Predikcija emisija štetnih plinova vozila korištenjem algoritama strojnog učenja s podatcima prikupljenih OBD2 uređajem

Vaiti, Tin

SVEUČILIŠTE U ZAGREBU
FAKULTET PROMETNIH ZNANOSTI

Tin Vaiti

PREDIKCIJA EMISIJA ŠTETNIH PLINOVA VOZILA KORIŠTENJEM ALGORITAMA STROJNOG UČENJA S PODATCIMA PRIKUPLJENIH OBD2 UREĐAJEM

DIPLOMSKI RAD

Zagreb, 2022.

SVEUČILIŠTE U ZAGREBU
FAKULTET PROMETNIH ZNANOSTI
POVJERENSTVO ZA DIPLOMSKI ISPIT

Zagreb, 2. lipnja 2022.

Zavod: **Zavod za inteligentne transportne sustave**
Predmet: **Rudarenje podataka**

DIPLOMSKI ZADATAK br. 6957

Pristupnik: **Tin Vaiti (0135250937)**
Studij: Inteligentni transportni sustavi i logistika
Smjer: Inteligentni transportni sustavi

Zadatak: **Predikcija emisija štetnih plinova vozila korištenjem algoritama strojnog učenja s podatcima prikupljenih OBD2 uređajem**

Opis zadatka:

Prometni sektor pokriva tri četvrtine ukupne svjetske emisije ugljikovog dioksida, od čega najviše odlazi na osobna vozila i autobuse. Kako bi se smanjio štetan utjecaj ugljikovog dioksida na klimatske promjene, potrebno je predvidjeti i upravljati emisijama štetnih plinova osobnih vozila. Zadatak ovog diplomskog rada je izrada algoritma koji poboljšava točnost predikcije emisije štetnih plinova iz vozila primjenom metoda strojnog učenja. Koristit će se javno dostupni skup povijesnih podataka o emisijama štetnih ispušnih plinova, a po potrebi i osobno prikupljeni podaci s On-Board Diagnostic II (OBD2) uređaja ugrađenog u vozilo. Nad korištenim skupom podataka izvršit će se analiza i pred obrada podataka, koji će se koristiti za razvoj, učenje i validaciju algoritma strojnog učenja za predikciju emisije štetnih pilnova.

Zadatak uručen pristupniku: 2. lipnja 2022.

Mentor:

prof. dr. sc. Tonči Carić

Predsjednik povjerenstva za diplomski ispit:

Sveučilište u Zagrebu
Fakultet prometnih znanosti

DIPLOMSKI RAD

PREDIKCIJA EMISIJA ŠTETNIH PLINOVA VOZILA KORIŠTENJEM ALGORITAMA STROJNOG UČENJA S PODATCIMA PRIKUPLJENIH OBD2 UREĐAJEM

Prediction of vehicle emissions using machine learning algorithms with data collected by the OBD2 device

Mentor: prof. dr. sc. Tonči Carić
Neposredni voditelj: dr. sc. Tomislav Erdelić

Student: Tin Vaiti
JMBAG: 0135250937

Zagreb, rujan 2022.

Sažetak

Naslov: Predikcija emisija štetnih plinova vozila korištenjem algoritama strojnog učenja s podatcima prikupljenih OBD2 uređajem

Konstantnim povećanjem broja vozila na cestama javlja se sve veći problem emisija štetnih ispušnih plinova. Navedeno je izazvalo globalni interes za primijenjeno istraživanje u područjima analize podataka i strojnog učenja nad podacima prikupljenim s vozila. Kako bi se olakšalo praćenje količine emisija štetnih ispušnih plinova razvijeni su različiti modeli procjene štetnih emisija. U ovom diplomskom radu izrađen je model predikcije emisija ugljikovog dioksida korištenjem metoda strojnog učenja. Za treiranje modela bili su korišteni javno dostupni podaci koji se sastoje od tehničkih podataka o samom vozilu, procijenjene emisije ugljičnog dioksida za nova laka vozila te ocjene potrošnje goriva specifične za određeno vozilo. Korišteni alat za izradu modela bio je programski jezik Python u kojemu se obradila predobrada podataka, vršilo se testiranje pet različitih regresija od kojih je regresija s povećanjem gradijenata ispala najpouzdanija i koja se koristila za testiranje modela na privatno prikupljenim podacima. Za testiranje modela koristili su se osobno prikupljeni podaci pomoću OBD-II (*eng.* On-Board Diagnostics, OBD) uređaja. Rezultati primjene modela bili su prikazani na manjem dijelu rute gdje je model uspješno predvidio količinu emisija u svakom dijelu rute koristeći ulazne parametre prikupljene OBD-II uređajem.

Ključne riječi: Predviđanje štetnih emisija vozila, strojno učenje, obrađivanje podataka, OBD-II

Abstract

Title: Prediction of vehicle emissions using machine learning algorithms with data collected by the OBD2 device

With the constant increase in the number of vehicles on the roads, the problem of harmful gas emissions is emerging. This has sparked global interest in applied research in the areas of data analysis and machine learning on data collected from vehicles. In order to facilitate the monitoring of the amount of harmful gas emissions, various models for the assessment of harmful emissions have been developed. In this thesis, a carbon dioxide emission prediction model was developed using machine learning methods. Publicly available data consisting of technical data about the vehicle itself, estimated carbon dioxide emissions for new light vehicles, and vehicle-specific fuel consumption ratings were used to train the model. The tool used to create the model was the Python programming language in which the data was processed, five different regressions were tested, of which the regression with increasing gradients turned out to be the most reliable and which was used to test the model on privately collected data. To test the model, personally collected data using an OBD-II (*eng.* On-Board Diagnostics, OBD) device was used. The results of the model application were shown on a smaller part of the route where the model successfully predicted the amount of emissions in each part of the route using the input parameters collected by the OBD-II device.

Keywords: Prediction of harmful vehicle emissions, machine learning, data processing, OBD-II

Sadržaj

1. Uvod

Tijekom proteklih nekoliko godina, proizvođači automobila počeli su voditi brigu oko smanjenja emisija i ukupnog korištenja resursa goriva koji su povezani s transportnom industrijom. Ovaj rastući problem potaknuo je vladine agencije i donositelje odluka da donesu propise i standarde o učinkovitosti i niskim emisijama, [1]. Štoviše, visoki troškovi nafte, zajedno s rastućom zabrinutošću oko onečišćenja okoliša i atmosfere, prisilili su proizvođače automobila na razvoj i marketing energetski učinkovitih vozila, usvajanjem strategija kao što su projektiranje učinkovitijih motora s manjim obujmom, smanjenje težine i koeficijenta otpora vozila, korištenje niskoprofilnih guma za smanjivanje otpora kotrljanja, dodavanje elektromotora uz motor na konvencionalno gorivo, itd., [2]. Budući da je u proteklom desetljeću narasla zabrinutost zbog klimatskih promjena, modeli procjene emisija CO2 i potrošnje goriva vozila sve su važniji. Navedeno je izazvalo globalni interes za primijenjeno istraživanje u područjima analize podataka i strojnog učenja nad podacima prikupljenim s vozila, [3].

Ovaj rad fokusira se na izradu modela predikcije ugljikovog dioksida vozila, kao jednog od glavnih izvora onečišćenja u cestovnom prometu. S modelom predikcije moguće je zamijeniti fizičko testiranje emisije ugljikovog dioksida na vozilima što donosi uštedu u vremenu i povećava ekonomsku isplativost. Predstavljena metodologija temelji se na korištenju algoritama strojnog učenja naučenih na javno dostupnim povijesnim skupom podataka ugrađene dijagnostike druge generacije (OBD-II) (Vlada Kanade: Ocjene potrošnje goriva, 2022., [4]). Metodologija je podijeljena u tri glavna koraka: prethodna obrada podataka, učenje modela i primjena modela na vlastito prikupljenim podacima. Rad je podijeljen u 7 cjelina:

1. Uvod
2. Model strojnog učenja
3. OBD-II uređaji
4. Obrada podataka
5. Implementacija modela predikcije
6. Rezultati

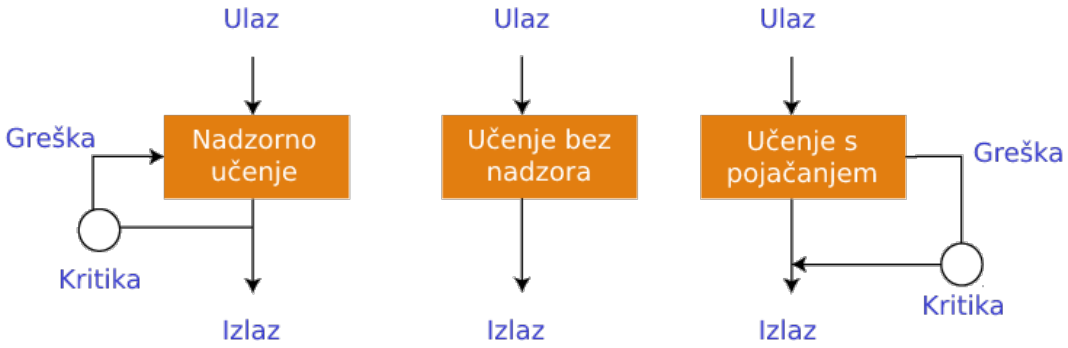
7. Zaključak

U drugom poglavlju kategorizirani su modeli strojnog učenja, te detaljnije opisani ne nadzirani i nadzirani modeli strojnog učenja.
U trećem poglavlju opisane su dvije generacije OBD uređaja, princip rada i protokoli OBD2 uređaja te podaci koji se mogu prikupljati istim.
Četvrto poglavlje opisuje korištene podatke iz skupa podataka za učenje modela te prikaz i opis koda za predobradu podataka.
Peto poglavlje opisuje korištene regresije za model, prikaz osobno prikupljenih podataka te korištena ruta za demonstraciju podataka.
Poglavlje rezultata dijeli se na rezultate regresije gdje su opisani i prikazani rezultati pokazatelja točnosti, prikaza rezultata na privatnim podacima te korelacije podataka na manjoj ruti koja sadrži razne aspekte vožnje.

2. Modeli strojnog učenja

Model strojnog učenja definiran je kao matematički prikaz rezultata procesa obuke. Strojno učenje je proučavanje različitih algoritama koji se mogu automatski poboljšati kroz iskustvo, podatke iz prošlosti i pomoću tih informacija izgraditi model. Model strojnog učenja sličan je računalnom softveru dizajniranom za prepoznavanje obrazaca ili ponašanja na temelju prethodnog iskustva ili podataka. Algoritam za učenje otkriva uzorke unutar podataka za obuku i daje model koji bilježi te uzorke i predviđa nove podatke. Model strojnog učenja može se također shvatiti kao program koji je osposobljen za pronalaženje uzoraka unutar novih podataka i stvaranje predviđanja. Ti su modeli predstavljeni kao matematička funkcija koja prima zahtjeve u obliku ulaznih podataka, obavlja predviđanje na temelju tih ulaznih podataka, a zatim daje izlaz kao odgovor. Prvo se ti modeli uče na skupu podataka, a zatim im se daje algoritam koji u pozadini stvara logičku povezanost datih podataka. Na temelju različitih skupova podataka, postoje tri modela učenja za algoritme. Svaki algoritam strojnog učenja smješta se u jedan od tri modela, [5]:

- Nadzirano učenje
- Učenje bez nadzora
- Učenje s pojačanjem



Slika 1: Modeli strojnog učenja
Izvor: [6]

Nadzirano učenje dalje se dijeli u dvije kategorije, [5]:

- Regresija
- Klasifikacija

Učenje bez nadzora također je podijeljeno u sljedeće kategorije, [5]:

- Grupiranje
- Pravilo udruživanja
- Smanjenje dimenzionalnosti

2.1. Nadzirani modeli strojnog učenja

Učenje pod nadzorom je najjednostavniji model strojnog učenja za razumijevanje gdje se ulazni podaci nazivaju podacima o obuci te imaju poznatu oznaku ili imaju rezultat kao izlaz. Učenje pod nadzorom radi na principu ulazno-izlaznih parova. Zahtijeva stvaranje funkcije koja se može uvježbati pomoću skupa podataka uvježbavanja, a zatim se primjenjuje na nepoznate podatke i postiže određene prediktivne performanse.
Model učenja pod nadzorom može se implementirati na jednostavnim problemima iz stvarnog života. Na primjer, postoji skup podataka koji se sastoji od dobi i visine; tada se može izgraditi model učenja pod nadzorom kako bi se predvidjela visina osobe na temelju njezine dobi. Modeli nadziranog učenja dalje se klasificiraju u dvije kategorije, a to su regresija i klasifikacija.

2.1.1. Regresija

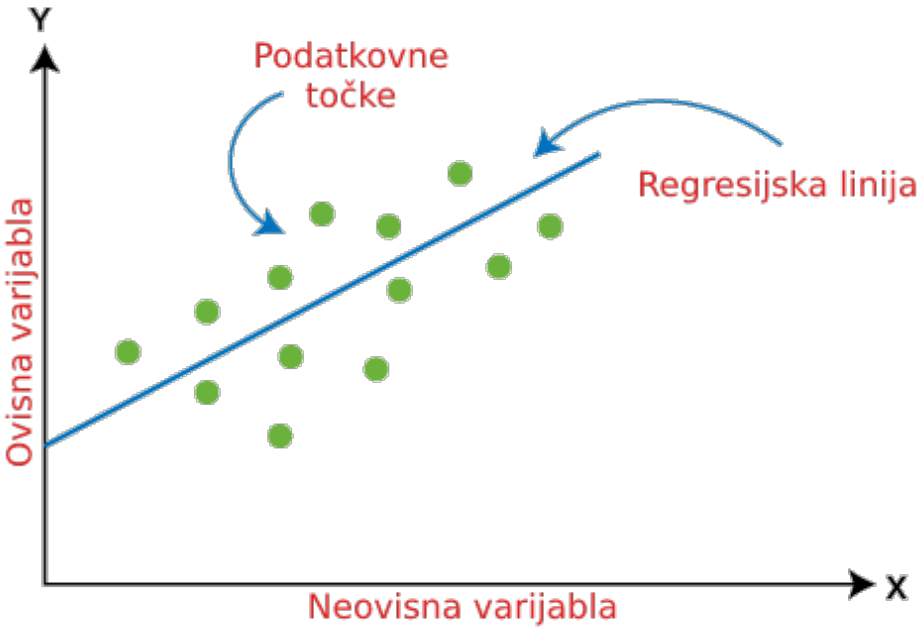
U problemima regresije, izlaz je kontinuirana varijabla. Neki često korišteni regresijski modeli su sljedeći:

a) Linearna Regresija

Linearna regresija može se proširiti na najjednostavniji model strojnog učenja u kojem se pokušava predvidjeti jedna izlazna varijabla pomoću jedne ili više ulaznih varijabli. Prikaz linearne regresije je linearna jednadžba koja kombinira skup ulaznih vrijednosti (X) i predviđeni izlaz (Y) za skup tih ulaznih vrijednosti. Predstavljen je u obliku linije i opisuje se jednadžbom: grupiranje, pravilo udruživanja i smanjenje dimenzionalnosti.

$$Y = a + bX \tag{1}$$

Primjerice na slici 3 može se vidjeti glavni cilj modela linearne regresije što je pronaći liniju koja najbolje odgovara točkama podataka.

