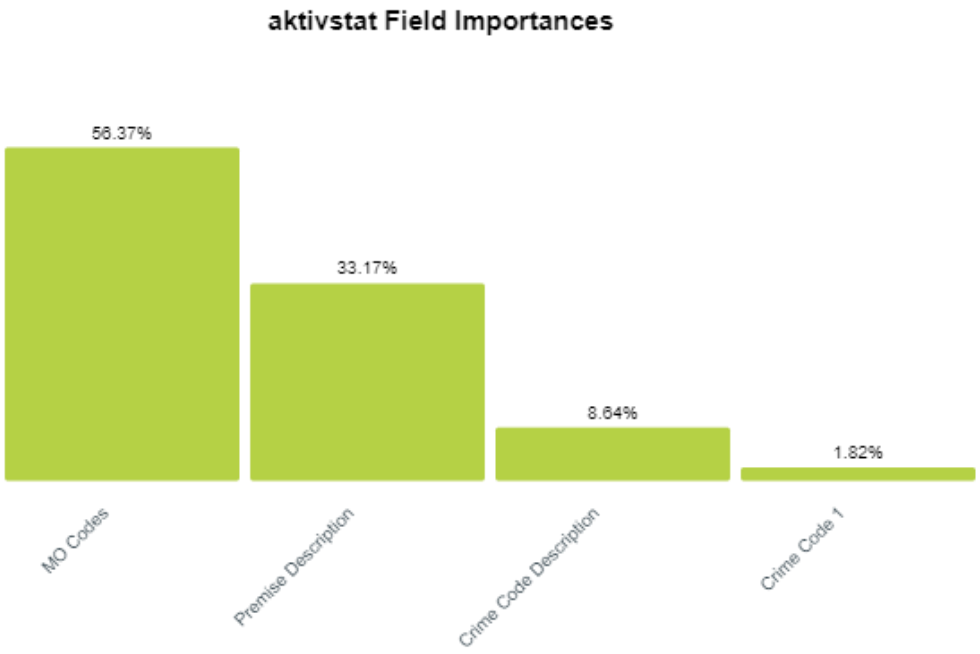
```
Data distribution:
77th Street: 6.55% (110605 instances)
Central: 2.32% (67096 instances)
Devonshire: 4.67% (74013 instances)
Foothill: 3.85% (60939 instances)
Harbor: 4.48% (70954 instances)
Hollenbeck: 3.64% (57592 instances)
Hollywood: 4.56% (72239 instances)
Mission: 5.07% (80249 instances)
N Hollywood: 5.45% (86405 instances)
Newton: 1.70% (74385 instances)
Northeast: 4.84% (76627 instances)
Olympic: 4.43% (70133 instances)
Pacific: 1.29% (83763 instances)
Rampart: 4.23% (67010 instances)
Southeast: 5.27% (83517 instances)
Southwest: 6.45% (102259 instances)
```

```
Topanga: 4.63% (73291 instances)
Van Nuys: 4.76% (75456 instances)
West LA: 4.19% (66388 instances)
West Valley: 4.28% (67797 instances)
Wilshire: 4.01% (63598 instances)


Predicted distribution:
77th Street: 22.62% (548540 instances)
Central: 1.19% (51068 instances)
Harbor: 4.88% (77374 instances)
Hollenbeck: 0.46% (7348 instances)
Hollywood: 0.21% (3364 instances)
N Hollywood: 12.33% (179530 instances)
Newton: 5.79% (91700 instances)
Olympic: 3.81% (60371 instances)
Pacific: 18.94% (300060 instances)
Rampart: 8.98% (142239 instances)
Southeast: 2.81% (44489 instances)
Southwest: 1.29% (20479 instances)
Topanga: 2.94% (46629 instances)
Wilshire: 0.70% (11125 instances)
```

Slika 10. Model Summary Report (vlastita izrada)

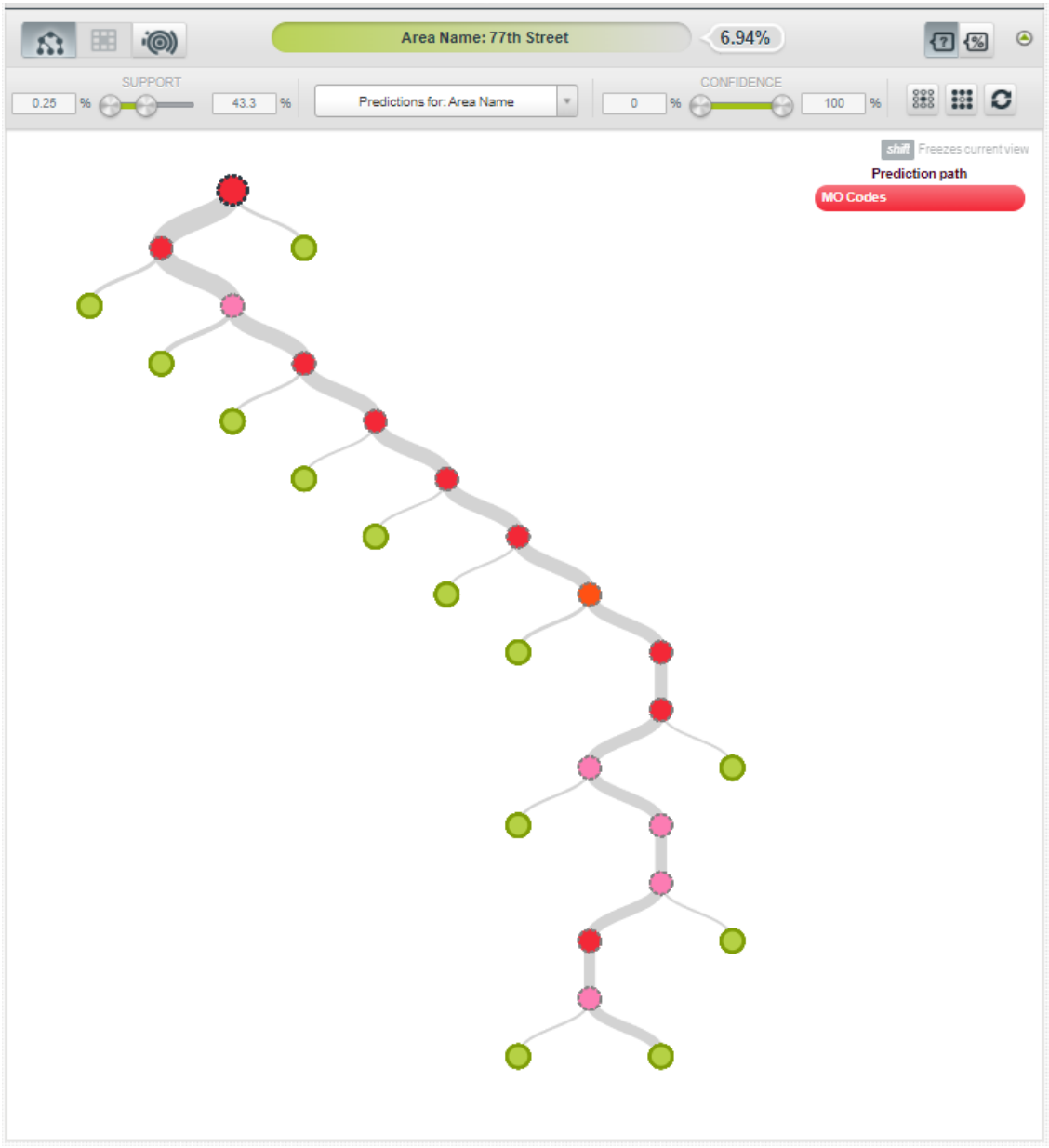
Točnost drugog modela je slična kao i kod prvog modela.



