Analiza podataka, njihovih veza i mogućnost smanjena dimenzionalnosti i primena algoritma mašinskog učenja nad skupom podataka

Opis problema

Problem koji se rešava je analiza podataka potrošnje goriva i emisije CO_2. Ovaj skup podataka sadrži informacije o specifikacijama vozila, potrošnji goriva i emisiji CO_2, prikupljene radi analize ekološkog uticaja vozila i predviđanja njihove emisije CO_2 korišćenjem regresionih modela.

Emisija CO_2 igra ključnu ulogu u analizi uticaja vozila na životnu sredinu, jer direktno doprinosi globalnom zagrevanju i klimatskim promenama. Kako bi se ublažili ovi efekti, mnoge zemlje uvode stroge propise koji zahtevaju smanjenje emisije štetnih gasova, čime se automobilske kompanije podstiču na razvoj ekološki prihvatljivijih vozila. Pored ekoloških razloga, emisija CO_2 ima i ekonomski značaj. Potrošnja goriva, koja je usko povezana sa emisijom CO_2, utiče na troškove vozača i zavisnost od fosilnih goriva. Smanjenjem emisije ne samo da se štiti životna sredina, već se i optimizuju troškovi potrošnje goriva. Iako CO_2 nije toksičan, njegov porast u atmosferi doprinosi pogoršanju kvaliteta vazduha u urbanim sredinama, što može negativno uticati na zdravlje ljudi.

Zbog toga se u automobilskoj industriji sve više ulaže u razvoj održivih tehnologija, poput električnih i hibridnih vozila, kao i u optimizaciju motora i poboljšanje efikasnosti goriva. Praćenje i predviđanje emisije CO_2 nije važno samo za regulative i ekonomiju, već i za dugoročnu održivost i razvoj tehnologija koje će smanjiti negativan uticaj na životnu sredinu.

Skup podataka

Podaci koji će biti korišćeni u ovom projektu preuzeti su sa:
https://www.kaggle.com/datasets/brsahan/vehicle-co2-emissions-dataset/data

Zavistan podataka ovog skupa je emisija CO_2 izražena u g/km. Promenljive koje utiču na emisiju CO_2 i koriste se za predikciju:

- **Fuel Consumption City (L/100 km)** – Potrošnja goriva u gradskim uslovima, izražena u litrima na 100 km. Veća potrošnja obično znači veću emisiju CO_2, jer vozilo troši više goriva pri nižim brzinama i čestim zaustavljanjima.
- **Fuel Consumption Hwy (L/100 km)** – Potrošnja goriva na auto-putu, izražena u litrima na 100 km. Na većim brzinama i konstantnoj vožnji potrošnja može biti niža nego u gradu, što može smanjiti emisiju CO_2.
- **Fuel Consumption Comb (L/100 km)** – Kombinovana potrošnja goriva, koja predstavlja prosečnu potrošnju u kombinovanim uslovima vožnje (grad + auto-put). Koristi se kao opšti pokazatelj efikasnosti potrošnje goriva.
- **Engine Size (L)** – Zapremina motora, izražena u litrima. Veći motori obično troše više goriva i emituju više CO_2, ali efikasnost može zavisiti od tehnologije motora i tipa goriva.
- **Cylinders** – Broj cilindara u motoru. Veći broj cilindara obično znači snažniji motor i veću potrošnju goriva, što može rezultirati većom emisijom CO_2.
- **Fuel Type** – Tip goriva koji vozilo koristi. Različiti tipovi goriva sagorevaju na različite načine i emituju različite količine CO_2. Na primer:
    - **X** – Regularni benzin
    - **Z** – Premium benzin
    - **D** – Dizel
    - **E** – Etanol (E85)
    - **N** – Prirodni gas (CNG)

Algoritmi

Smanjenje dimenzionalnosti treba odraditi preko selekcije karakteristika. Koristiti Pearsonovu korelaciju da se odredi povezanost između karakteristika i zavisne promenljive (CO_2 emisija). Pearsonova korelacija meri linearni odnos između dve promenljive i ako neka promenljiva ima vrlo nisku korelaciju sa zavisnom promenljivom ona se može isključiti jer ne doprinosi preciznosti predikcije.

Odabir modela izvšiti testiranjem performansi više modela regresije. Testirati nad sledećim modelima:

Linear Regression – osnovni model koji procenjuje linearne odnose između varijabli

Lasso Regression – verzija linearne regresije sa L1 regularizacijom koja eliminiše manje značajne karakteristike

Ridge Regression – linearni model sa L2 regularizacijom koji smanjuje problem prekomernog uklapanja (overfitting)

Decision Tree – drvo odlučivanja koje segmentira podatke na osnovu pravila.

Random Forest – ansambl metoda koja koristi više stabala odlučivanja radi bolje generalizacije.

Gradient Boosting – model koji iterativno poboljšava predikcije kombinujući jednostavna stabla odlučivanja.

XGBoost – optimizovana verzija gradient boosting-a, poznata po visokoj preciznosti.

Koristiti **K-means klasterizaciju** u projektu kako bi se vozila grupisala u odnosu na njihove karakteristike. Jedan od ključnih izazova kod K-Means klasterizacije je izbor optimalne vrednosti za **K**. Zato treba primeniti **Elbow metod**, koji analizira promenu vrednosti **intra-klasterske varijance (WCSS - Within-Cluster Sum of Squares)** kako bi se pronašao optimalan broj klastera.

Tehnologije

Programski jezik koji će se koristiti je Python. Izabran je zbog svojih mnogobrojnih biblioteka koje omogućavaju i olakšavaju implementaciju projekta. Neke od tih biblioteka su pandas, sklearn, numpy i matplotlib. Pandas omogućava da se podaci učitaju i da se formira data frame. Sklearn omogućava treniranje modela koji će se koristiti za predikciju i za klasterizaciju. Numpy za rad sa numeričkim izrazima i matplotlib za vizuelnu analizu podataka. Okruženje koje će se koristiti za izradu projekta je Visual Studio Code.

Cilj

Cilj projekta je analiza podataka kroz pronalaženje veza nezavisnih sa zavisnim podacima, donošenje zaključaka na osnovu tih veza i vizuelni prikaz najznačajnijih svojstava koji utiču na predikciju. Na osnovu dobijenog vizuelnog prikaza odabrati najznačajnija svojstva i koristiti ih za treniranje modela. Testirati više različitih modela regresije koji su opisani u delu *Algoritmi*, prikazati rezultate testiranja i izabrati onaj sa najboljim performansama. Potrebno je kreirati model i utvrditi tačnost njegove predikcije. Pre primene klasterizacije pronaći optimalan broj klastera primenom Elbow metode. Aplikacija će posedovati UI preko kog će biti omogućen unos svojstava od značaja na osnovu kojih će se vršiti predikcija CO_2 i grupisanje vozila. Korisniku će biti omogućen vizuelni prikaz analize podataka.