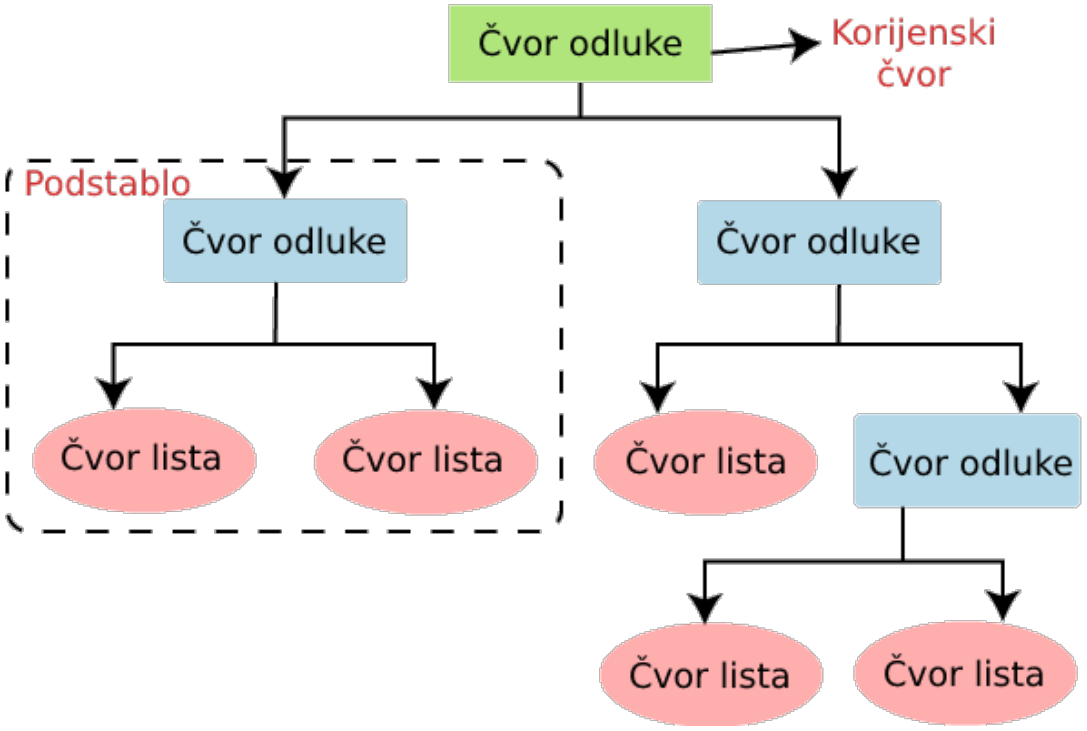


Slika 2: Prikaz rada linearne regresije
Izvor: [6]

Linearna regresija je proširena na višestruku linearnu regresiju (pronalaženje ravnine) i polinomijalnu regresiju (pronalaženje krivulje), [7].

b) Stablo odluke

Stabla odlučivanja popularni su modeli strojnog učenja koji se mogu koristiti za probleme regresije i klasifikacije. Stablo odluka koristi strukturu odluka poput stabla zajedno s njihovim mogućim posljedicama i ishodima.



Slika 3: Prikaz rada stabla odluke
Izvor: [8]

Svaki unutarnji čvor predstavlja test na atributu, a svaka se grana koristi za predstavljanje ishoda testa. Što više čvorova stablo odlučivanja ima, rezultat će biti točniji. Prednost stabala odlučivanja je u tome što su intuitivna i laka za implementaciju, ali im nedostaje točnosti. Stabla odlučivanja često se koriste u operacijskim istraživanjima, posebno u analizi odluka, strateškom planiranju i u strojnom učenju, [7].

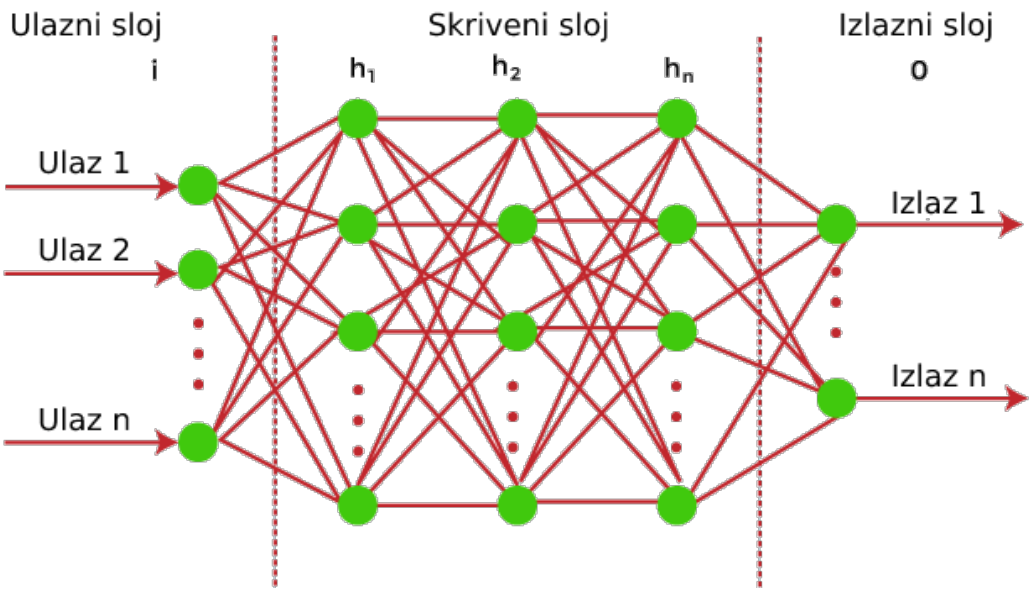
c) Nasumična šuma

Nasumična šuma je metoda učenja koja se sastoji od velikog broja stabala odlučivanja. Svako stablo odlučivanja u slučajnoj šumi predviđa ishod, a predviđanje s većinom glasova smatra se ishodom. Slučajni šumski model može se koristiti za probleme regresije i klasifikacije. Za zadatak klasifikacije, ishod slučajne šume uzima se iz većine glasova, dok se u zadatku regresije ishod uzima iz srednje vrijednosti ili prosjeka predviđanja koje generira svako stablo, [7].

d) Neuronske mreže

Neuronske mreže su podskup strojnog učenja i poznate su i kao umjetne neuronske mreže. Neuronske mreže sastoje se od umjetnih neurona i dizajnirane su tako da sliče strukturi i radu ljudskog mozga. Svaki umjetni neuron povezuje se s mnogim drugim neuronima u neuronskoj

mreži, a takvi milijuni povezanih neurona stvaraju sofisticiranu kognitivnu strukturu.



Slika 4: Prikaz slojeva neuronske mreže
Izvor: [6]

Neuronske mreže sastoje se od višeslojne strukture koja sadrži jedan ulazni sloj, jedan ili više skrivenih slojeva i jedan izlazni sloj, kao što se može vidjeti na slici 4. Kako je svaki neuron povezan s drugim neuronom, on prenosi podatke iz neurona jednog sloja na neurone drugog (sljedećih) sloja. Konačno, podaci dolaze do posljednjeg sloja ili izlaznog sloja neuronske mreže i generiraju izlaz. Neuronske mreže ovise o podacima za učenje i poboljšanje točnosti, međutim, savršeno uvježbana i precizna neuronska mreža može brzo grupirati podatke i postati moćan alat za strojno učenje i umjetnu inteligenciju. Jedna od najpoznatijih primjena neuronskih mreža je Googleov algoritam pretraživanja, [7].

2.1.2. Klasifikacija

Klasifikacijski modeli su druga vrsta tehnika nadziranog učenja, koje se koriste za generiranje zaključaka iz promatranih vrijednosti u kategoričkom obliku. Na primjer, model klasifikacije može identificirati je li e-pošta spam ili ne; hoće li kupac kupiti proizvod ili neće i slično. Klasifikacijski algoritmi koriste se za predviđanje dviju klasa i kategorizaciju izlaza u različite skupine. U klasifikaciji je dizajniran model klasifikatora koji klasificira skup podataka u različite kategorije, a svakoj kategoriji se dodjeljuje oznaka. Postoje dvije vrste klasifikacija u strojnom učenju:

- Binarna klasifikacija: Ako problem ima samo dvije moguće klase, naziva se binarni klasifikator. Na primjer, 1 ili 0, Da ili Ne.
- Klasifikacija s više klasa: Ako problem ima više od dvije moguće klase, radi se o klasifikatoru s više klasa. Na primjer ako se klasificiraju vozila po veličini i algoritam klasifikacije je pronašao 3 klase one mogu biti mala, srednja i velika vozila.

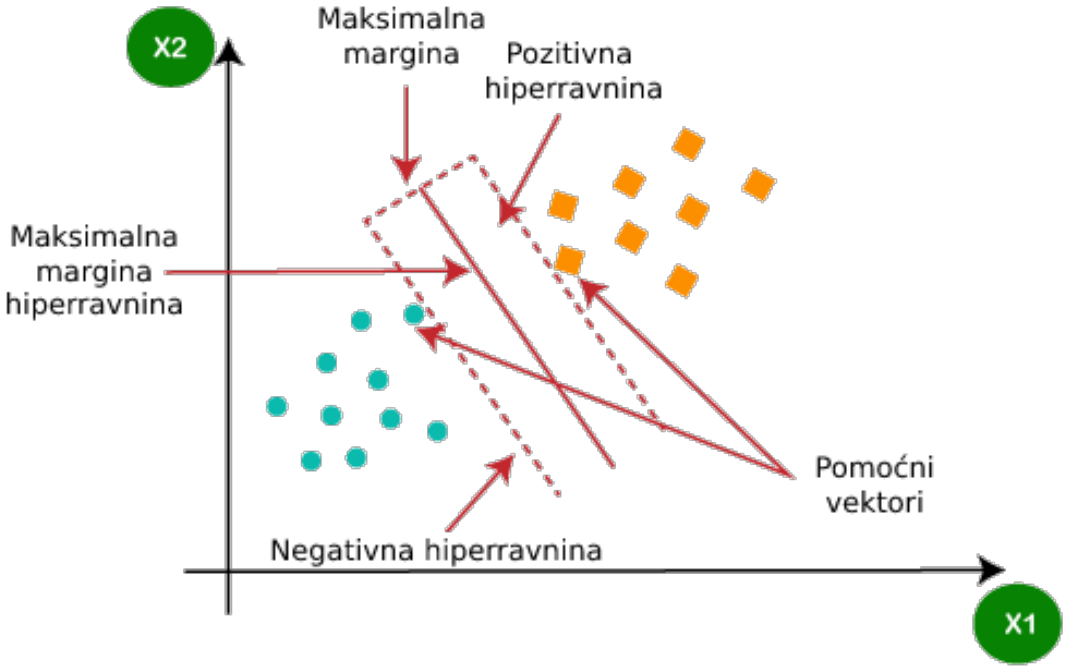
Neki popularni algoritmi klasifikacije su sljedeći:

a) Logistička regresija

Logistička regresija koristi se za rješavanje problema klasifikacije u strojnom učenju. Slični su linearnoj regresiji, ali se koriste za predviđanje kategoričkih varijabli. Može predvidjeti izlaz u binarnim slučajevima (Da ili Ne, 0 ili 1, True ili False i sl.). Rezultati logističkih regresija nisu točne vrijednosti, nego vjerojatnosne vrijednosti između 0 i 1, [9].

b) Metoda potpornih vektora

Metoda potpornih vektora (*eng.* Support Vector Machine, SVM) popularan je algoritam strojnog učenja koji se široko koristi za zadatke klasifikacije i regresije.



Slika 5: Prikaz rada algoritma strojno potpunih vekora
Izvor: [6]

Glavni cilj SVM-a je pronaći granice najbolje odluke u N-dimenzionalnom prostoru, koji može

razdvojiti točke podataka u klase, a granica najbolje odluke poznata je kao hiperravnina. SVM odabire ekstremni vektor za pronalaženje hiperravnine, a ti su vektori poznati kao potporni vektori, [9].

c) Naivni bayes

Naivni Bayes još je jedan popularan algoritam klasifikacije koji se koristi u strojnom učenju. Naziva se tako jer se temelji na Bayesovom teoremu i slijedi "naivnu" (neovisnu) pretpostavku između značajki koja je dana kao:

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)} \tag{2}$$

Svaki naivni Bayesov klasifikator pretpostavlja da je vrijednost određene varijable neovisna o bilo kojoj drugoj varijabli/značajki. Na primjer, ako voće treba klasificirati na temelju boje, oblika i okusa, tako bi se žuto, ovalno i kiselo prepoznalo kao limun. Ovdje je svaka značajka neovisna o drugim značajkama, [9].

2.2. Modeli strojnog učenja bez nadzora

Modeli strojnog učenja bez nadzora implementiraju proces učenja suprotan nadziranom učenju, što znači da omogućuje modelu učenje iz neoznačenog skupa podataka za obuku. Na temelju neoznačenog skupa podataka model predviđa izlaz. Koristeći učenje bez nadzora, model sam uči skrivene obrasce iz skupa podataka bez ikakvog nadzora. Modeli učenja bez nadzora uglavnom se koriste za izvođenje tri zadatka, a to su: grupiranje, pravilo udruživanja i smanjenje dimenzionalnosti.

2.2.1. Grupiranje

Grupiranje je tehnika učenja bez nadzora koja uključuje grupiranje ili grupiranje podatkovnih točaka u različite grupe na temelju vektorske udaljenosti točaka (podataka) jedne od druge. Objekti s njamanjom udaljenošću jednog od drugog ostaju u istoj skupini. Algoritmi klasteriranja mogu se široko koristiti u različitim zadacima kao što su segmentacija slike, analiza statističkih podataka, segmentacija tržišta itd.https://www.overleaf.com/project/62e8f224fbee65c698574e49 Neki često korišteni algoritmi klasteriranja su K-means klasteriranje, hijerarhijsko klasteriranje, DBSCAN, sl., [10].

Slika 6: Prikaz grupacije podataka
Izvor: [6]

2.2.2. Pravilo udruživanja

Pravilo udruživanja je tehnika učenja bez nadzora, koja pronalazi logičke odnose među varijablama unutar velikog skupa podataka. Glavni cilj ovog algoritma za učenje je pronaći ovisnost jedne podatkovne stavke o drugoj podatkovnoj stavci i mapirati te varijable u skladu s tim tako da se može generirati maksimalna dobit. Ovaj se algoritam uglavnom primjenjuje u analizi tržišne košarice, rudarenju korištenja weba, kontinuiranoj proizvodnji itd. Neki popularni algoritmi učenja pravila pridruživanja su A priori algoritam, Eclat, FP-algoritam rasta i drugi, [10],

2.2.3. Smanjenje dimenzionalnosti

Broj značajki, odnosno varijabli prisutnih u skupu podataka poznati su kao dimenzionalnost skupa podataka, a tehnika koja se koristi za smanjenje dimenzionalnosti poznata je kao tehnika redukcije dimenzionalnosti. Iako više podataka daje točnije rezultate, to također može utjecati na performanse modela, gdje se u takvim slučajevima koriste se tehnike smanjenja dimenzionalnosti, [10].

