# Predviđanje iznosa taxi vožnji u New Yorku pomoću pristupa strojnog i dubokog učenja

**Skladišta i rudarenje podataka**

**Moira Čekada**

# Uvod

U 2019. godini, New York je zabilježio preko 103 milijuna taksi vožnji, što je stvorilo opsežan skup podataka koji odražava složene obrasce urbane mobilnosti. Usluge prijevoza godišnje doprinose gradskom gospodarstvu s oko 15 milijardi dolara, zbog čega je optimizacija predviđanja cijena vožnje postala ključna za poboljšanje operativne učinkovitosti i korisničkog iskustva. Nedavne studije iz područja urbane znanosti o podacima (Zhang i sur., 2020.; Liu & Chen, 2021.) istaknule su stratešku važnost prediktivne analitike u različitim područjima poput otkrivanja prijevara, dinamičkog određivanja cijena i predviđanja potražnje. S obzirom na opseg i bogatstvo podataka koji obuhvaćaju varijable kao što su udaljenost vožnje, lokacije preuzimanja i ostavljanja putnika, doba dana i broj putnika, ova studija ima za cilj iskoristiti pristupe temeljene na podacima kako bi se poboljšala točnost procjene cijena vožnje. Takvi napori usklađeni su s globalnim trendovima u razvoju pametnih gradova, gdje je korištenje podataka o prijevozu u stvarnom vremenu ključno za održivo urbano planiranje i inteligentna rješenja za mobilnost.

### Opis problema

Točno predviđanje cijene taksi vožnje i dalje je značajan izazov u urbanim transportnim sustavima, posebno u dinamičnim okruženjima poput New Yorka, gdje faktori kao što su

prometne gužve, doba dana i udaljenost vožnje neprestano variraju. Netočne procjene cijena mogu dovesti do nezadovoljstva putnika, gubitka prihoda te neučinkovitosti u strategijama otpreme i određivanja cijena. U stvarnim scenarijima, i vozači i putnici imaju koristi od poznavanja očekivane cijene unaprijed; vozači mogu optimizirati odluke o ruti, dok putnici mogu bolje upravljati troškovima i izbjeći prijevare. Unatoč obilju povijesnih podataka o vožnjama koji se prikupljaju godišnje, mnogi sustavi za određivanje cijena i dalje se oslanjaju na statične cjenike koji ne odražavaju uvjete u stvarnom vremenu. Stoga postoji hitna potreba za prediktivnim okvirom koji može pouzdano procijeniti cijene taksi vožnji koristeći povijesne i kontekstualne podatke, omogućujući time pametnije odluke o urbanoj mobilnosti za gradske planere, platforme za prijevoz i putnike.

Ciljevi

1. Predvidjeti iznose taksi vožnji na temelju povijesnih podataka o vožnjama iz New Yorka.
2. Klasificirati taksi vožnje u unaprijed definirane cjenovne kategorije radi lakše analize i donošenja odluka.

3. Implementirati i ocijeniti različite modele strojnog učenja za točnost procjene cijena vožnje.
4. Istražiti i analizirati ključne čimbenike koji utječu na varijabilnost cijena taksi vožnji.
5. Procijeniti performanse modela koristeći odgovarajuće regresijske i klasifikacijske metrike.
6. Usporediti tradicionalne, ansambl i pristupe dubokog učenja na stvarnom skupu podataka.
7. Pružiti uvide i preporuke za poboljšanje sustava za predviđanje cijena vožnji na platformama za urbanu mobilnost.

## Pregled podataka

Ova studija koristi skup podataka Komisije za taksije i limuzine New Yorka (TLC), s posebnim fokusom na zapise o vožnjama žutih taksija za 2019. godinu, koji se sastoje od 12 SQLite datoteka – po jedna za svaki mjesec. Te datoteke zajedno sadrže preko 103 milijuna zapisa o vožnjama, nudeći sveobuhvatan uvid u taksi operacije u jednoj od najprometnijih metropolitanskih prometnih mreža na svijetu. Skup podataka, preuzet s Kagglea, pruža detaljne podatke na razini vožnje, uključujući vremena preuzimanja i ostavljanja putnika, udaljenost vožnje, komponente cijene, vrste plaćanja i lokacijske zone. Ovaj opsežan i stvaran skup podataka omogućuje robusnu analizu i modeliranje predviđanja cijena taksi vožnji, odražavajući dinamiku urbanog prijevoza u New Yorku.