Slika 11. Prikaz važnosti atributa koji utječu na predikciju (vlastita izrada)

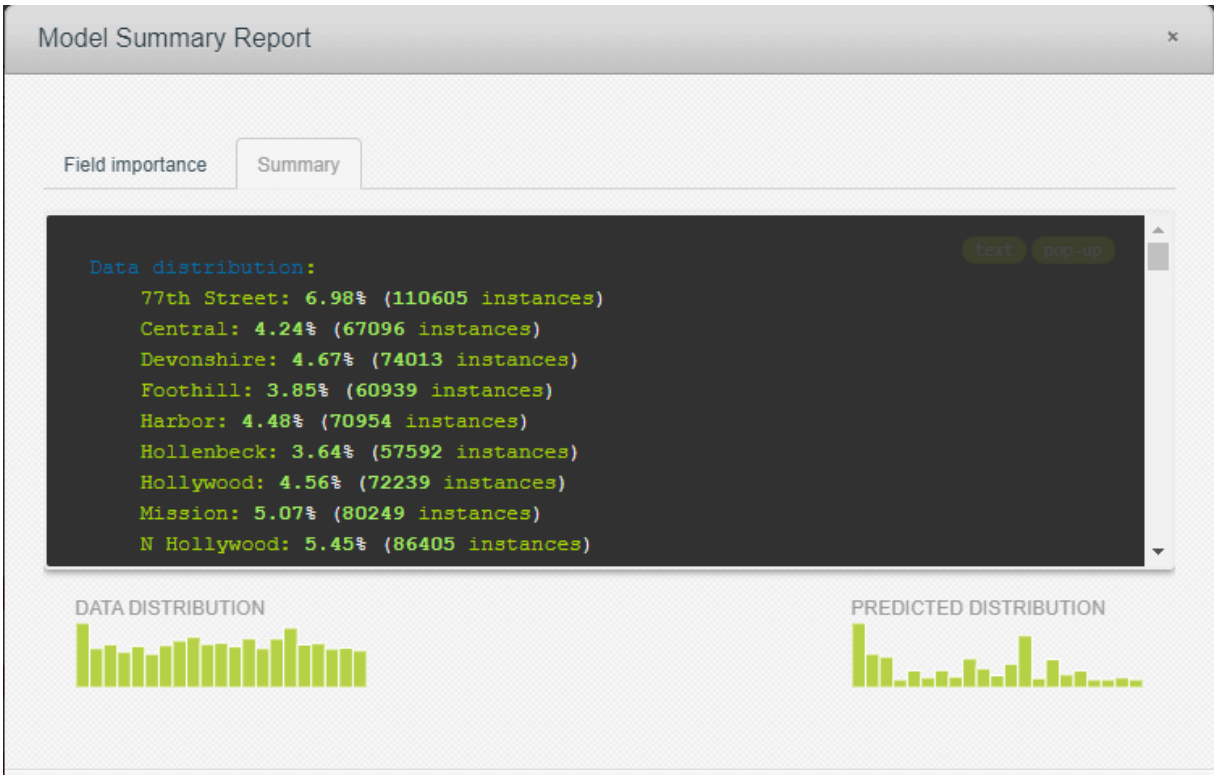
Na slici 11 može se uočiti da u odnosu na prvi model gdje su samo tri varijable koje utječu na predikciju ovoga modela ovdje postoji jedna varijabla više. Prva i najvažnija varijabla je i dalje „MO Codes“ (kod postupaka počinjenih za vrijeme istrage zločina), ali ona sada ima 56,37% utjecaja na model. Sljedeća manje važnija varijabla je „Premise Description“ (opis koda okolice) i ona ima 33,17% utjecaja na model. Pretposljednja varijabla je varijabla „Crime Code Description“ (opis koda zločina) i ona utječe samo 8,64% na model. Naposlijetku, posljednja i najmanje utjecajna varijabla, je varijabla „Crime Code 1“ (kod koji opisuje počinjeni zločin) i ona utječe samo 1,82% na model.

4.2.3. Nestatističko obrezivanje



Slika 12. Treći model stabla odlučivanja (vlastita izrada)

Na slici 12 prikazana je grana stabla koje je dobiveno metodom nestatističkog obrezivanja (eng. *No statistical pruning*). Varijable na ovoj grani su identične kao i na prethodna dva modela. Zeleni listovi također predstavljaju mjesta zločina. Pouzdanost trećeg modela, isto kao i kod prvog i drugog modela, iznosi 6,94 %.