2.3. Učenje s pojačanjem

U učenju s pojačanjem, algoritam uči akcije za zadani skup stanja koje dovode do ciljnog stanja. To je model učenja temeljen na povratnim informacijama koji uzima povratne signale nakon svakog stanja ili radnje interakcijom s okolinom. Ova povratna informacija djeluje kao nagrada (pozitivna za svaku dobru radnju i negativna za svaku lošu radnju), a cilj je maksimizirati

pozitivne nagrade kako bi se poboljšala izvedba. Ponašanje modela u učenju s potkrepljenjem slično je ljudskom učenju, budući da ljudi uče stvari kroz iskustva kao povratne informacije i u interakciji s okolinom. Ispod su neki popularni algoritmi koji spadaju u učenje s pojačanjem su sljedeći, [11]:

- Q-učenje - jedan od Q-učenja popularnih algoritama za učenje bez modela koji se temelji na Bellmanovoj jednadžbi.
- Stanje-akcija-nagrada-stanje-akcija (*eng.* State-Action-Reward-State-Action, SARSA) - on-policy algoritam temeljen na Markovljevom procesu odlučivanja. Koristi radnju koju izvodi trenutni Markovljev proces za učenje Q-vrijednosti.
- Duboka Q mreža (*eng.* Deep Q Network, DQN) - Q-učenje unutar neuronske mreže. U osnovi se koristi u okruženju velikog prostora stanja gdje bi definiranje Q-tablice bio složen zadatak. Dakle, u takvom slučaju, umjesto da se koristi Q-tablica, neuronska mreža koristi Q-vrijednosti za svaku akciju na temelju stanja.

3. OBD2 uređaji

Ugrađena dijagnostika je pojam koji se odnosi na sposobnost samodijagnostike i informiranja vozila. OBD sustavi daju vlasniku vozila ili tehničaru za popravak pristup statusu različitih podsustava vozila. Količina dijagnostičkih informacija dostupnih putem OBD-a uvelike je varirala od njegovog uvođenja u ranim 1980-im ovisno o verzijama računala u vozilu. Rane verzije OBD-a jednostavno bi osvijetlile svjetlo indikatora neispravnosti ako bi se pojavio problem, ali ne bi pružile nikakve informacije o prirodi problema. Suvremene OBD implementacije koriste standardizirani digitalni komunikacijski priključak za pružanje podataka u stvarnom vremenu uz standardiziranu seriju dijagnostičkih kodova kvarova, koji omogućuju osobi da brzo identificira i otkloni kvarove unutar vozila, [12].

3.1. Prva generacija OBD uređaja

Regulatorna namjera OBD-I bila je potaknuti proizvođače automobila da dizajniraju pouzdane sustave za kontrolu emisije koji ostaju učinkoviti tijekom vijeka trajanja vozila. OBD-I je bio uglavnom neuspješan, budući da sredstva za prijavu dijagnostičkih informacija specifičnih za emisije nisu bila standardizirana. Tehničke poteškoće s dobivanjem standardiziranih i pouzdanih informacija o emisijama iz svih vozila dovele su do nemogućnosti učinkovite provedbe godišnjeg programa testiranja, [12].

Slika 7: Prikaz OBD-I priključka
Izvor: [13]

Dijagnostički kodovi kvarova (*eng.* Diagnostic Trouble Code, DTC) OBD-I vozila obično se mogu pronaći bez skupog alata za skeniranje. Svaki je proizvođač koristio vlastiti konektor dijagnostičke veze (*eng.* Diagnsotic Link Connector, DLC), lokaciju DLC-a, definicije DTC-a i postupak za očitavanje DTC-ova iz vozila. DTC-ovi s OBD-I automobila često se očitavaju kroz treptanje svjetla 'Check Engine Light' (CEL) ili 'Service Engine Soon' (SES). Spajanjem određenih pinova dijagnostičkog konektora, svjetlo 'Check Engine' će treptati dvoznamenkastim brojem koji odgovara određenom stanju greške, [12].

3.2. Druga generacija OBD uređaja

OBD-II je poboljšanje u odnosu na OBD-I u mogućnostima i standardizaciji. Standard OBD-II specificira vrstu dijagnostičkog konektora i broj izlaznih pinova, dostupne električne signalne protokole i format poruka. Također pruža popis kandidata za parametre vozila za praćenje zajedno s načinom kodiranja podataka za svaki. U konektoru se nalazi igla koja omogućuje napajanje alata za ispitivanje iz akumulatora vozila, što eliminira potrebu za zasebnim spajanjem alata za ispitivanje na izvor napajanja. Neki tehničari ipak mogu spojiti alat za skeniranje na pomoćni izvor napajanja kako bi zaštitili podatke u neuobičajenom slučaju da vozilo doživi gubitak električne energije zbog kvara. Konačno, standard OBD-II pruža proširiv popis DTC-

ova. Kao rezultat ove standardizacije, jedan uređaj može postavljati upite putnom računalu (računalima) u bilo kojem vozilu.

Slika 8: Prikaz OBD-II priključka
Izvor: [14]

Standardizacija OBD-II potaknuta je zahtjevima za emisije. Iako se preko njega moraju prenositi samo kodovi i podaci koji se odnose na emisije, većina je proizvođača napravila konektore za OBD-II uređaje kao jedini u vozilu putem kojeg se dijagnosticiraju svi sustavi vozila. OBD-II dijagnostički kodovi kvarova su četveroznamenkasti kodovi, ispred njih stoji slovo: P za pogon (motor i prijenos), B za karoseriju, C za šasiju i U za mrežu [15].

3.2.1. Princip rada OBD-II uređaja

U automobilu postoje razni senzori: senzori za mjerenje vanjske temperature zraka, lambda senzor, senzori tlaka u razvodniku itd. Svaki od ovih senzora šalje signal na centralno računalo automobila, Upravljačkoj jedinici motora (*eng.* Engine Control Unit, ECU). ECU koristi te informacije za prilagodbu različitih elemenata rada motora, npr. ubrizgavanje goriva u motor. Ako je informacija koju ECU dobije od jednog od svojih senzora izvan standardnog intervala točnosti, u memoriju se sprema DTC. Također ECU šalje signal i vozaču kao trepćuće svjetlo odgovarajuće oznake na instrument ploči, [16].

3.2.2. Podaci prikupljeni OBD-II uređajem

OBD-II uređaj može očitati više od 200 podataka iz automobila, pri čemu se upit šalje kao heksadekadska kombinacija dijagnostičke usluge i PID-a (*eng.* Parameter IDs). Postoji 10 dijagnostičkih usluga opisanih u najnovijem OBD-II standardu SAE J1979, prikazanih u tablici 1. PID-ovi su kôdovi koji se koriste za traženje podataka s vozila, a koriste se kao dijagnostički alat. SAE standard J1979 definira raznovrsne PID-ove koji se koriste u svakom automobilu, ali proizvođači automobila također mogu definirati dodatne PIDove specifične za njihova vozila, [16].

Tablica 1: Dijagnostičke usluge

Usluga	**Opis**
01	Prikaz trenutnih podataka
02	Prikaz zamrznutih podataka
03	Prikaz pohranjenih dijagnostičkih kôdova s greškama
04	Brisanje pohranjenih dijagnostičkih kôdova
05	Rezultati ispitivanja praćenja senzora za kisik
06	Rezultati ispitivanja ostalih komponenti sustava
07	Prikaz dijagnostičkih kôdova za probleme koji su na čekanju
08	Kontrola za upravljanje on-board sustavom
09	Prikaz podataka o vozilu
0A	Trajni dijagnostički kôdovi s problemima

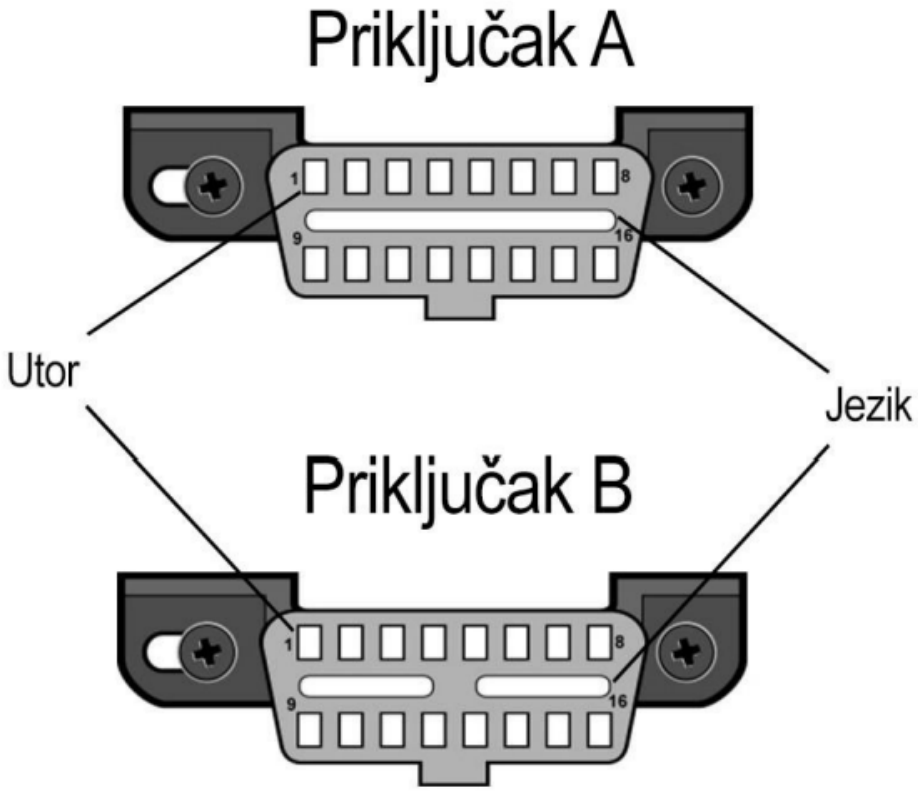
U ovom radu koristili su se podaci očitani s OBD-II uređaja prikazani tablicom 2.

Tablica 2: Komande i mjerne jedinice korištenih podataka

Ime Komande	Mjerna jedinica
Brzina vozila	km/h
Vrijeme proteklo od paljenja motora	s
Intenzitet potrošnje goriva	km/L
Razina goriva u spremniku	postotak
Praćenje paljenja motora	-
Voltaža upravljačkog modula	V
Pozicija papučice gasa	postotak
Okretaji motora u minuti	rpm
Temperatura ulja u motoru	°C
Maseni protok zraka	g/s
Apsolutno opterećenje	postotak
Pritisak goriva	kPa
Barometarski pritisak	kPa
Tlak u šini za gorivo	kPa
Tlak usisnog goriva	kPa
Temperatura rashladne tekućine u motoru	°C
Temperatura zraka	°C

3.3. Protokoli OBD-II uređaja

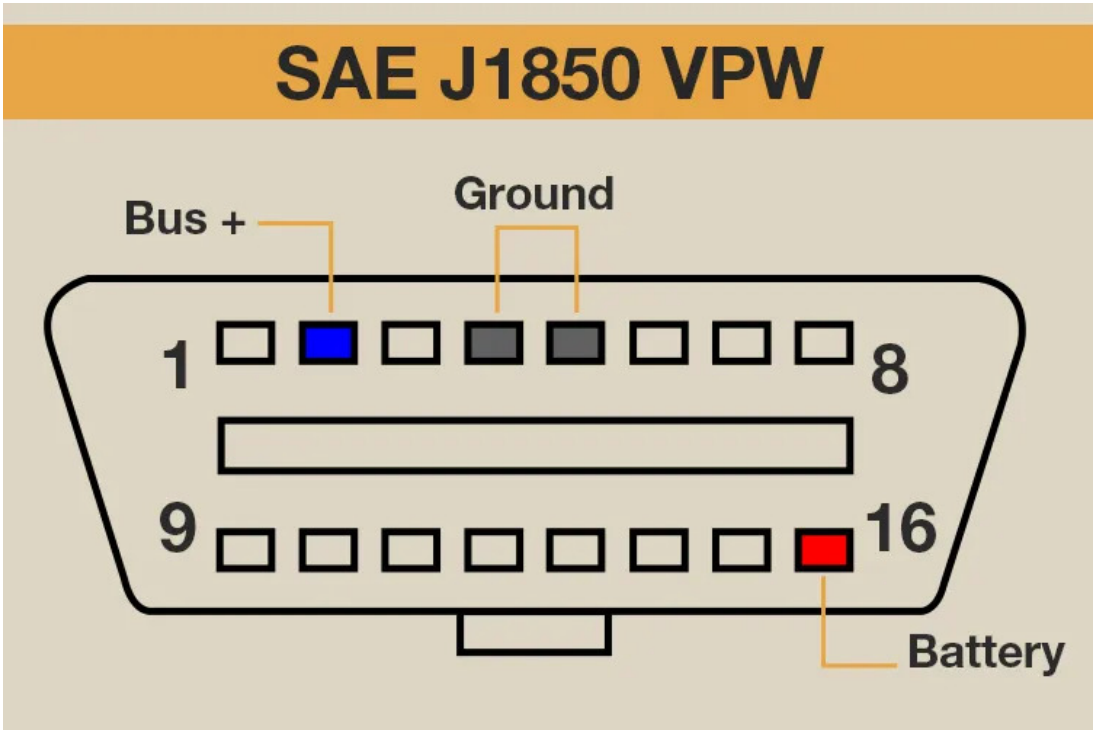
OBD-II sustav općenito ima pet protokola i dvije vrste priključka. Različiti modeli koriste različite protokole. Moguće je da automobil ima priključak tipa A ili priključak tipa B. Oba imaju fizičku razliku u svojim utorima. Konektori tipa A imaju 16 utora raspoređenih u dva reda. Svaki red ima 8 utora, a po sredini ih dijeli jedan veliki utor zvan „jezik“. Konektori tipa B također imaju 16 utora, ali jezik im se dijeli na dva dijela.



Slika 9: Prikaz A i B konektora
Izvor: [17]

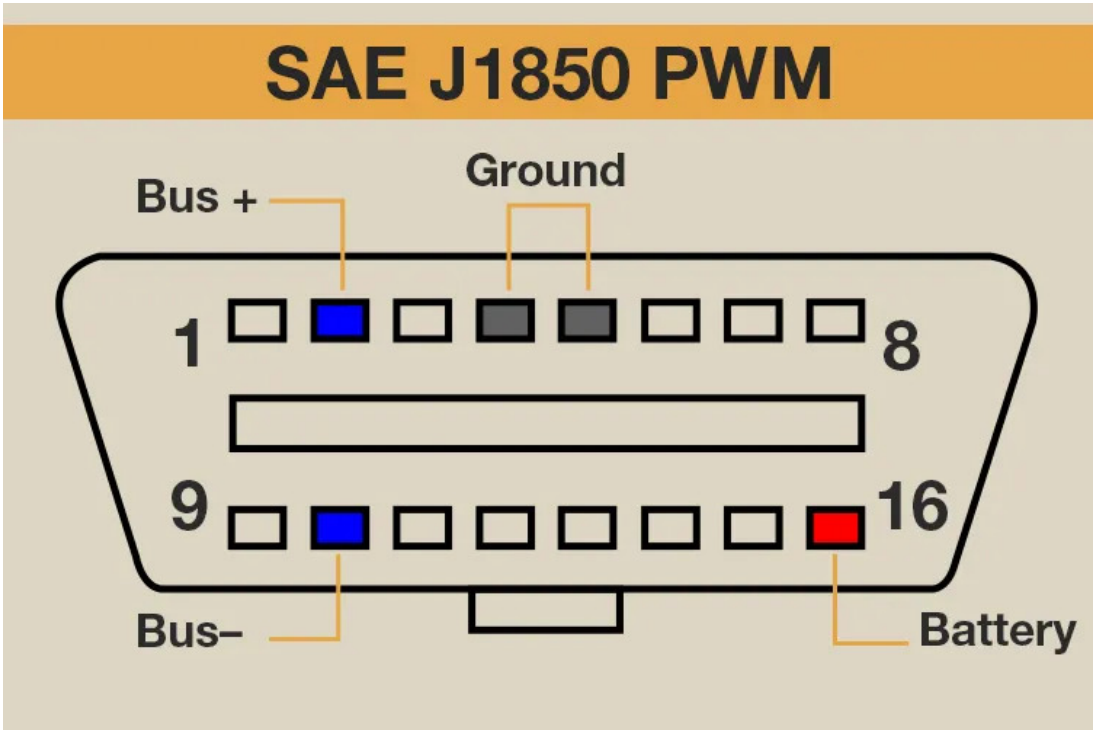
Postoji 5 vrsta OBD-II protokola:

1. SAE J1850 VPW - Pin 2 je obavezan. Konektor također mora imati materijalne kontakte unutar pinova 4, 5 i 16. General Motors prvenstveno koristi ovaj protokol.