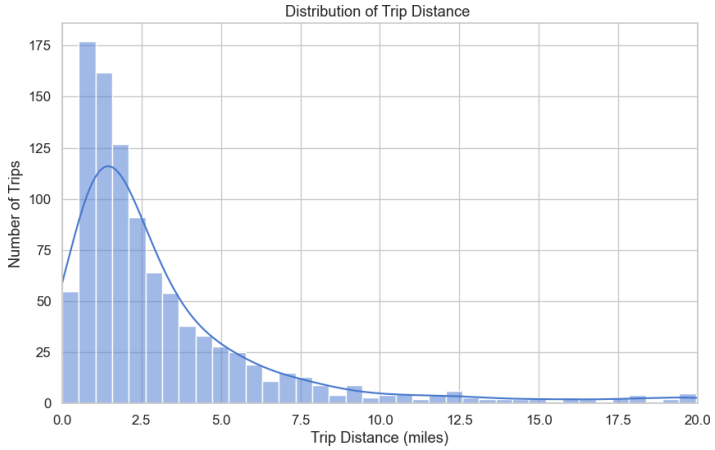
## Skup podataka o vožnjama žutih taksija u NYC-u (2019.)

| Field Name | Description |
|---|---|
| `VendorID` | Code for TPEP provider: 1 = Creative Mobile Technologies, 2 = VeriFone Inc. |
| `tpep_pickup_datetime` | Date and time when the meter was turned on (trip started). |
| `tpep_dropoff_datetime` | Date and time when the meter was turned off (trip ended). |
| `passenger_count` | Number of passengers (entered by the driver). |
| `trip_distance` | Distance traveled in miles (from taximeter). |
| `PULocationID` | Taxi Zone ID where trip began. |
| `DOLocationID` | Taxi Zone ID where trip ended. |
| `RateCodeID` | Pricing code (e.g., 1 = standard, 2 = JFK, 5 = negotiated fare, etc.). |
| `store_and_fwd_flag` | Y = stored before sending; N = sent in real time. |
| `payment_type` | Payment method: 1 = Credit card, 2 = Cash, 3 = No charge, etc. |

| | |
|---|---|
| fare_amount | Base fare calculated by time and distance. |
| extra | Surcharges like rush hour ($0.50–$1.00). |
| MTA_tax | Fixed $0.50 MTA tax. |
| improvement_surcharge | $0.30 fee added at the start of every trip (since 2015). |
| tip_amount | Tip provided (only for credit card payments). |
| tolls_amount | Sum of tolls paid during the trip. |
| total_amount | Final trip cost (excluding cash tips). |