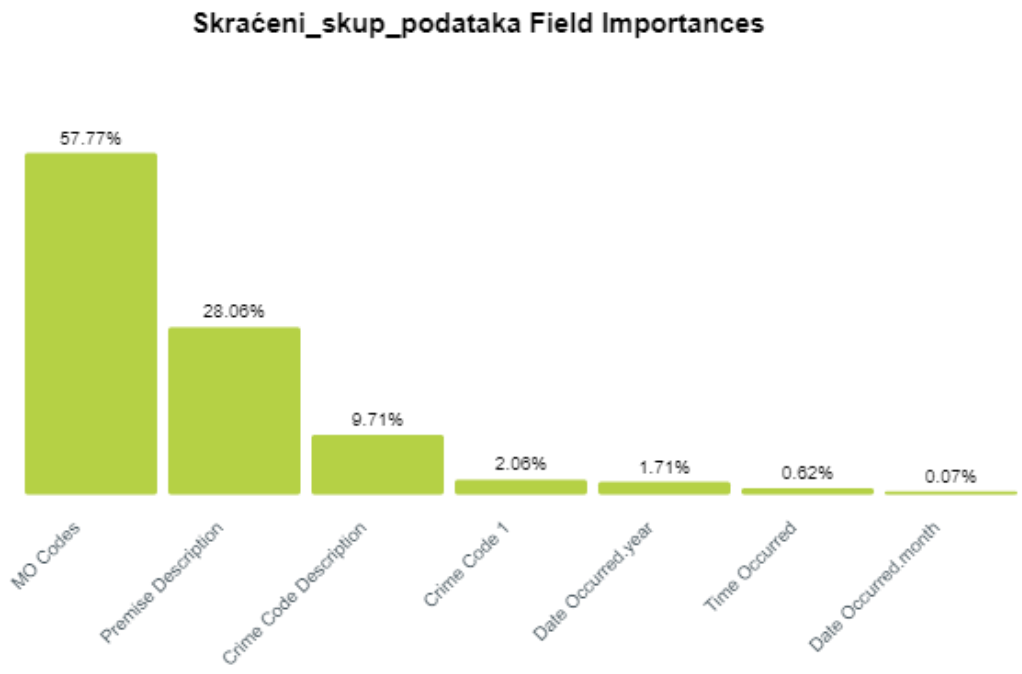
Slika 13. Točnost trećeg modela (vlastita izrada)

```
Data distribution:
77th Street: 6.98% (110605 instances)
Central: 4.24% (67096 instances)
Devonshire: 4.67% (74013 instances)
Foothill: 3.85% (60939 instances)
Harbor: 4.48% (70954 instances)
Hollenbeck: 3.64% (57592 instances)
Hollywood: 4.56% (72239 instances)
Mission: 5.07% (80249 instances)
N Hollywood: 5.45% (86405 instances)
Newton: 4.70% (74385 instances)
Northeast: 4.84% (76627 instances)
Olympic: 4.43% (70133 instances)
Pacific: 5.29% (83763 instances)
Rampart: 4.23% (67010 instances)
Southeast: 5.27% (83517 instances)
Southwest: 6.45% (102259 instances)
Topanga: 4.63% (73291 instances)
Van Nuys: 4.76% (75456 instances)
West LA: 4.19% (66388 instances)
West Valley: 4.28% (67797 instances)
Wilshire: 4.01% (63598 instances)
```

```
Predicted distribution:
77th Street: 15.75% (249491 instances)
Central: 8.23% (130424 instances)
Devonshire: 7.49% (118618 instances)
Foothill: 0.65% (10257 instances)
Harbor: 4.15% (65726 instances)
Hollenbeck: 2.43% (38510 instances)
Hollywood: 4.24% (67169 instances)
Mission: 2.66% (42204 instances)
N Hollywood: 7.08% (112207 instances)
Newton: 4.67% (73951 instances)
Northeast: 2.95% (46810 instances)
Olympic: 5.80% (91827 instances)
Pacific: 12.67% (200775 instances)
Rampart: 2.30% (36509 instances)
Southeast: 6.89% (109144 instances)
Southwest: 3.19% (50535 instances)
Topanga: 4.10% (64965 instances)
Van Nuys: 1.20% (18973 instances)
West LA: 0.39% (6111 instances)
West Valley: 2.44% (38704 instances)
Wilshire: 0.72% (11406 instances)
```

Slika 14. Model Summary Report (vlastita izrada)

Točnost trećeg modela je slična kao i kod prvog i drugog modela, a ona varira u ovisnosti područja za koje promatramo.



Slika 15. Prikaz važnosti atributa koji utječu na predikciju (vlastita izrada)