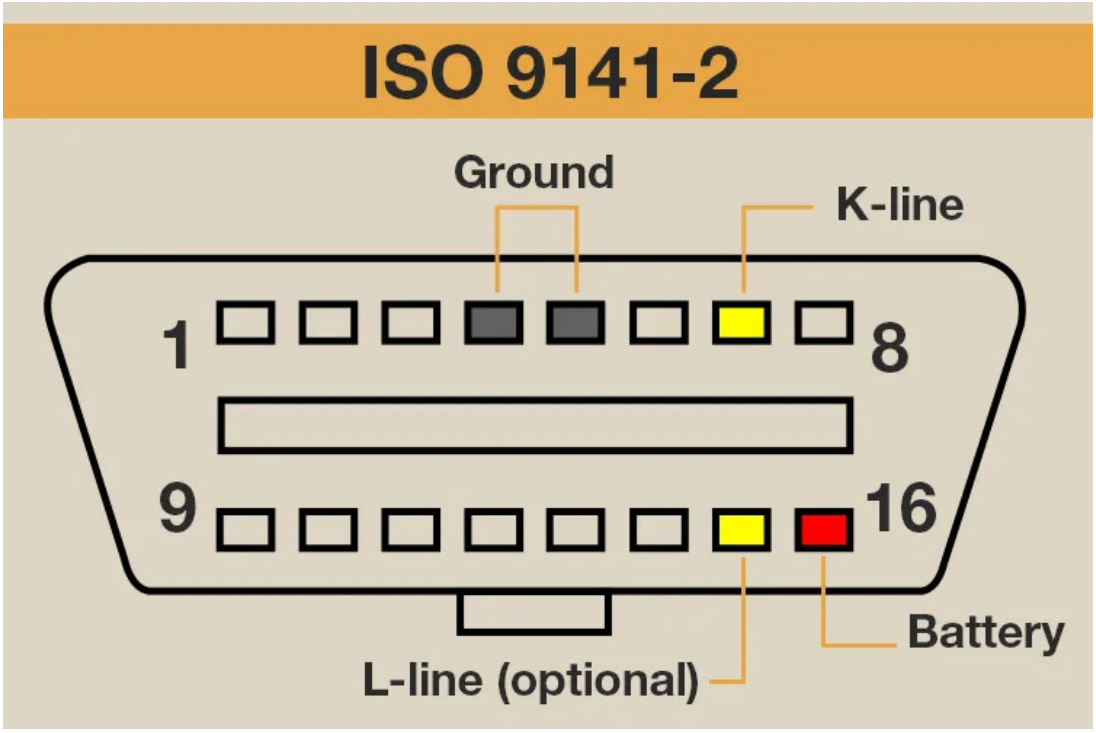
Slika 10: Protokol SAE J1850 VPW
Izvor: [18]

2. SAE J1850 PWM - Mora imati pinove 2, 4, 5, 10 i 16. Obično Ford koristi ovaj protokol.



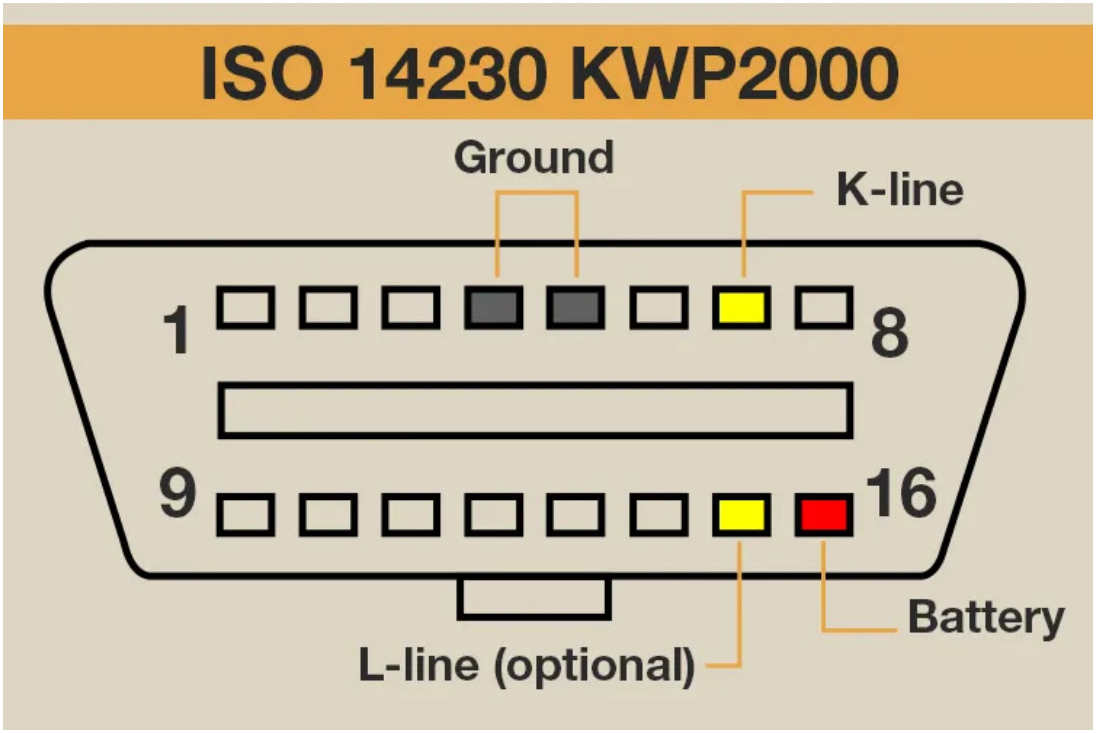
Slika 11: Protokol SAE J1850 PWM
Izvor: [18]

3. ISO 9141-2 - Pin 7 je obavezan (pinovi 4, 5 i 16 moraju se imati). ISO 9141-2 koristi se u vozilima Chrysler, europskim i azijskim vozilima.



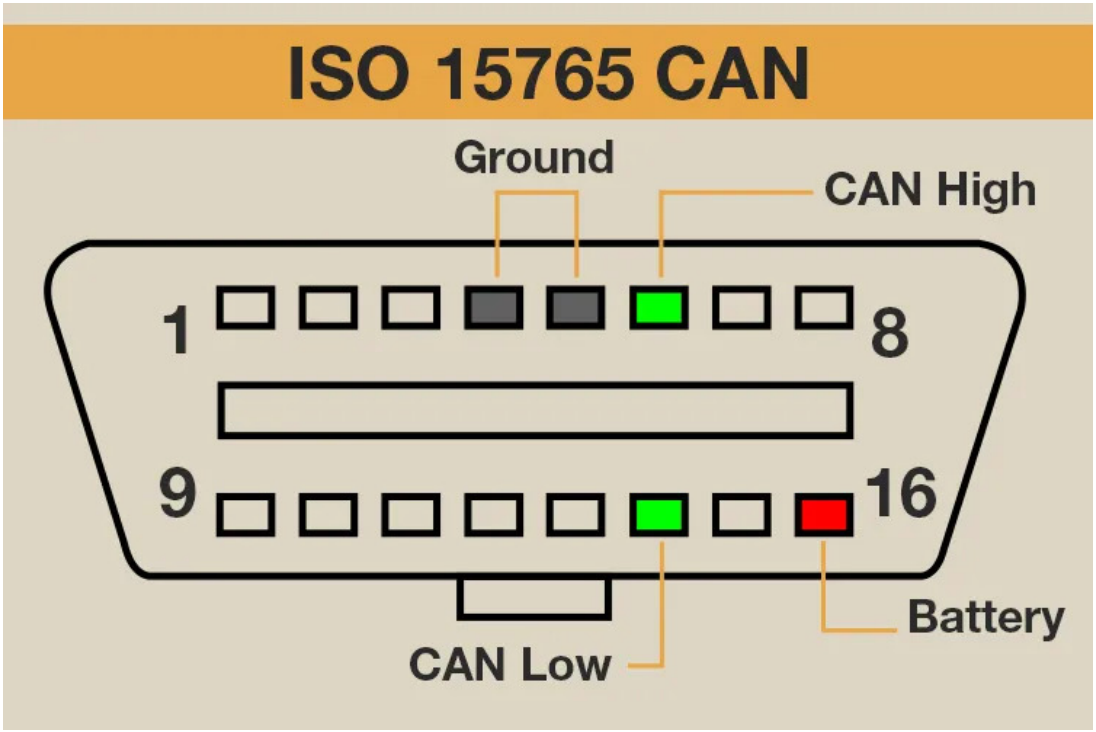
Slika 12: Protokol ISO 9141-2 i KWP2000
Izvor: [18]

4. ISO 14230 KWP2000 - Isti uzorak igala kao ISO 9141-2, ali ga možete pronaći samo u azijskim vozilima.



Slika 13: Protokol ISO 9141-2 i KWP2000
Izvor: [18]

5. ISO 15765 CAN - Pinovi 6 i 14 trebali bi biti tamo, dok bi metalni kontakti trebali imati pinove 4, 5 i 16. Ovaj se protokol može pronaći u američkim vozilima od 2008, [16].



Slika 14: Protokol ISO 15765-4/SAE J2480 (CAN)
Izvor: [18]

4. Obrada podataka

Ovaj dokument koristi javno dostupan skup podataka Vlade Kanade koji izdvaja značajke vozila na temelju standarda OBD-II. Korištene značajke za učenje algoritama strojnog učenja, s njihovim odgovarajućim opisima prikazane su u tablici 3.

Tablica 3: Značajke javno dostupnog skupa podataka koji se koristi za predikciju emisija

Značajka	**Mjerna jedinica**	**Raspon vrijednosti**
Klasa vozila	Broj (Integer)	1 - 10
Volumen motora	Litre (Float)	1.2 - 8
Cilindar	Broj (Integer)	3 - 16
Prijenos	Broj (Integer)	0 - 22
Tip goriva	Broj (Integer)	0 - 3
Potrošnja goriva	Litre (Float)	4 - 26.1
Emisija CO2	Broj (Integer)	0 - 608

Klasu vozila definiraju se putem tablica koje se sastoje od godine izrade vozila, potrošnje goriva, emisije ugljikovog dioksida te same ocjene. Vozila su grupirana u 10 klasa gdje svaka klasa reprezentira ocjenu koja govori koliko je određeno vozilo jaki zagađivač stakleničkih plinova. Kako se postrožuju zakoni o količini dopuštene emisije štetnih ispušnih plinova vozila, sukladno se povećavaju i kriteriji po kojima se ocjenjuju vozila. Tako primjerice novo vozilo iz 2013. godine s potrošnjom od 7.2 - 8.3 L/100 km i proizvodnjom 138 - 162 g/km ugljikovog dioksida ocijenilo bi se ocjenom 9, dok bi se isto to vozilo po kriterijima 2023. godine ocijenilo s ocjenom 7, [19]. Volumen motora je izražena u litrama goriva koje troši, značajka cilindar predstavlja broj cilindara u motoru vozila, prijenos je povezan s vrstom mjenjača (automatski ili ručni) i brojem stupnjeva prijenosa. Značajka tip goriva sadrži tri moguća slučaja goriva, i to obični benzin, vrhunski benzin i dizel. Na kraju je značajka potrošnja goriva čija je mjerna jedinica u litrama. Značajke se pretvaraju u numeričke varijable predstavljanjem značajke rasponom brojeva jer su predstavljene kao kategoričke značajke (riječi) u izvornom skupu podataka. Zadnja značajka je emisija ugljikovog dioksida koja će se u radu koristiti kao podatak za provjeru točnosti predikcije, a označava količinu emisije CO2 koju je proizvođač automobila izmjerio za određeni model automobila.

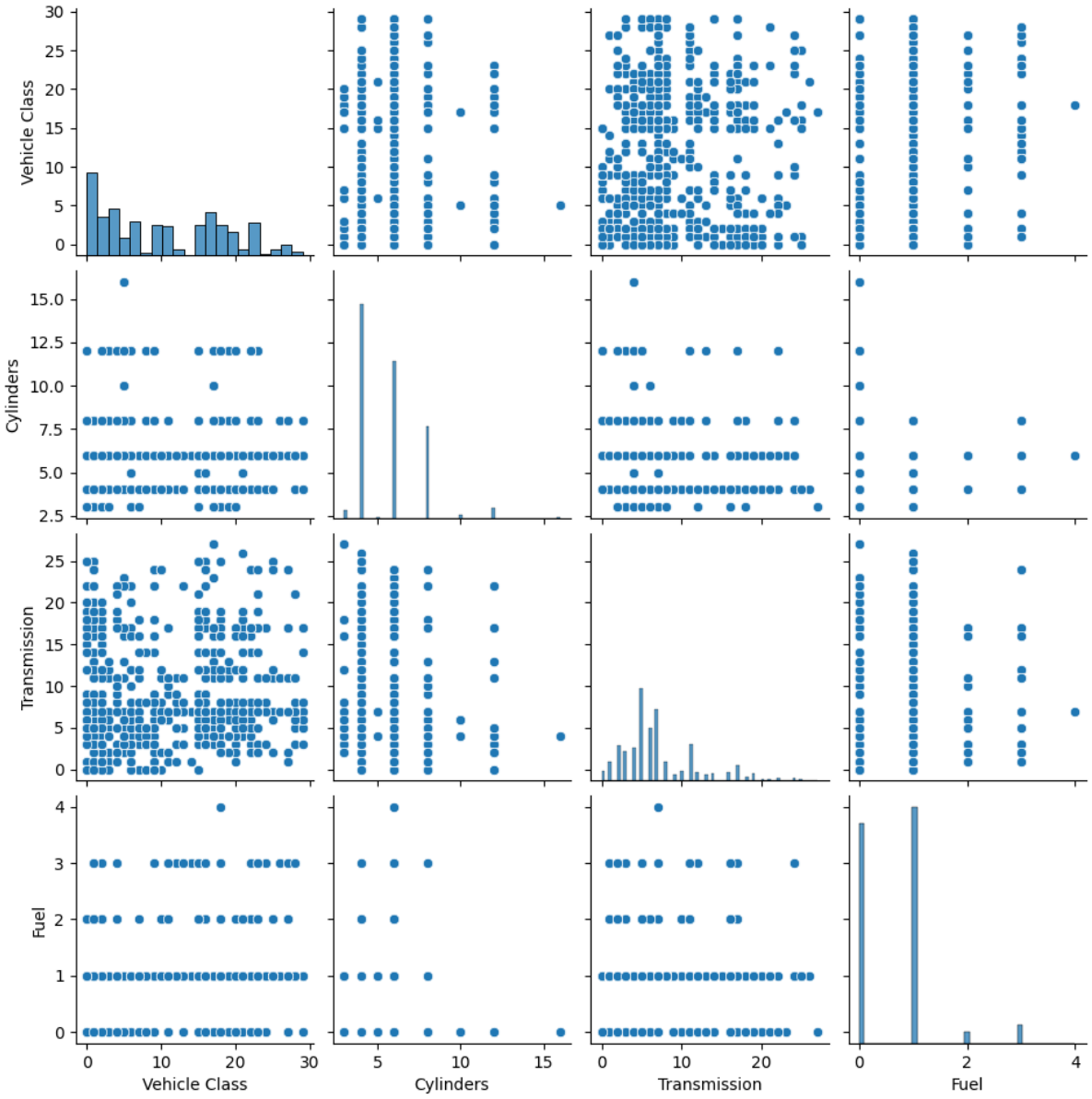
Dio jednostavnog inženjeringa značajki je određivanje značajki za obuku algoritama strojnog

učenja. U istraživanju su zanemarene dvije vrste značajki. Prvi tip je vezan uz proizvođača vozila i naziv modela vozila. Te su značajke isključene zbog fokusa istraživanja na modeliranju klasa emisija vozila bez obzira na proizvođačev povijesni razvoj vozila. Uz ovo isključenje, neka neučinkovita ili velika vozila istog proizvođača neće utjecati na druga, učinkovitija vozila.
Skup podataka nalazi se u obliku csv datoteke koja sadrži 8288 stupaca s podacima te je prvih pet prikazano na slici 15. Prilikom predobrade podataka bio je pronađen samo jedan stupac koji je imao nepotpune vrijednosti te se zbog toga i odbacio.

	Year	Vehicle Class	Engine Size	Cylinders	Transmission	Fuel	Fuel Consumption	CO2_Emissions
1	2022	0	2.4	4.00000	0	0	9.9	200
2	2022	1	3.5	6.00000	1	0	12.6	263
3	2022	1	2	4.00000	1	0	11	232
4	2022	1	2	4.00000	1	0	11.3	242
5	2022	0	2	4.00000	1	0	11.2	230

Slika 15: Prikaz prvih 5 podataka iz skupa

Također su pronađeni ekstremi u podacima što su bila sportska vozila s velikim obujmom motora, potrošnjom goriva i velikim brojem cilindra kojih je bilo sveukupno 188, a mogu se uočiti na slici 16. Slika također pokazuje povezanost podataka za lakšu predodžbu prilikom predobrade kako bi se raspoznala njihova povezanost. Povezanost između godina, obujma motora, potrošnje goriva i CO2 emisije nije bila pronađena zbog čega nisu prikazani na slici.



Slika 16: Prikaz povezanosti podataka

4.1. Python i strojno učenje

U ovom radu koristio se Python za izradu modela štetnih ispušnih plinova. Python je programski jezik visoke razine opće namjene koji je postao popularan u posljednje vrijeme. Omogućuje pisanje kôda u manje redaka što nije moguće s drugim jezicima. Važna značajka Pythona je da podržava više programskih paradigmi. Python nudi veliki skup biblioteka koje su napisali drugi programeri i mogu se koristiti za raznovrsne namjene. Glavne značajke Pythona su: lako se nauči, programski jezik visoke razine, prenosivost, objektno orijentirani jezik i druge značajke, [5].

Izraz strojno učenje odnosi se na automatizirano otkrivanje smislenih obrazaca u podacima. U posljednjih nekoliko desetljeća postaje uobičajeni alat u gotovo svim zadacima koji zahtijevaju ekstrakciju informacija iz velikih skupova podataka. Postoje razni primjeri tehnologija koje se temelje na strojnom učenju: tražilice uče kako donijeti najbolje rezultate, automobili su opremljeni sustavima za sprječavanje nesreća koji su izgrađeni pomoću algoritama strojnog učenja i slično, [5].

U ovom radu korištene su sljedeće biblioteke koje nudi Python za obradu podataka i rad s algoritmima strojnog učenja:

- **Numpy -** je biblioteka za programski jezik Python, koja dodaje podršku za velike, višedimenzionalne nizove i matrice, zajedno s velikom zbirkom matematičkih funkcija visoke razine za rad s tim nizovima, [20].
- **Pandas -** je softverska biblioteka napisana za programski jezik Python za manipulaciju i analizu podataka. Konkretno, nudi strukture podataka i operacije za manipuliranje numeričkim tablicama i vremenskim serijama, [21].
- **Sklearn -** je besplatna softverska biblioteka strojnog učenja za programski jezik Python. Sadrži različite algoritme za klasifikaciju, regresiju i grupiranje; uključujući metodu potpornih vektora, nasumične šume, povećanje gradijenta, k-srednje vrijednosti i DBSCAN, a dizajniran je za međuoperativnost s Python numeričkim i znanstvenim bibliotekama NumPy i SciPy, [22].

4.2. Predobrada podataka

Koristeći otvorene podatke sa sobom donosi i jednu manu koja je da često sadrže puno nepotrebnih podataka koji bi smanjili kvalitetu modela predikcije. Zbog toga se uvodi predobrada podataka koja omogućuje korištenje određenih stupca i podatka koji su potrebni kako bi se postigla maksimalna kvaliteta rezultata. Skup podataka sadrži 14 stupaca od koji se za učenje modela koristi njih 7, koji se mogu očitati iz tablice 3.

```python
class DataPreprocessing:

    def __init__(self, file_path):
        self.file_path = file_path
        self.preprocessed_data = None

    def _data_processing(self):
        use_cols = ['Vehicle Class', 'Engine Size', 'Cylinders', 'Transmission',
                    'Fuel', 'Fuel Consumption', 'CO2']
        df = pd.read_csv(self.file_path, sep=',', encoding='cp1252', low_memory=False, usecols=use_cols)

        df = df.rename(columns={'CO2': 'CO2 Ratings'})
        df = df.dropna()

        self._factorize_data(df, ['Vehicle Class', 'Transmission', 'Fuel'])

        self.preprocessed_data = df

    def get_preprocessed_data(self):
        self._data_processing()
        return self.preprocessed_data

    @staticmethod
    def _factorize_data(df, column_names):
        for name in column_names:
            df[name] = pd.factorize(df[name])[0]
```

Slika 17: Predobrada podataka

Na slici 17 prikazan je programski kôd za preprocesiranje podataka koji se sastoji od jedne klase zvane *DataPreprocessing* i 4 funkcije:

- *__init__* - inicijalna funkcija koja prilikom poziva klase traži putanju gdje se podaci nalaze
- *_data_preprocessing* - funkcija koja uzima samo određene stupce iz skupa podataka nužne za učenje modela, miče prazna polja i poziva funkciju za faktoriziranje podataka
- *_factorize_data* - funkcija koja pretvara tekstualne podatke koji se nalaze u skupu podataka u brojeve kako bi ih algoritam (računalo) moglo raspoznati, [23]
- *get_preprocessed_data* - funkcija koja se poziva dalje u algoritmu kada se treba pristupiti skupu podataka

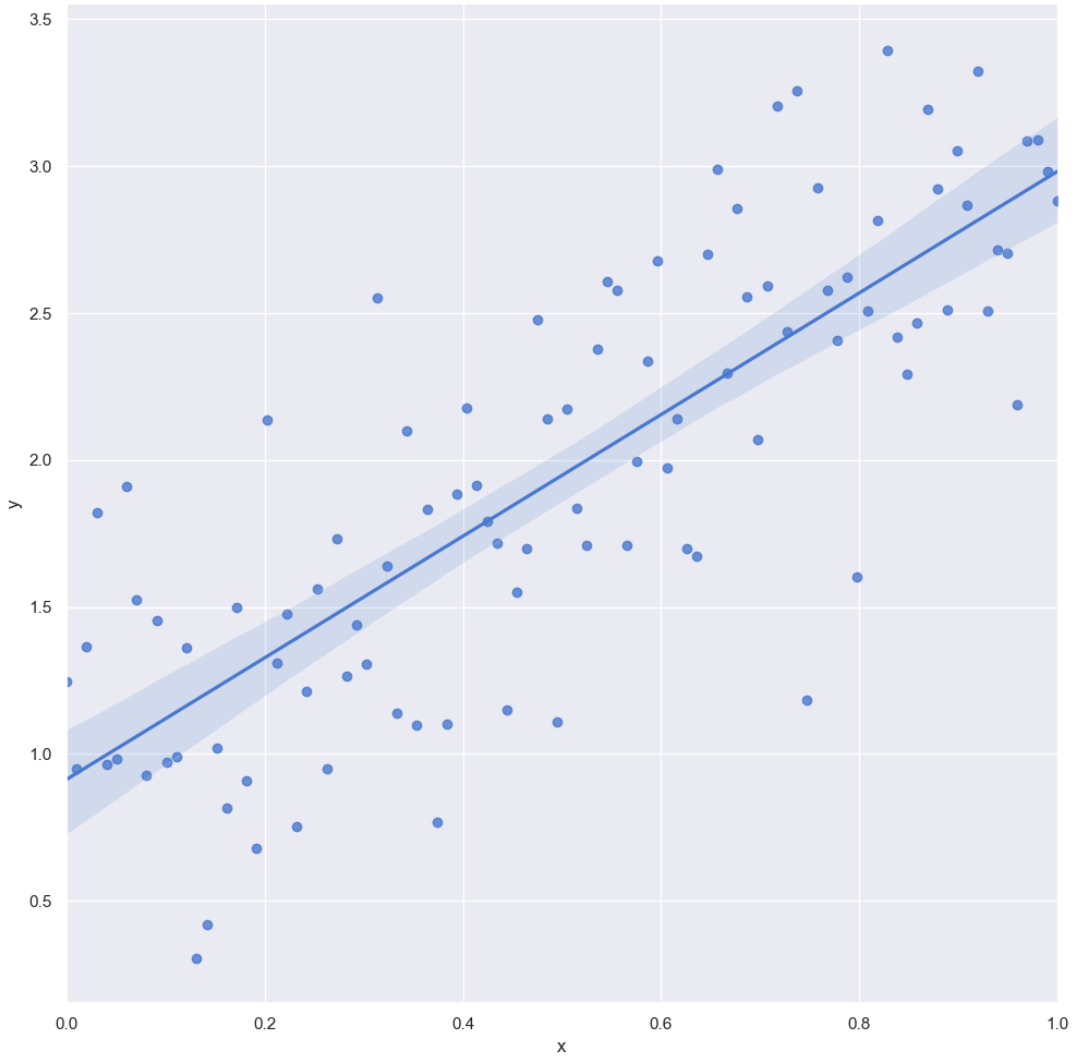
5. Implementacija modela predikcije

Za izradu modela koristilo se 5 različitih vrsta regresija pomoću kojih su dobiveni rezultati. Ponajprije je bilo potrebno podijeliti skup podataka u grupu za učenje i grupu za učenje, gdje je uzeta jedna petina podataka za testiranje, a ostatak za učenje. Separacija podataka se kod stojnog učenja koristi kako bi se lakše procijenila učinkovitost algoritma što će se poslije u radu pokazati pomoću pokazatelja točnosti.

5.1. Korištene regresije

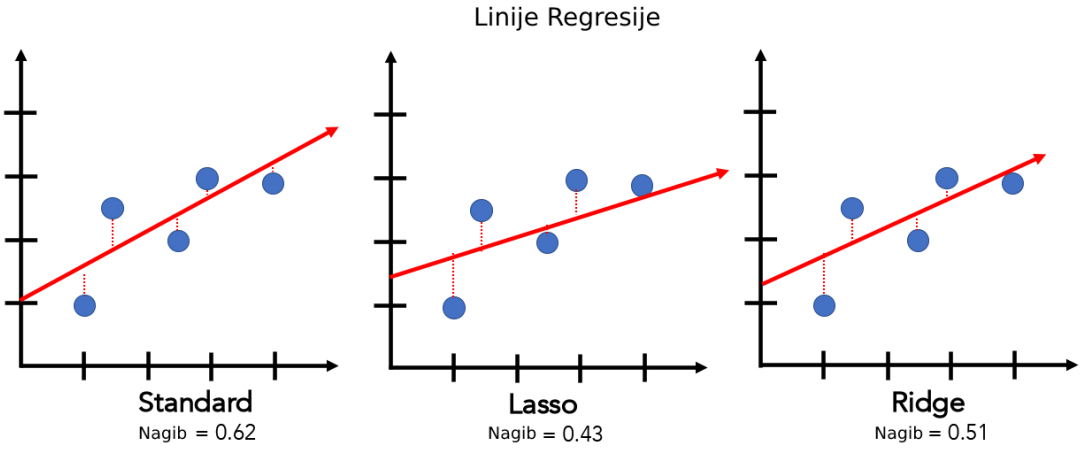
Za učenje modela primijenile su se sljedeće regresije:

Bayesova linearna regresija - To je vrsta uvjetnog modeliranja u kojem se srednja vrijednost jedne varijable opisuje linearnom kombinacijom skupa drugih varijabli, s ciljem dobivanja posteriorne vjerojatnosti koeficijenata regresije (kao i drugih parametara koji opisuju distribuciju) i u konačnici dopušta predviđanje uzorka, ovisno o opaženim vrijednostima regresora, X. Najjednostavnija i najčešće korištena verzija ovog modela je normalni linearni model, [7].



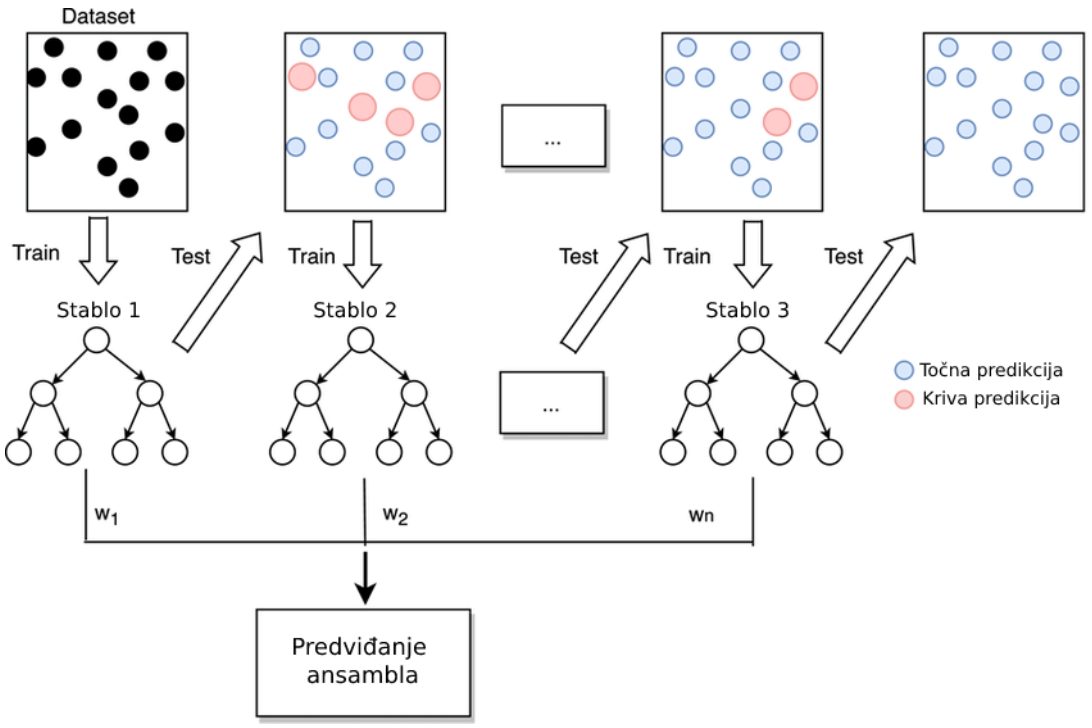
Slika 18: Prikaz Bayesove regresije
Izvor: [24]

Lasso regresija - U statistici i strojnom učenju, Lasso (operator najmanjeg apsolutnog skupljanja i odabira; također Lasso ili LASSO) je metoda regresijske analize koja izvodi odabir varijabli i regularizaciju kako bi se poboljšala točnost predviđanja i interpretabilnost rezultirajućeg statističkog modela. Lasso je izvorno formuliran za modele linearne regresije, [7].