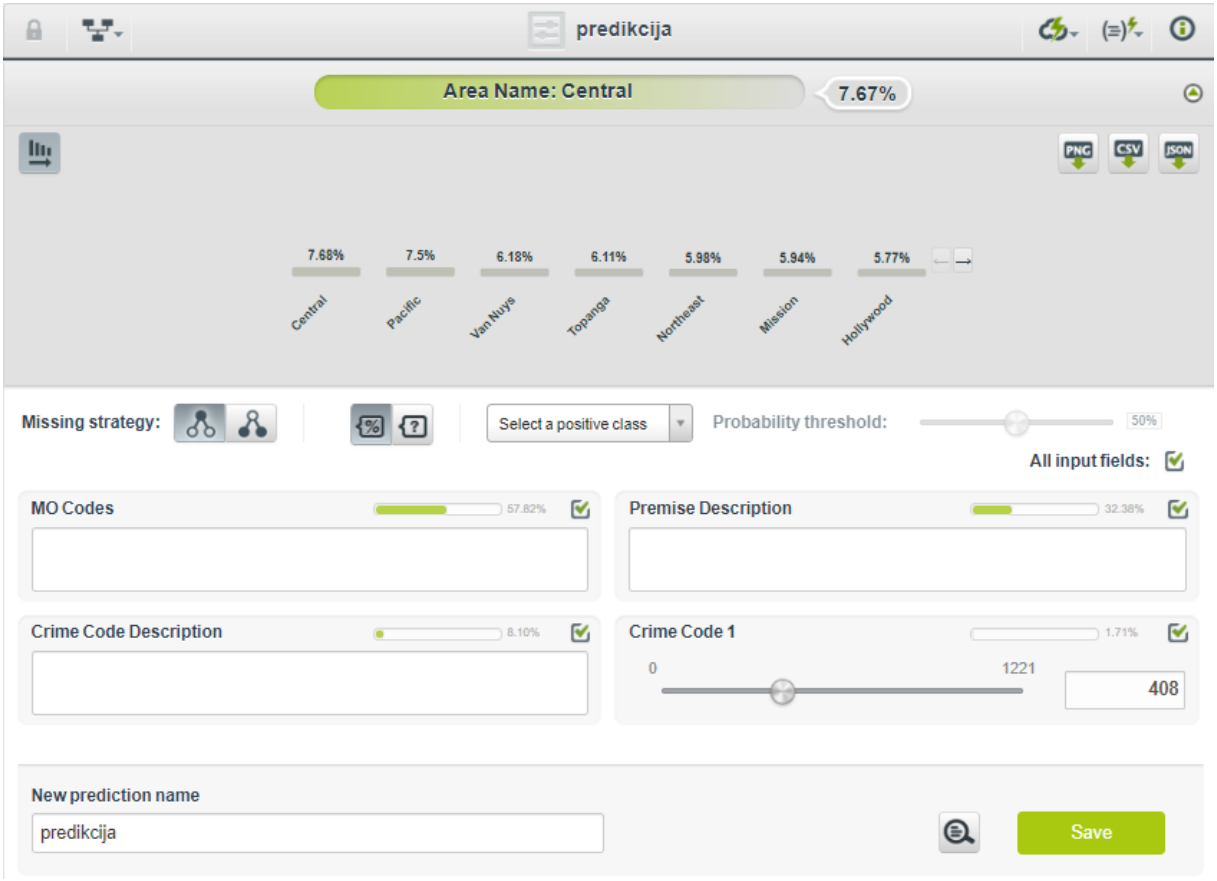
Na slici 15 može se uočiti da u odnosu na prvi i drugi model ima najviše varijabla koje su važne u predikciji. Prva i najvažnija varijabla je i dalje „MO Codes“ (kod postupaka počinjenih za vrijeme istrage zločina), ona ima 57,77% utjecaja na model. Sljedeće manje važnije varijable su „Premise Description“ (opis koda okolice) sa 28,08%, zatim varijabla „Crime Code Description“ (opis koda zločina) sa 9,71%, zatim „Crime Code 1“ (kod koji opisuje počinjeni zločin) sa 2,06% utjecaja na model. Naposlijetku slijede najmanje važne varijable, a to su „Date Occured.year“ (prikazuje godinu kada se zločin dogodio) sa 1,71%, zatim „Time Occured“ (točni sati u kojima je počinjen zločin) sa 0,62% te na kraju varijabla „Date Occured.month“ (prikazuje mjesec kada se zločin dogodio) sa 0,07% utjecaja na model.

4.2.4. Predikcija pomoću stabla odlučivanja

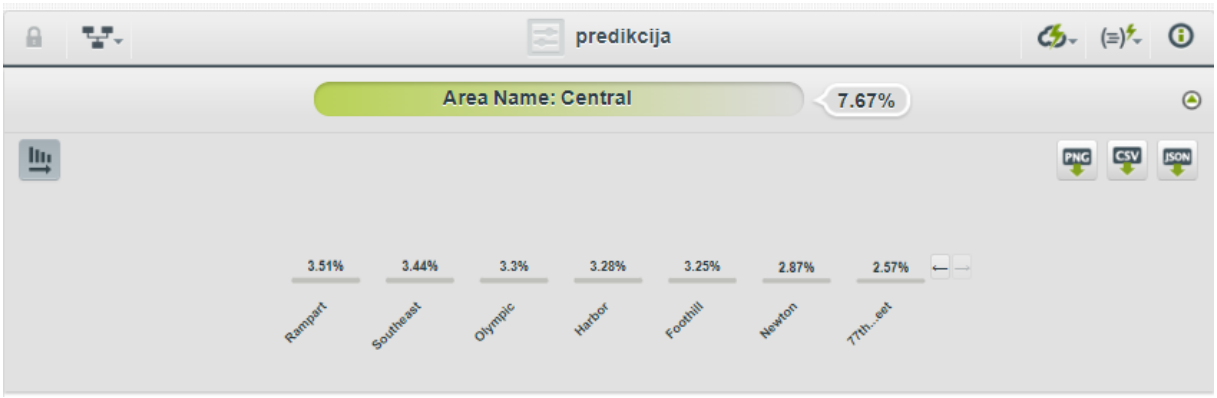
Može se uočiti, kod sva tri modela stabla odlučivanja u ovisnosti o vrsti obrezivanja, da je pouzdanost bila identična, a točnost približno slična. Iz tog razloga ne postoji idealan model od ta tri nad kojim bi se napravilo predviđanje pa je svejedno koji model odabrati. Budući da se posljednje u „BigML“-u kreirao model stabla odlučivanja koristeći ne statističku vrstu obrezivanja možemo nad tim modelom kreirati i predviđanje.

Na temelju odabranog modela stabla odlučivanja kojeg smo izradili, pritiskom na gumb „Predict“ kreira se predviđanje. Ispod u postavkama postave se vrijednosti atributa za koje želimo napraviti predviđanje. Pritiskom na gumb „Save“ sprema se model. Na vrhu slika 16 i 17 prikazane su kakve su predviđene vrijednosti izlaza. Na slici 16 može se vidjeti da je najveći postotak predviđanja da će se zločin dogoditi u području „Central“, dok sa slike 17. možemo vidjeti da je najmanji postotak da će se zločin dogoditi u području „77th Street“.