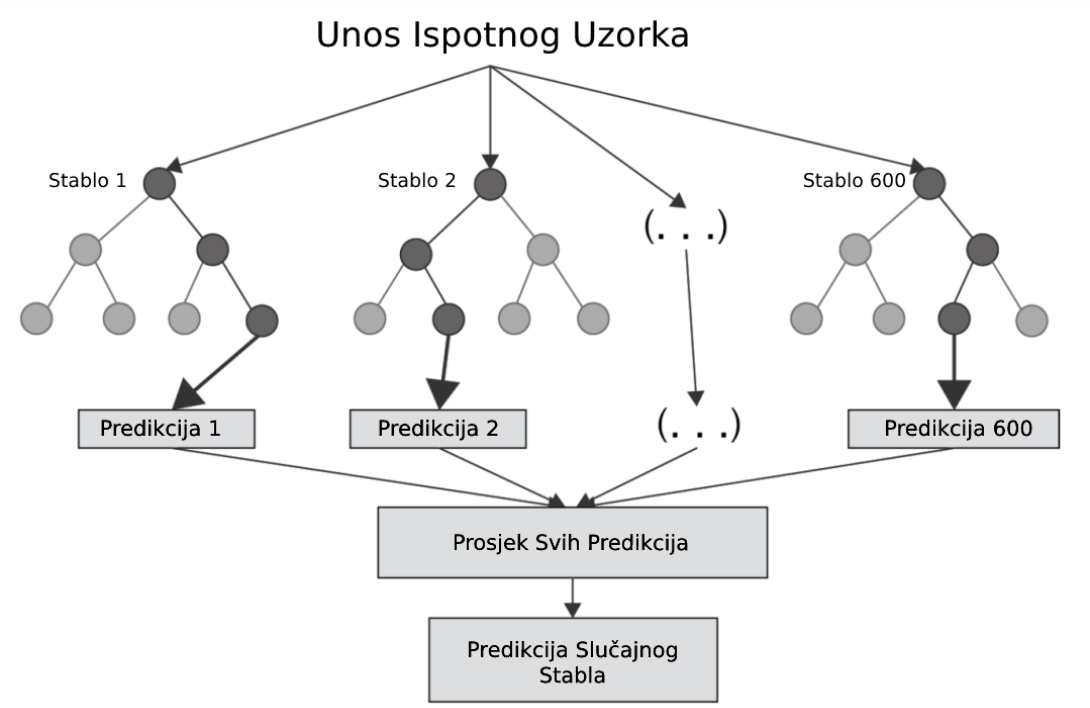
Slika 19: Prikaz odnosa Lasso regresije
Izvor: [25]

Regresija s povećanjem gradijenta - To je tehnika strojnog učenja koja se, između ostalog, koristi u zadacima regresije i klasifikacije. Daje model predviđanja u obliku skupa slabih modela predviđanja, koji su obično stabla odlučivanja. Kada je stablo odlučivanja "slab učenik", rezultirajući algoritam naziva se stablo pojačano gradijentom; i obično nadmašuje slučajnu šumu. Gradijentno pojačani model stabala izgrađen je u fazamas, ali generalizira druge metode dopuštajući optimizaciju proizvoljne diferencijabilne funkcije gubitka, [7].



Slika 20: Prikaz rada regresije s povećanjem gradijenta
Izvor: [26]

Nasumična šuma - Također zvana šuma nasumičnog odlučivanja je skupna metoda učenja za klasifikaciju, regresiju i druge zadatke koja funkcionira na način da se konstruiran veliki broj stabala odlučivanja tijekom učenja modela. Za zadatke klasifikacije, izlaz nasumične šume je klasa koju je odabrala većina stabala. Za regresijske zadatke vraća se srednja vrijednost ili prosječno predviđanje pojedinačnih stabala. Šume nasumičnog odlučivanja ispravljaju naviku stabala odlučivanja da se pretjerano prilagođavaju svom skupu za obuku. Nasumične šume općenito nadmašuju stabla odlučivanja, ali njihova je točnost niža od stabala s pojačanim gradijentom. Međutim, karakteristike podataka mogu utjecati na njihovu izvedbu, [7].



Slika 21: Prikaz rada regresije nasumičnog stabla
Izvor: [27]

Linearna regresija - Na kraju je korištena linearna regresija koja je opisana u poglavlju 2.1.1..

5.2. Implementacija na stvarnim podacima

Za implementaciju su se koristili podaci prikupljeni OBD-II uređajem na osobnom vozilu. Radi se o automobilu Opel Insignia u kojem je bio priključen OBD-II uređaj tijekom regularne gradske vožnje i nekoliko dužih putovanja po autocesti. Prikupljeno je sveukupno 14495 seta podataka čija se struktura može vidjeti na slici 22.

Za model potrebna su dva tipa podataka, konstantni dio koji se sastoji od osnovnih informacija o automobilu (broj cilindra, veličina motora, tip goriva i dr.) te promjenjivi dio koji se sastoji od potrošnje goriva. Analizirajući prikupljene podatke informacija o potrošnji goriva nije dostupna zbog ograničenja koje dolazi kod korištenja OBD-II uređaja. Svaki proizvođač može ograditi podatke koji se mogu očitavati OBD-II uređajem zbog čega se na slici 22 može vidjeti u nekim poljima vrijednost -1, što označava nemogućnost prikupljanja podataka.

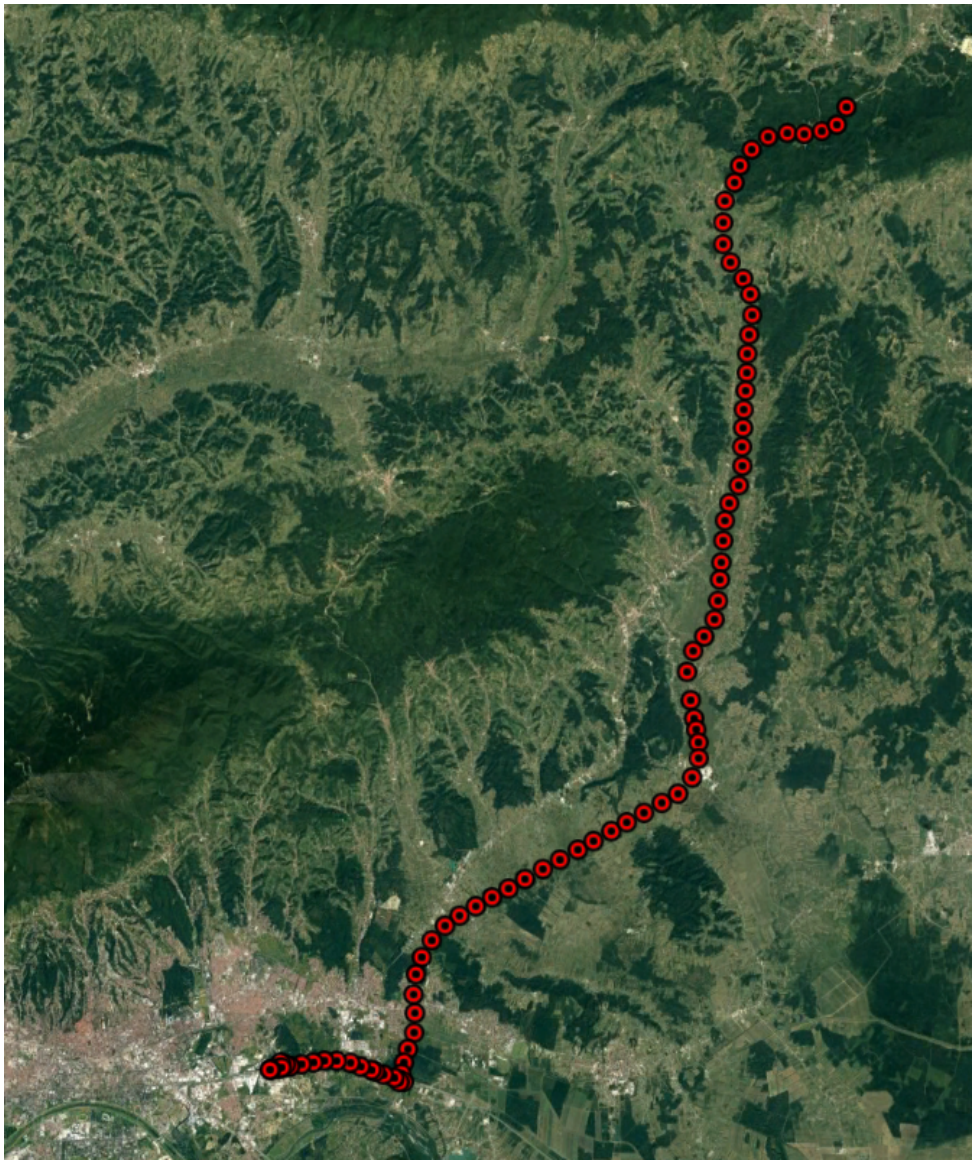
Slika 22: Prikaz prikupljenih podataka na web poslužitelju
Izvor: [16]

Potrošnju goriva u trenutnom vremenu je moguće izračunati ako je poznata brzina vozila (*eng.* Vehicle Speed Sensor, VSS), maseni protok zraka (*eng.* Mass Air Flow, MAF) (što je poznato u setu podataka) te nekoliko konstanta. Prva konstanta je omjer zraka i goriva u motoru. U modernim vozilima s niskim emisijama, odnos zrak/gorivo održava se na konstantnom kemijskom idealnom omjeru od 14,7 g zraka na 1 g benzina. Druga potrebna konstanta je gustoća benzina gdje ona donekle varira ovisno o vrsti goriva i temperaturi okoline, ali s obzirom na točnost prikaza, sljedeća konstanta dobro funkcionira za dizel gorivo: 840 grama po litri. Na kraju se te dvije konstante pomože s brzinom vozila te se maseni protok zraka pomnoži s 3600 (broj sekundi u satu) i dobiva se sljedeća jednadžba:

$$\frac{L}{100km} = \frac{3600 * MAF}{10878 * VSS} \tag{3}$$

Koristeći jednadžbu iznad dobiva se potrošnja goriva što je promjenjivi dio modela. Implementacijom konstantnog i promjenjivog dijela uspješno se dobiva predikcija količine CO2 u određenom vremenu.

Za demonstraciju koristila se ruta prikazana na slici 23 koja se sastoji od 1460 specifičnih podataka kondenziranih u 86 točaka gdje svaka točka sadrži podatke o lokaciji, id točke, brzine vozila, maseni protok zraka, potrošnje goriva i predviđenoj emisiji CO2.



Slika 23: Prikaz korištene rute

6. Rezultati

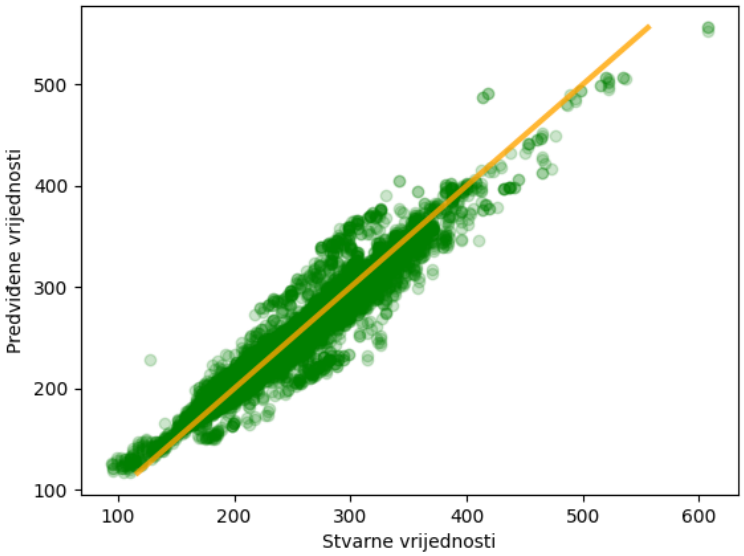
Rezultati diplomskog rada sastoje se od dva dijela, rezultata regresija, gdje se prikazuju usporedbe predviđenih i stvarnih vrijednosti te se pomoću pet različitih pokazatelja točnosti odabira najbolja regresija. Drugi dio sastoji se od prikaza uspješne implementacije najboljeg modela na privatno prikupljenim podacima OBD-II uređaja. Opisane su specifikacije cijele rute, te zbog velike količine podataka prikupljenih OBD-II uređajem za ljepši prikaz izabrao se manji dio rute za prikaz promjene podataka kroz vrijeme.

6.1. Rezultati regresija

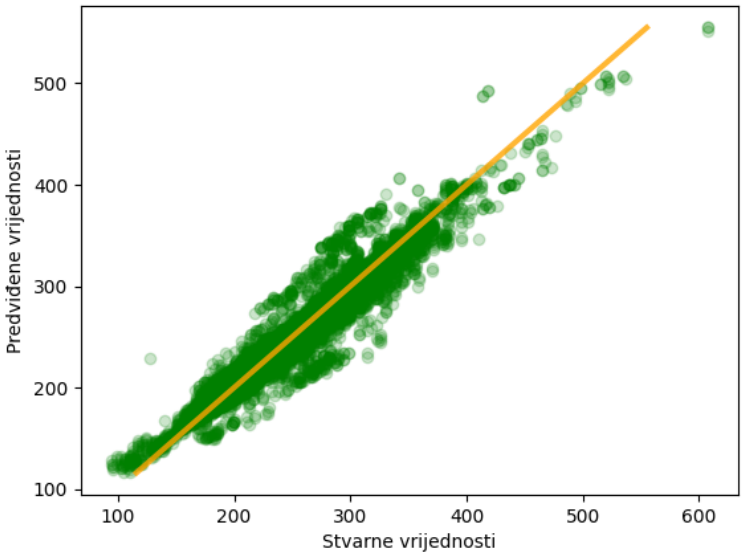
Kako bi se mogla ispitati točnost korištenog modela koristili su se pokazatelji točnosti modela koje nudi paket Sklearn u Pythonu. Koristili su se sljedeći pokazatelji točnosti:

- Ocjena točnosti
- Objašnjeni rezultat varijance
- Srednja kvadratna logaritamska pogreška
- R2 rezultat
- D2 rezultat apsolutne pogreške

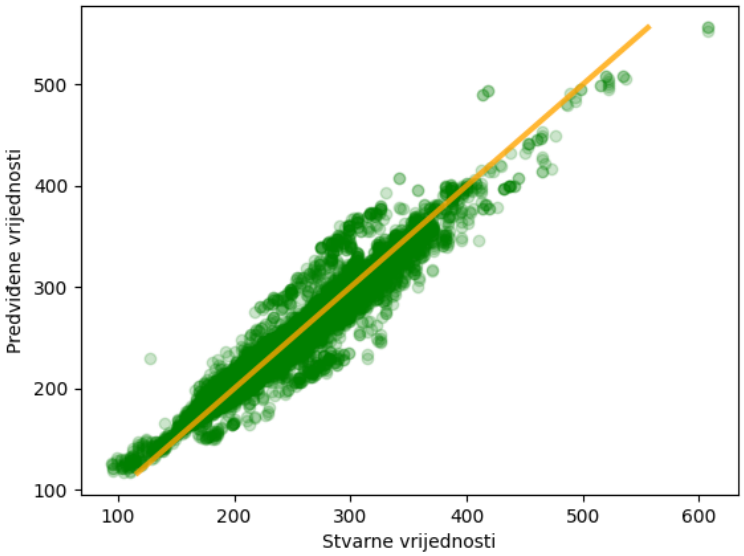
Glavni način rada svih pokazatelja točnosti je da uzimaju točnu vrijednost iz skupa podataka, što je u radu bio stupac emisija CO2 koji je označavao stvarno izmjerene vrijednosti emisija od proizvođača, te ih uspoređuje s predviđenim vrijednostima modela predikcije. Pokazatelji točnosti su kao ulazne parametre koristili dvije varijable (točna i predviđena vrijednost) gdje se svaka od njih sastojala od 8288 podataka. Izgled usporedbe predviđenih i stvarnih vrijednosti može se vidjeti na slikama ispod.



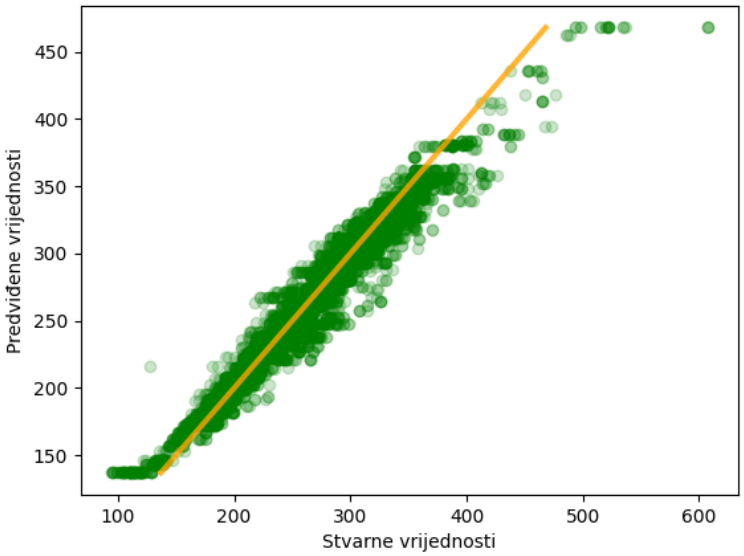
Slika 24: Rezultati Linearne regresije



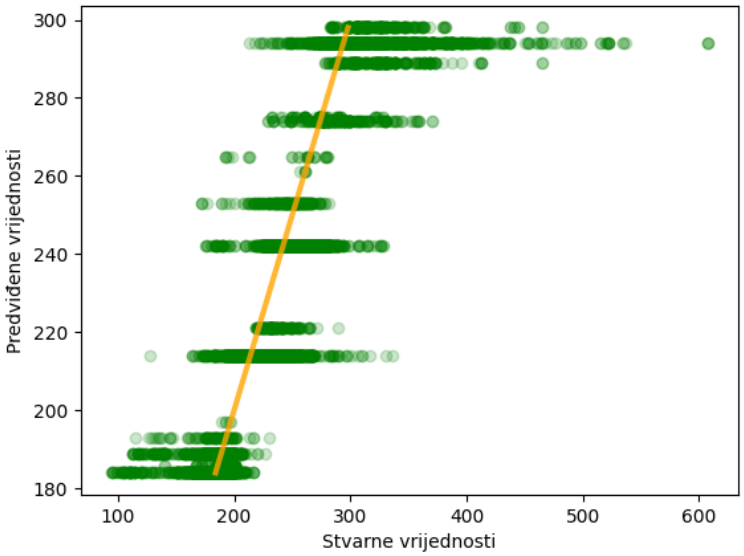
Slika 25: Rezultati Bayesove regresije



Slika 26: Rezultati Lasso regresije



Slika 27: Rezultati Regresije s povećanjem gradijenata



Slika 28: Rezultati Regresije nasumične šume

Ocjena točnosti: To je kriterij koji se koristi za izračunavanje točnosti ili broja točnih predviđanja za rezultat, a koji matematički predstavlja omjer zbroja svih predviđenih stvarno pozitivnih (TP) i stvarno negativnih rezultata (TN), a može se prikazati jednadžbom:

$$Accuracy = \sum \frac{TP}{TN} \tag{4}$$

U rezultatima analize najbolji rezultat koji se može postići je 0, a najlošiji je 1, [28].

Rezultat objašnjene varijance: U statistici, objašnjena varijacija mjeri udio u kojem matematički model uzima u obzir varijaciju (disperziju) danog skupa podataka. Komplementarni dio ukupne varijacije naziva se neobjašnjiva ili rezidualna varijacija. Objašnjena varijanca može se označiti s r^2. U analizi varijance se naziva eta kvadrat, a u regresijskoj analizi naziva se koeficijent determinacije (R^2). Ta tri pojma su u osnovi sinonimi, osim što R2 pretpostavlja da su promjene u ovisnoj varijabli posljedica linearnog odnosa s nezavisnom varijablom, a eta kvadrat nema ovu temeljnu pretpostavku. Rezultat objašnjene varijance može se prikazati jednadžbom:

$$r^2 = R^2 = \eta^2 \tag{5}$$

U rezultatima analize najbolji rezultat koji se može postići je 1, a najlošiji je 0, [29].

Srednja kvadratna logaritamska pogreška: Može se tumačiti kao mjera omjera između stvarnih i predviđenih vrijednosti. Način rada algoritma je da će tretirati male razlike između malih stvarnih i predviđenih vrijednosti približno isto kao i velike razlike između velikih stvarnih i predviđenih vrijednosti. Koristi sljedeću jednadžbu:

$$\text{MSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2} \tag{6}$$

U jednadžbi *n* je ukupan broj opažanja, *p od i* je predviđena vrijednost, *a od i* je točna vrijednost, te je *log(x)* prirodni logaritam vrijednosti. U rezultatima analize najbolji rezultat koji se može postići je 0, a najlošiji je 1, [30].

R2 rezultat: U statistici, koeficijent determinacije je udio varijacije ovisne varijable koji je predvidljiv iz nezavisne varijable. To je pokazatelj koji se koristi u kontekstu statističkih modela čija je glavna svrha, predviđanje budućih ishoda ili testiranje hipoteza, na temelju drugih povezanih informacija. Omogućuje mjeru koliko dobro model replicira promatrane ishode, na temelju udjela ukupne varijacije ishoda objašnjenih modelom, a u pozadini koristi jednadžbu:

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{7}$$

U rezultatima analize najbolji rezultat koji se može postići je 1, a najlošiji je 0, [31].

D2 rezultat apsolutne pogreške: U statistici, D2 apsolutna pogreška je mjera pogrešaka između uparenih opažanja koja izražavaju isti fenomen. Primjeri Y u odnosu na X uključuju usporedbe predviđenog u odnosu na promatrano, naknadno vrijeme u odnosu na početno vrijeme i jednu tehniku mjerenja u odnosu na alternativnu tehniku mjerenja. Izračunava se kao zbroj apsolutnih pogrešaka podijeljen s veličinom uzorka, a u pozadini koristi jednadžbu:

$$D^2(y,\hat{y}) = 1 - \frac{\text{dev}(y,\hat{y})}{\text{dev}(y,y_{\text{null}})} \tag{8}$$

U rezultatima analize najbolji rezultat koji se može postići je 1, a najlošiji je 0, [32].

Koristeći navedene pokazatelje točnosti rezultati se mogu očitati u tablici ispod.

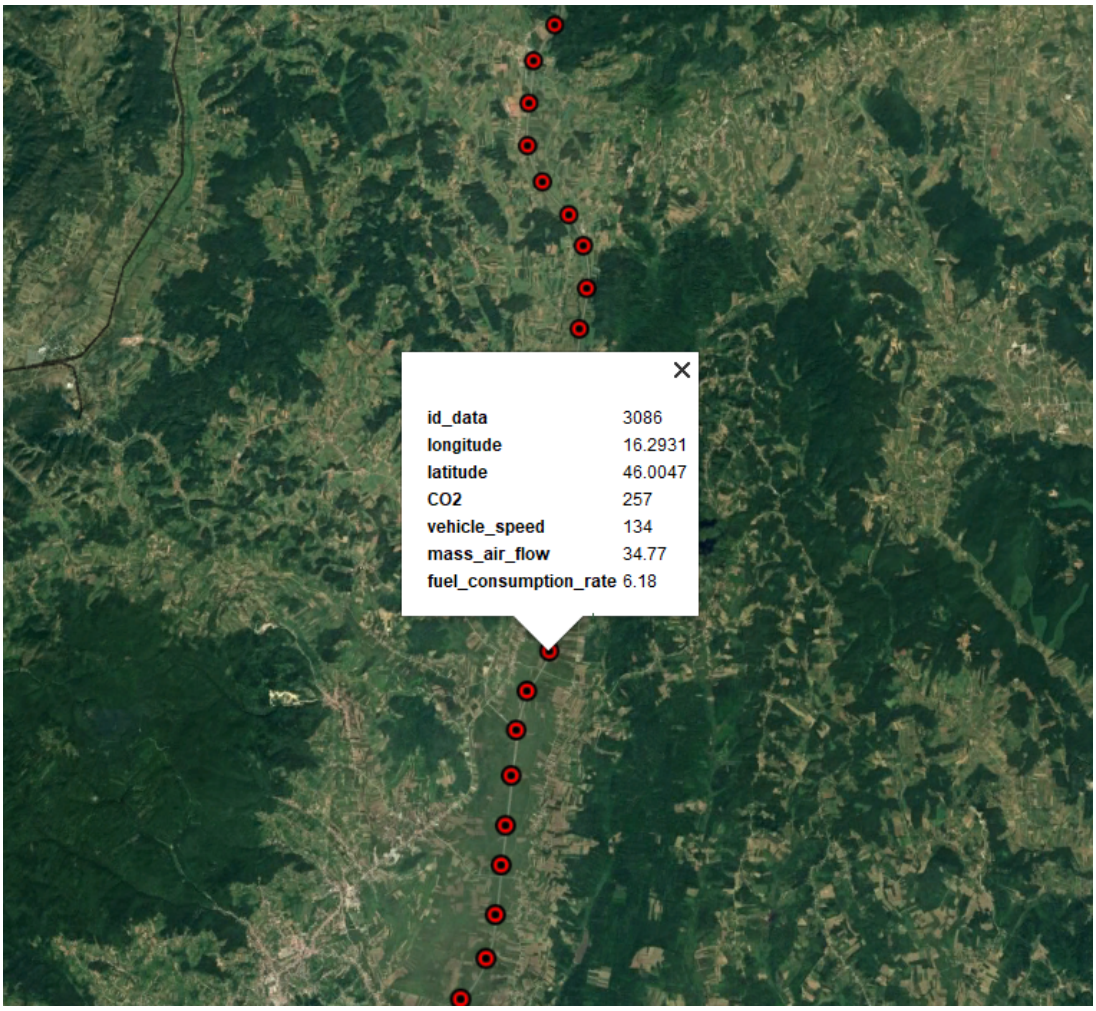
Tablica 4: Rezultati pokazatelja točnosti

	Ocjena točnosti	Objašnjeni rezultat varijance	Srednja kvadratna logaritamska pogreška	R2 rezultat	D2 rezultat apsolutne pogreške
Linearna regresija	0.026	0.915	0.004	0.913	0.740
Bayesova regresija	0.024	0.917	0.004	0.917	0.714
Lasso regresija	0.026	0.916	0.004	0.912	0.738
Regresija s povećanjem gradijenata	0.039	0.953	0.002	0.953	0.809
Regresija nasumične šume	0.043	0.708	0.015	0.689	0.546

Kako bi se osigurali najkvalitetniji rezultati implementacije modela bilo je potrebno odabrati najbolju regresiju za korištenje, te analizirajući dobivene rezultate regresija i pokazatelja točnosti može se primijetiti kako su rezultati Linearne, Bayesove i Lasso regresije dosta slični. Uspoređujući njihove grafove daje se dojam kako su rezultati identični no analizirajući rezultate pokazatelja točnosti u tablici 4 vidi se kako svi pokazatelji imaju malu devijaciju osim srednje kvadratne logaritamske pogreške. Regresija s povećanjem gradijenata i regresija nasumične šume najviše odstupaju, te se mogu smjestiti na dva kraja, gdje uspoređujući ih s drugima regresija nasumične šume ima najlošije rezultate, a regresija s povećanjem gradijenata ukazala se kao najtočnija i koristila se u radu za implementaciju modela na privatnim podacima.

6.2. Prikaz rezultata na privatnim podacima

Koristeći odabranu regresiju na privatnim podacima je obavljena primjena gdje je na određenoj ruti od 86 točaka uspješno predviđena emisija CO2. Na slici 29 se može vidjeti primjer kako je u točki id-a 3086 automobil išao brzinom 134 km/h, imao potrošnju goriva 6.18 L/100 km, maseni protok zraka 34.77 g/s te mu je predviđena emisija CO2 u tom trenutku 257 grama.



Slika 29: Prikaz određene točke

U tablici 6 može očitati kako je ukupna emisija CO2 na prikazanoj ruti iznosila 12.07 kg s duljinom rute od 52.34 km, prosječnom brzinom vozila od 110.05 km/h te s trajanjem vožnje od 30 minuta i 45 sekundi, te je prosječna emisija iznosila 230.67 g/km. Vozilo je potrošilo 16.81 L s prosječnom potrošnjom od 8.8 L/100 km, te je maseni protok zraka na ruti iznosio 62.93 kg s prosjekom od 34.11 g/s.