Iako su ti postoci predviđanja zločina po područjima grada Los Angelesa mali, oni svakako mogu pomoći tamošnjoj policiji da mogu bolje rasporediti policijske patrole po gradu te na taj način povećati efikasnost i smanjiti učinak nedostatka radnika u policiji. Iz navedenog razloga, možemo reći da je prediktivni model ipak u konačnici dobar zato što unatoč lošijoj pouzdanosti i točnosti ipak ima smisla i bio bi od neke koristi policiji.



Slika 16. Prikaz rezultata predikcije područja gdje će se prije dogoditi zločin (vlastita izrada)

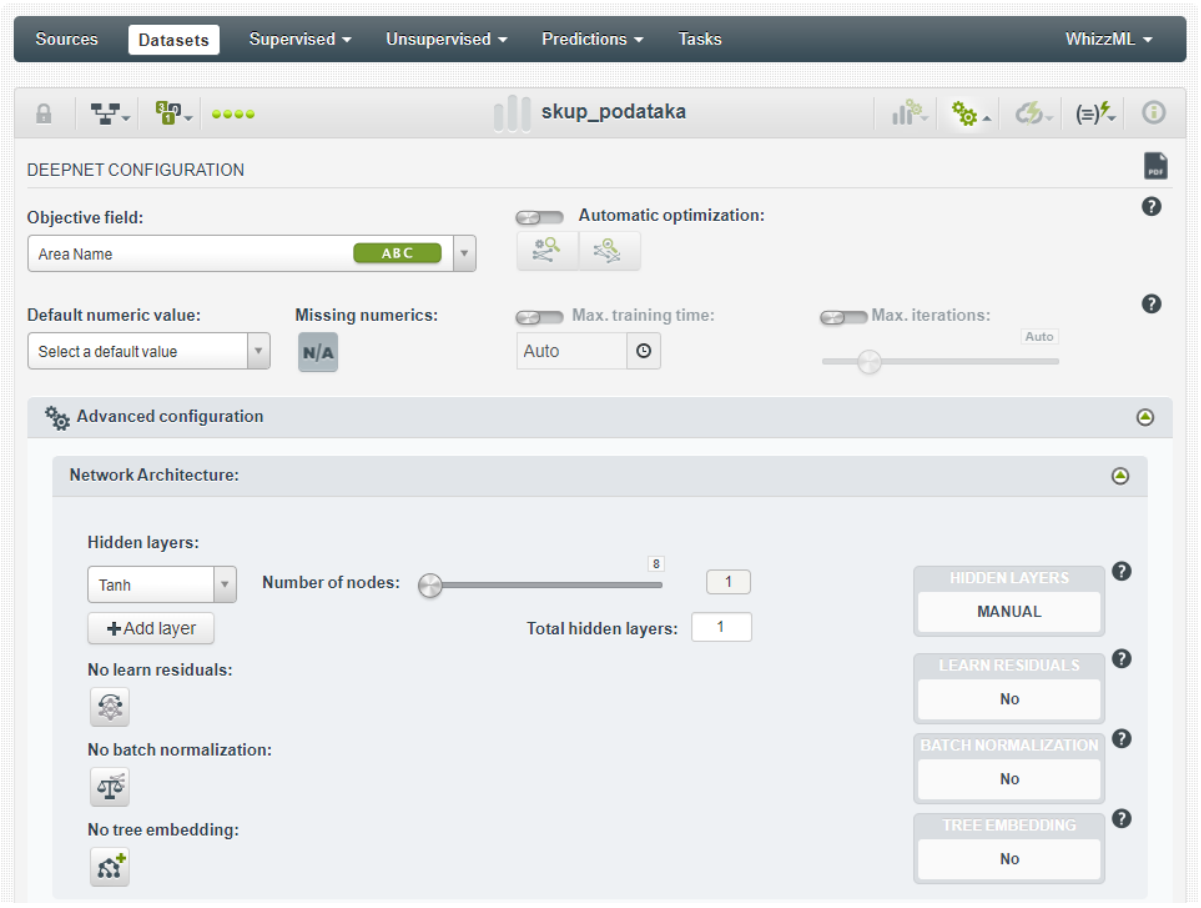


Slika 17. Prikaz rezultata predikcije područja gdje će se prije dogoditi zločin (vlastita izrada)

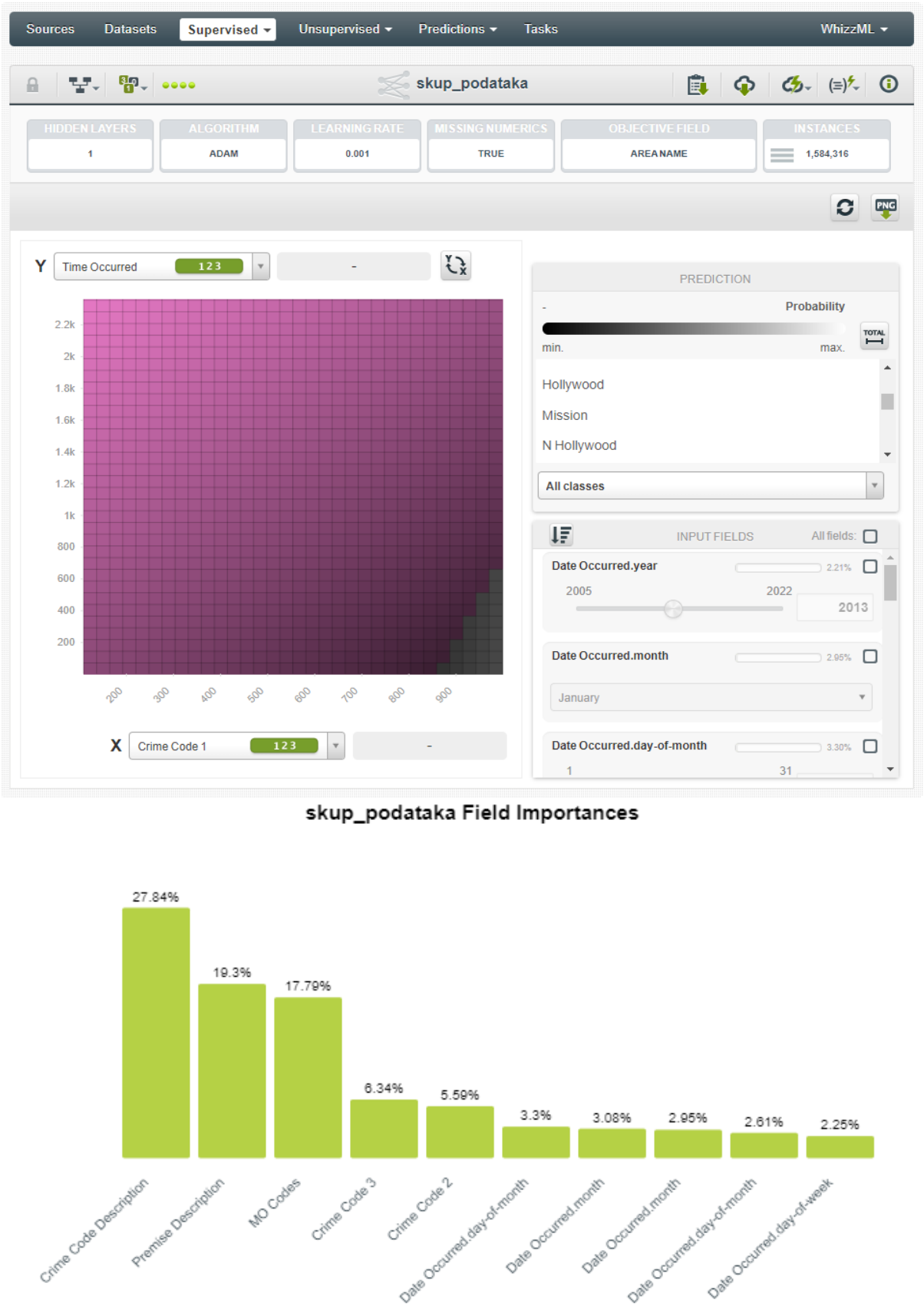
4.3. Neuronske mreže

Ovo poglavlje donosi objašnjenje predviđanja putem algoritama neuronske mreže u alatu BigML (opcija Deepnet). I dalje se koristi isti skup podataka. Predviđat će se područje u kojem se zločin dogodio s obzirom na vrijeme i vrstu zločina.

U odabranom skupu podataka potrebno je kliknuti na opciju „Configure“ te dalje na „Deepnet“, odnosno neuronska mreža. Sljedeći korak je konfigurirati postavke za analizu. Ne želimo postavke ostaviti na „Automatski“. Potrebno je odabrati zavisni atribut, onaj kojeg predviđamo. U našem slučaju to je „Area name“. Isključuje se automatska optimizacija, kako bi se mogao varirati broj neurona. Zadana vrijednost je 64, no u našem slučaju koristit će se vrijednosti 8, 9 i 10. Tri su sloja neurona. Prvi je sloj ulaznih neurona, tj. jedan ulazni neuron predstavlja jednu ulaznu varijablu. Drugi je sloj u kojem se nalaze neuroni koji obrađuju podatke. Taj sloj možemo varirati. Posljednji sloj predstavlja sloj u kojem se nalazi ono što predviđamo. [2] Broj neurona određuje se kao aritmetička sredina neurona na ulazu i izlazu, te je traženi broj neurona 9. Posljednji korak je varirati broj neurona za ± 1. Navedeno se vidi na sljedećim snimkama zaslona.



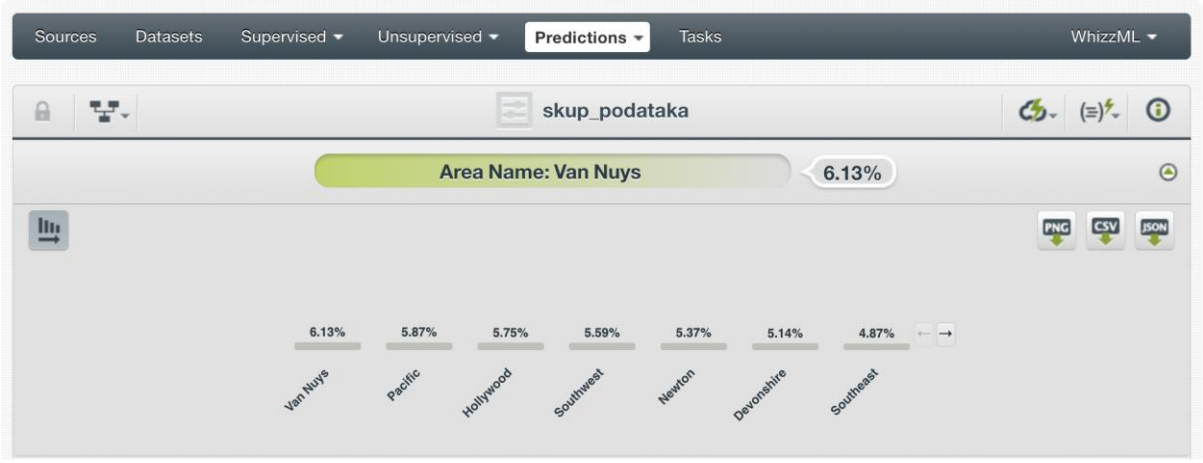
Slika 18. Deepnet konfiguracija (vlastita izrada)



Slika 19. 9 neurona (vlastita izrada)