Tablica 5: Podaci cijele rute

	Ukupno	**Prosjek**	**Minimum**	**Maksimum**
CO2 [g/km]	12.07 [kg]	230.67	17	479
Brzina vozila [km/h]	52.34 [km]	110.05	5	135
Maseni protok zraka [g/s]	62.93 [kg]	34.11	6.94	106.49
Potrošnja goriva [L/100km]	16.81 [L]	8.8	4.13	52.44

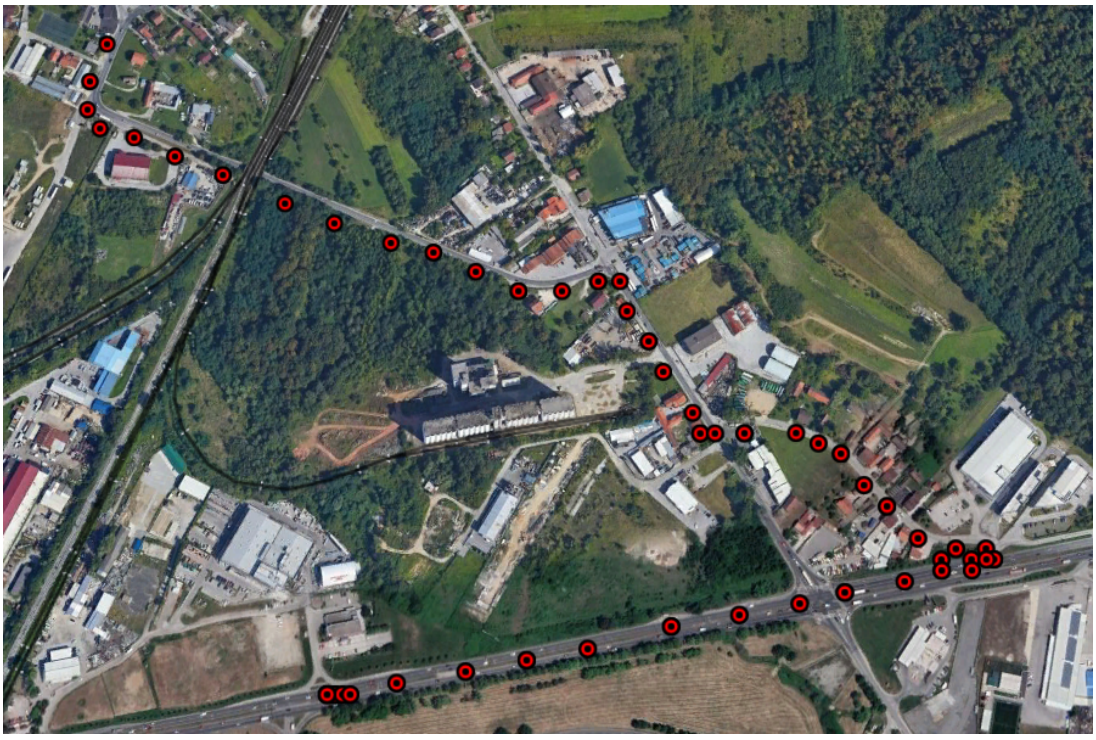
Zbog velike količine podataka za prikaz podataka u vremenu koristila se druga manja ruta koja je prikazana na slici 30 na kojoj je bilo dijelova gdje je vozilo mirovalo, ubrzavalo i vozilo

konstantnom brzinom. Podaci o ruti mogu se očitati u tablici ispod.

Tablica 6: Podaci manje rute

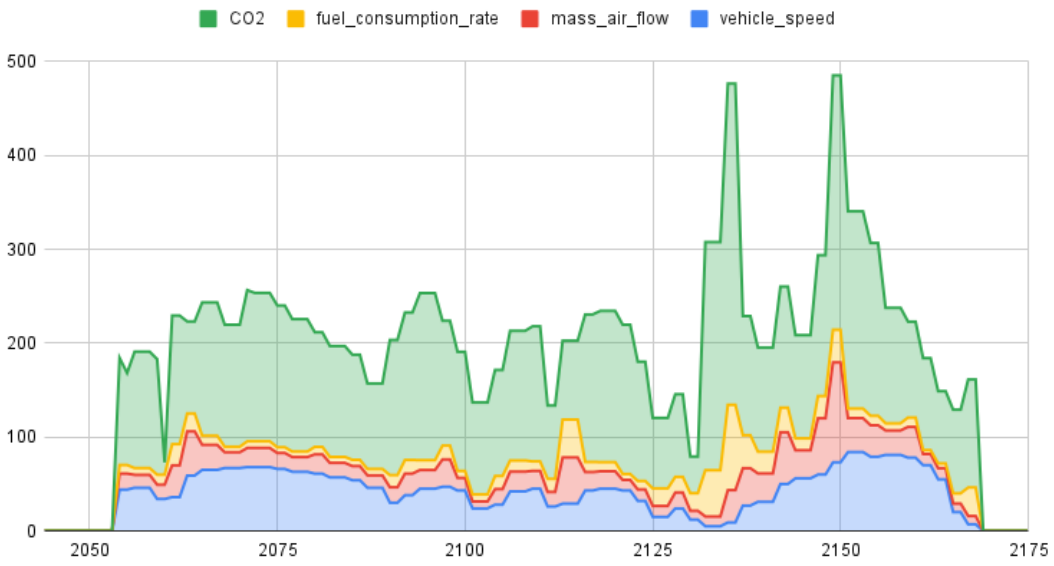
	Ukupno	**Prosjek**	**Minimum**	**Maksimum**
CO2 [g/km]	192 [g]	99.15	0	342
Brzina vozila [km/h]	1.94 [km]	33.26	0	84
Maseni protok zraka [g/s]	1295.75 [g]	17.75	0	106.49
Potrošnja goriva [L/100km]	228 [mL]	11.75	0	90.97

Micanjem duplikata manja ruta se sastoji od 66 točaka i proteže se duljinom od 1.94 km dok vožnja traje 2 minute i 13 sekundi. Ukupna emisija CO2 iznosila je 192g, potrošnja goriva 228 mL s prosječnom brzinom od 33.26 km/h.



Slika 30: Prikaz manje rute

Na slici 31 moguće je očitati ovisnost podataka o rezultatu modela gdje se može izdvojiti kako na drugom dijelu grafa gdje dolazi do ubrzanja vozila model izbacuje veliku količinu emisije CO2. Drugo se može izdvojiti kako kada se vozilo kreće konstantnom brzinom bez naglih ubrzanja ili koćenja rezultati emisije također nemaju nagla odstupanja. Iz grafa se može zaključiti da nagle promjene, primjerice velika potrošnja goriva u kratkom vremenu prilikom ubrzavanja prouzrokuju nagli porast emisija, tako i smanjenje brzine i potrošnje goriva smanjuju količinu emisije. Model također prilikom mirovanja vozila izbacuje da je emisija CO2 jednaka nuli, što u praksi ne bi nužno bila istina jer motor i dalje radi te i dalje ispušta štetne plinove.



Slika 31: Prikaz podatka manje rute u vremenu

7. Zaključak

U ovom diplomskom radu razvijen je simulacijski model predikcije emisije ugljikovog dioksida pomoću podataka OBD-II uređaja i strojnog učenja. Model je izrađen koristeći Python programski jezik koji pojednostavljuje koncept kodiranja programskih simulacija. Skup podataka korišten za učenje modela je javno dostupan skup podataka koji sadrži 8288 podataka o individualnim vozilima. Detaljnije podaci se sastoje od osnovnih parametara vozila kao što su godina proizvodnje, veličina motora, broj cilindra i slično te od CO2 emisije koja je u radu bila korištena za usporedbu s dobivenim rezultatima modela.

Prilikom izrade modela koristilo se pet različitih vrsta regresija, te su točnosti istih bile testirane pomoću pet pokazatelja točnosti, te je najučinkovitija regresija bila regresija s povećanjem gradijenata.

Za implementaciju modela su se koristili privatno prikupljeni podaci gdje je na osobnom vozilu bio priključen OBD-II uređaj gdje su se prikupljali podaci o vožnji nad kojima je vršena implementacija. Podaci su prikazani na ruti koja se sastojala od 1460 specifičnih podataka koji su sadržali potrebne ulazne podatke za korišteni model. Ruta je bila duljine od 52.34 km, te je vožnja trajala 30 minuta i 45 sekundi, a model je za cijelu rutu predvidio da se emitiralo ukupno 12.07 kg emisije CO2 s prosječnom emisijom od 230.67 g/km.

Ovaj diplomski rad i njegov praktični dio mogu se primijeniti u nastavi za demonstraciju izrade i primjene modela strojnog učenja, za analizu emisija ugljikovog dioksida pomoću podataka prikupljenih OBD-II uređajem, za analizu emisija ugljikovog dioksida na određenim rutama nakon vožnje te kao edukacijski materijali budućim ITS stručnjacima pri razvoju sličnih modela strojnog učenja. Navedeno predstavlja izazov i motivaciju za daljnji rad i istraživanja u ovom području s ciljem postizanja još boljih rezultata točnosti modela. Kao prijedlog za daljnji rad predlaže se unapređenje sustava kako bi se uspješno prepoznalo da vozilo u stanju mirovanja i daje emitira ispušne plinove te za kvalitetniju implementaciju mogli bi se prikupiti podaci s više vrsta vozila.

Popis literature

[1] Wallace D. *Environmental policy and industrial innovation: Strategies in Europe, the USA and Japan*. Routledge, 2017.

[2] Chan C. C., The State of the Art of Electric, Hybrid, and Fuel Cell Vehicles. Preuzeto s: https://ieeexplore.ieee.org/document/4168013. [Pristupio: Kolovoz 2022].

[3] Kan Z., Tang L., Kwan M. and Zhang X., Estimating vehicle fuel consumption and emissions using GPS big data. Preuzeto s: https://pubmed.ncbi.nlm.nih.gov/29561813/. [Pristupio: Kolovoz 2022].

[4] Fuel consumption ratings. https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6. [Pristupio: Kolovoz 2022].

[5] Meena G., Sharma D., and Mahris M., 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE). Preuzeto s: http://toc.proceedings.com/54355webtoc.pdf. [Pristupio: Kolovoz 2022].

[6] Machine Learning Models. https://www.javatpoint.com/. [Pristupio: Kolovoz 2022].

[7] Stambaugh R. F., Predictive regressions, Journal of Financial Economics. Preuzeto s: https://www.sciencedirect.com/science/article/abs/pii/S0304405X99000410. [Pristupio: Kolovoz 2022].

[8] Decision tree. https://ai-pool.com/a/s/decision-trees. [Pristupio: Kolovoz 2022].

[9] Thaiparnit S., Chumuang N., and Ketcham M., A comparitive study of clasification liver dysfunction with machine learning. Preuzeto s: https://ieeexplore.ieee.org/document/8692808. [Pristupio: Kolovoz 2022].

[10] G Bonaccorso, Machine learning algorithms. https://books.google.hr/books?hl=en&lr=&id=_-ZDDwAAQBAJ&oi=fnd&pg=PP1&dq=Giuseppe+Bonaccorso.

+Machine+learning+algorithms.+Packt+Publishing+Ltd,+2017&ots=epjAB5DD5E&sig=tLnumHabGmCyDWSHbO3ecVIqOW4&redir_esc=y#v=onepage&q=Giuseppe%20Bonaccorso.%20Machine%20learning%20algorithms.%20Packt%20Publishing%20Ltd%2C%202017&f=false. [Pristupio: Kolovoz 2022].

[11] Watkins C. and Dayan P., Q-learning, Machine learning. Preuzeto s: https://link.springer.com/article/10.1007/BF00992698. [Pristupio: Kolovoz 2022].

[12] Rimpas D., Papadakis A., and Samarakou M., Technologies and Materials for Renewable Energy, Environment and Sustainability. Preuzeto s: https://www.sciencedirect.com/science/article/pii/S2352484719308649. [Pristupio: Kolovoz 2022].

[13] OBD-I. https://www.joom.com/ro/products/5ce29e628b2c3701010bd3b5. [Pristupio: Kolovoz 2022].

[14] OBD-II. http://run-odo-run.com/hyundai. [Pristupio: Kolovoz 2022].

[15] Hilpert h.,Thoroe L. and Schumann M., Real-Time Data Collection for Product Carbon Footprints in Transportation Processes Based on OBD2 and Smartphones. Preuzeto s: https://ieeexplore.ieee.org/document/5718558. [Pristupio: Kolovoz 2022].

[16] Vaiti T., Izrada sustava za online prikupljanje podataka s vozila opremljenih OBD uređajema. Preuzeto s: https://dabar.srce.hr/islandora/object/fpz%3A2095. [Pristupio: Kolovoz 2022].

[17] On-Board Diagnostics (OBD-II) Port. https://slideplayer.com/slide/4502179/14/images/5/Con. [Pristupio: Kolovoz 2022].

[18] Simple Ways To Identify Your Vehicle's OBD2 Protocol. https://www.obdadvisor.com/obd2-protocol-supported-vehicle/. [Pristupio: Kolovoz 2022].

[19] Greenhouse Gas Rating. Preuzeto s: https://www.epa.gov/greenvehicles/greenhouse-gas-rating. [Pristupio: Kolovoz 2022].

[20] Numpy the fundamental package for scientific computing with python. https://numpy.org/. [Pristupio: Kolovoz 2022].

[21] Pandas data analysis and manipulation tool. https://pandas.pydata.org/. [Pristupio: Kolovoz 2022].

[22] Sklearn machine learning in python. https://scikit-learn.org/stable/. [Pristupio: Kolovoz 2022].

[23] Lasso and Ridge regression. https://www.geeksforgeeks.org/python-pandas-series-factorize/. [Pristupio: Kolovoz 2022].

[24] Bayesian Linear Regression Models with PyMC3. https://www.quantstart.com/articles/Bayesian-Linear-Regression-Models-with-PyMC3/. [Pristupio: Kolovoz 2022].

[25] Lasso and Ridge regression. https://thaddeus-segura.com/lasso-ridge/. [Pristupio: Kolovoz 2022].

[26] Flow diagram of gradient boosting machine learning method. https://www.researchgate.net/figure/Flow-diagram-of-gradient-boosting-machine-learning-method-The-ensemble-classifiers_fig1_351542039. [Pristupio: Kolovoz 2022].

[27] Diving into the Deep learning : Random Forest Algorithm. https://www.linkedin.com/pulse/diving-deep-learning-random-forest-algorithm-shubham-gupta. [Pristupio: Kolovoz 2022].

[28] Sklearn accuracy classification score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html. [Pristupio: Kolovoz 2022].

[29] Sklearn explained variance regression score function. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.explained_variance_score.html. [Pristupio: Kolovoz 2022].

[30] Sklearn mean squared logarithmic error regression loss. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_log_error.html. [Pristupio: Kolovoz 2022].

[31] Sklearn r2 regression score function. https://scikit-learn.org/stable/modules/model_evaluation.html#r2-score. [Pristupio: Kolovoz 2022].

[32] Sklearn d2 regression score function. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.d2_tweedie_score.html. [Pristupio: Kolovoz 2022].

Popis slika

Popis tablica

Sveučilište u Zagrebu
Fakultet prometnih znanosti
Vukelićeva 4, 10000 Zagreb

IZJAVA O AKADEMSKOJ ČESTITOSTI I SUGLASNOSTI

Izjavljujem i svojim potpisom potvrđujem da je _____diplomski rad_____
(*vrsta rada*)
isključivo rezultat mojega vlastitog rada koji se temelji na mojim istraživanjima i oslanja se na objavljenu literaturu, a što pokazuju upotrijebljene bilješke i bibliografija. Izjavljujem da nijedan dio rada nije napisan na nedopušten način, odnosno da je prepisan iz necitiranog rada te da nijedan dio rada ne krši bilo čija autorska prava. Izjavljujem, također, da nijedan dio rada nije iskorišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

Svojim potpisom potvrđujem i dajem suglasnost za javnu objavu završnog/diplomskog rada pod naslovom PREDIKCIJA EMISIJA ŠTETNIH PLINOVA VOZILA KORIŠTENJEM ALGORITAMA STROJNOG UČENJA S PODATCIMA PRIKUPLJENIH 0BD2 UREĐAJEM , u Nacionalni repozitorij završnih i diplomskih radova ZIR.

Student/ica:

U Zagrebu, 09.09.2022 _____

(ime i prezime, *potpis*)