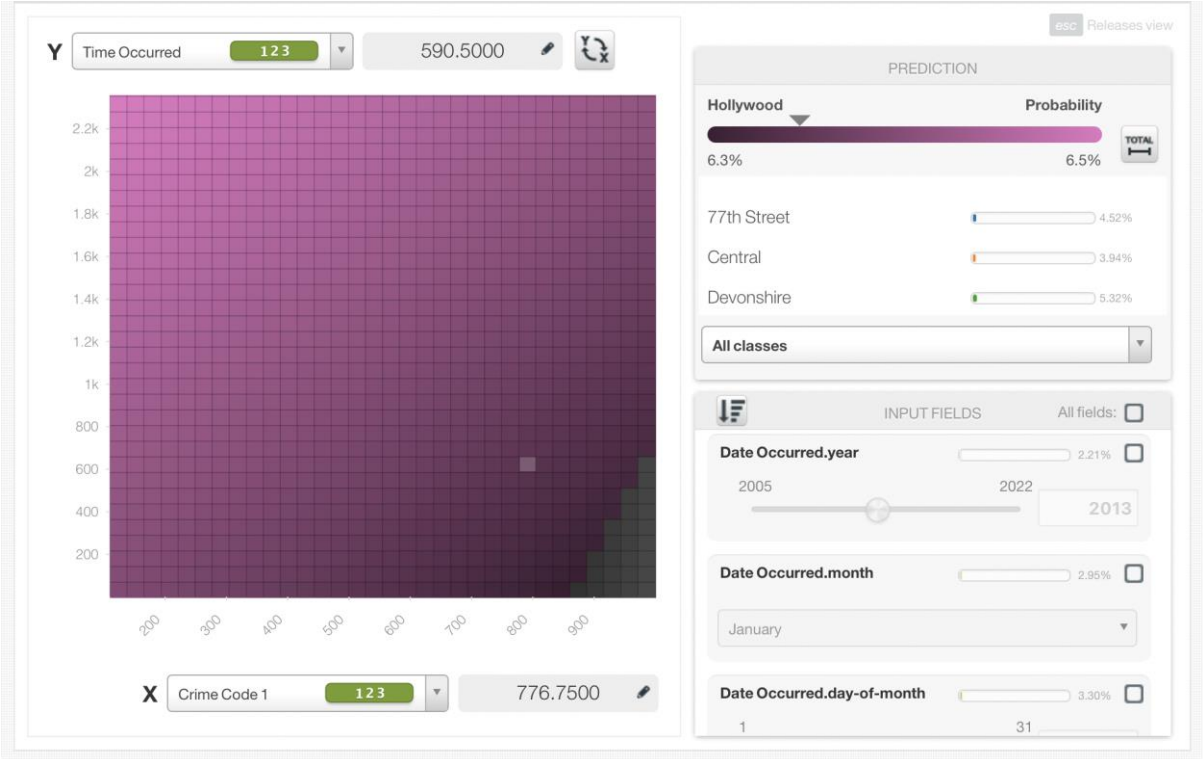
Broj neurona variran je te su izrađeni modeli sa 8, 9 i 10 neurona. Nakon variranja broja sakrivenih neurona koristit će se model sa 9 neurona s obzirom na raspodjelu varijabli, što se može vidjeti u „Deepnet Summary Report“. Modeli su pokazivali vrlo sličnu razinu pouzdanosti.

U nastavku će se pokazati primjer predviđanja primjenom opcije „Predict“. Ovdje se mogu varirati svi parametri te vidjeti kako utječu na krajnji ishod. Parametar godine stavlja se na 2010, dan u tjednu ponedjeljak, 15. u mjesecu. Dobiva se sljedeće predviđanje:



Slika 20. Predviđanje opcijom "Predict"

Predvidjeti možemo i sa početnog zaslona, grafički. Ondje se mogu varirati samo 2 varijable, na x-osi i y-osi. Tako primjerice se možemo pozicionirati u jedan od kvadratića i dobiti slijedeću predikciju:



Slika 21. Predikcija sa samo 2 varijable

S obzirom na veliki broj mogućih lokacija, ovaj model nije trivijalan za predviđanje, već se za svaku lokaciju dobiva predviđanje od najviše 5-6%.

5. Interpretacija i evaluacija modela

Cilj modeliranja našeg skupa podataka bio je odrediti gdje je najveća mogućnost da se određeni zločin dogodi u određeno vrijeme. Dakle, tražili smo povezanost određenog zločina s mjestom i vremenom zločina.

Prediktivnim modeliranjem, tj. modeliranjem algoritmima stabla odlučivanja i neuronskih mreža, došli smo do zaključka kako na sam ishod predviđanja najviše utječu varijable „MO Codes“ (kod postupaka počinjenih za vrijeme istrage zločina) za stablo odlučivanja, odnosno „Crime Code Description“ (opis koda zločina) za neuronske mreže. Drugi najvažniji faktor je „Premise Description“ (opis koda okolice). Korištenjem obaju prediktivnih algoritama dobili su se slični rezultati. Pokušali smo predvidjeti na kojem mjestu će se dogoditi pojedini zločin, no u niti jednom slučaju se za niti jedno područje nije dobio postotak vjerojatnosti veći od 10%.

Klasterskom analizom istance se grupiraju u klastere prema sličnim vrijednostima atributa. Analiza je ponovljena više puta prilikom čega je variran broj klastera. Odabrano je grupiranje u osam klastera jer je takvo grupiranje dalo najbolji ratio_ss. Klasteri se razlikuju prema tome gdje se i kada dogodio pojedini zločin. Tako su primjerice u klasterima 0 i 2 zločini grupirani oko 2015. godine, u klasterima 6 i 7 oko 2011. godine itd. Istance unutar klastera povezane su i prema vrsti zločina. U klasteru 2 najviše je istanci povezanih s provalama i krađama, dok se u klasteru 1 nalaze kaznena djela vezana uz napade i vandalizam. Svi klasteri osim klastera 3 imaju između 150.000 i 290.000 istanci, dok se u klasteru 3 nalazi samo oko 6000 istanci.

Zaključujemo da ovim metodama nije moguće uspostaviti jasniju povezanost između vrste zločina i mjesta i vremena na kojem se zločin dogodio, barem ne na ovom skupu podataka, tj. za niti jedan zločin u određeno vrijeme ne možemo sa sigurnošću reću na kojem mjestu će se dogoditi – za svaki dio grada vjerojatnost je otprilike jednaka.

6. Zaključak

U ovom radu koristilo se više funkcionalnosti alata “BigML“ kako bi se rudarili podatci koji nas zanimaju, odnosno na temelju kojih se bazira cijeli ovaj rad. Za predviđanje su korištene tri metode analize podataka, a to su klaster metoda, stablo odlučivanja i neuronske mreže.

U analizi podataka pomoću klastera u svakoj grupi bilo je prisutno stršilo koje smo ignorirali jer sadrži manje od 0,5% instanci te smo na temelju ratio_ss utvrdili koji je optimalan broj klastera .Ispostavilo se da je u našem slučaju 8, te smo uz pomoć njega analizirali podatke. Korištenjem algoritma stabla odlučivanja za izradu prediktivnog modela, odnosno predviđanja područja zločina napravljena su tri modela u ovisnosti o vrsti obrezivanja. Moglo se uočiti da su točnost i pouzdanost kod sva tri modela bili identični. Pouzdanost je u sva tri modela bila 6,94%, dok je točnost varirala ovisno za koje područje se promatralo. Kod predviđanja koje je napravljeno pomoću trećeg modela stabla odlučivanja, može se vidjeti da su rezultati, odnosno postoci koji govore koliki je postotak gdje će se sljedeći put dogoditi neki zločin jako mali. Najveći postotak je bio za područje "Central" dok je najmanji bio za područje "77th Street". Korištenjem algoritma neuronskih mreža predviđanja su se mogla ostvariti grafički i ručnom konfiguracijom. U oba slučaja bilo je teško predviđati s obzirom na veliki broj vrijednosti izlaznih varijabli. Za svaku lokaciju dobilo se predviđanje od najviše 6%.

Moramo biti svjesni da u realnosti ne postoji savršeni model koji može predvidjeti područje u kojem će se i kada dogoditi neki zločin, ali može se reći da je dobiveni prediktivni model ipak dobar jer može policiji pomoći da otprilike raspodijeli vlastite resurse po područjima unutar grada.

7. Popis literature

1. A. Kumar, A. Verma, G. Shinde i ostali, "Crime Prediction Using K-Nearest Neighboring Algorithm", 2020 International Conference on Emerging Trends in Information Technology and Engineering, 2020., adresa: https://ieeexplore.ieee.org/document/9077832 (preuzeto: 7.12.2021)

2. B. Kliček, D. Oreški, materijal s predavanja iz kolegija Otkrivanje znanja u podacima ak. god. 2021./2022., preuzeto s e-learning sustava Moodle (preuzeto: 11.8.2022.)

3. E. Ahishakiye, E. O. Omulo, D. Taremwa, I. Niyonzima, "Crime Prediction Using Decision Tree (J48) Classification Algorithm", International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Broj 06 – Izdanje 03, 2017., adresa: https://www.ijcit.com/archives/volume6/issue3/Paper060308.pdf (preuzeto: 4.12.2021)

4. J. Gramlich, "What we know about the increase in U.S. murders in 2020“, Pew Research Centre, 2021, adresa: https://www.pewresearch.org/fact-tank/2021/10/27/what-we-know-about-the-increase-in-u-s-murders-in-2020/ (preuzeto 5.1.2021)

5. N. H. M. Shamsuddin, N. A. Ali, R. Alwee, "An Overview on Crime Prediction Methods", Universiti Teknologi Malaysia,Faculty of Computing, 2017., adresa: https://ieeexplore.ieee.org/document/8075335 (preuzeto: 4.12.2021)

6. N. Hotz, "What is CRISP DM?“, Data Science Process Alliance, 2022, adresa: https://www.datascience-pm.com/crisp-dm-2/ (preuzeto: 11.8.2022)

7. N. Hotz, "What is SEMMA?“, 2021, adresa: https://www.datascience-pm.com/semma/ (preuzeto 11.8.2022)

8. Police Executive Research Forum, "The Workforce Crisis, and What Police Agencies Are Doing About It“, Washington, D.C., 2019, adresa: https://www.policeforum.org/assets/WorkforceCrisis.pdf (preuzeto 5.1.2021)

9. R. Iqbal, P. H. S. Panahy, "An Experimental Study of Classification Algorithms for Crime Prediction", Indian Journal of Science and Technology, 2013., adresa: https://www.researchgate.net/publication/256198632_An_Experimental_Study_of_Classification_Algorithms_for_Crime_Prediction (preuzeto: 4.12.2021)

10. S. A. Chun, V. A. Paturu, S. Yuan i ostali, "Crime Prediction Model using Deep Neural Networks", Proceedings of dg.o 2019: 20th Annual International Conference on Digital Government Research, 2019., adresa: https://doi.org/10.1145/3325112.3328221 (preuzeto: 4.12.2021)

11. S. Kim, P. Joshi, P. S. Kalsi, P. Taheri, "Crime Analysis Throufg Machine Learning", Simon Fraser University, Fraser International College, 2018., adresa:

https://www.researchgate.net/publication/330475412_Crime_Analysis_Through_Machine_Learning (preuzeto: 5.12.2021)
12. V. Ingilevicha, S. Ivanovb, "Crime rate prediction in the urban environment using social factors", ITMO University, Sankt Peterburg, Ruska Federacija, 2018., adresa: https://www.sciencedirect.com/science/article/pii/S1877050918315667 (preuzeto: 7.12.2021)
13. X. Zhang, L. Liu, L. Xiao, J. Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots", IEEE Access, 2020, adresa: https://ieeexplore.ieee.org/document/9211482 (preuzeto: 5.12.2021)

8. Popis slika

9. Popis tablica

10. Popis priloga

U poglavlju prilozi navedene su web adrese na kojima se nalaze alati iz kojih je preuzet skup podataka s kojim smo radili te u kojem smo modelirali s istim podacima.

1. Kaggle (2022.). Dostupno 8.1.2022. na: https://www.kaggle.com/
2. BigML (2022.) alat za online rudarenje podataka, Dostupno 8.1.2022. na: https://bigml.com/