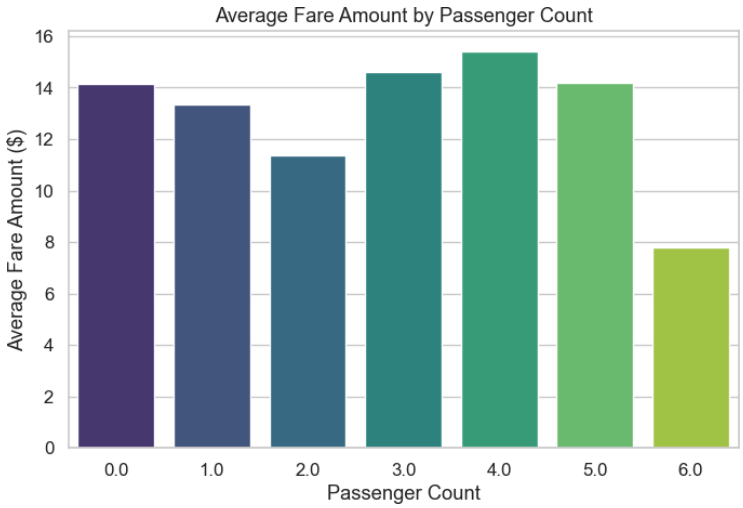
# EDA PROCESS

**Fig 1: Distribution of Trip Distance**



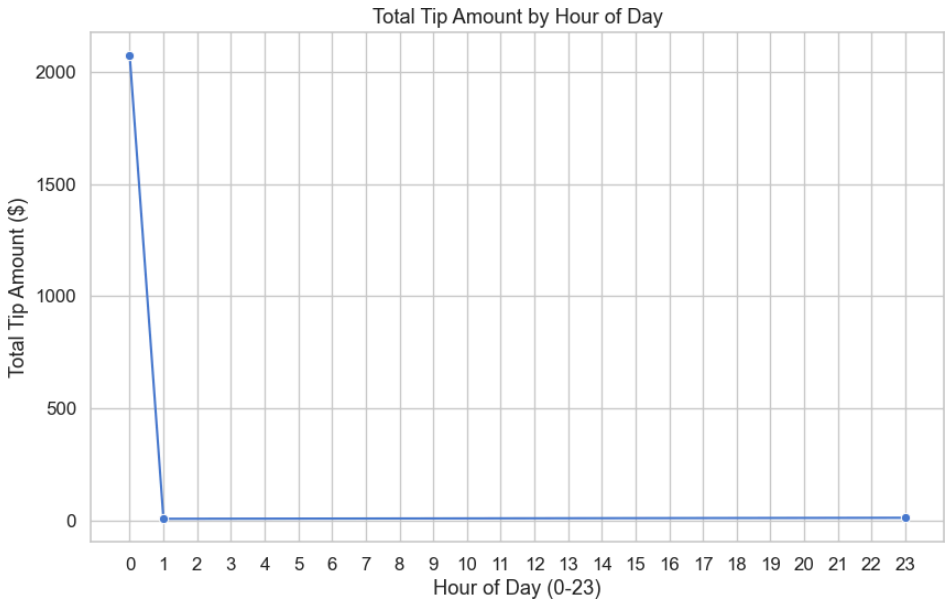

Histogram pokazuje desno iskošenu distribuciju, s većinom vožnji grupiranih oko 1 do 1,5 milja. Kako se udaljenost povećava, broj vožnji naglo opada, formirajući dugi rep prema 20 milja. To ukazuje da su kratke vožnje daleko češće od dugih.

**Fig 2 : Average Fare Amount by Passenger Count**



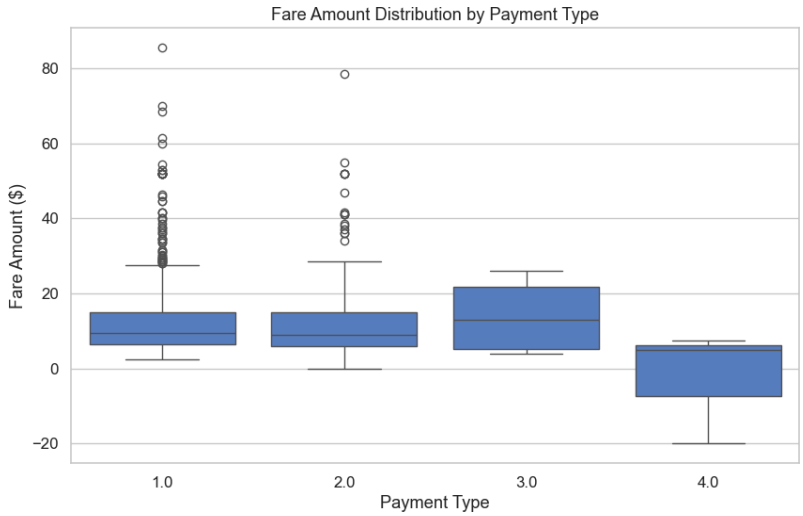

Grafikon pokazuje da prosječna cijena vožnje raste s brojem putnika, dostižući vrhunac s 4 putnika na oko 15,50 USD. Zatim se smanjuje za 5 putnika na oko 14 USD i dodatno pada za 6 putnika na otprilike 7,50 USD, što ukazuje na moguće prilagodbe cijena ili nedosljednosti u podacima.

**Fig 3 : Total Tip Amount by Hour of Day**



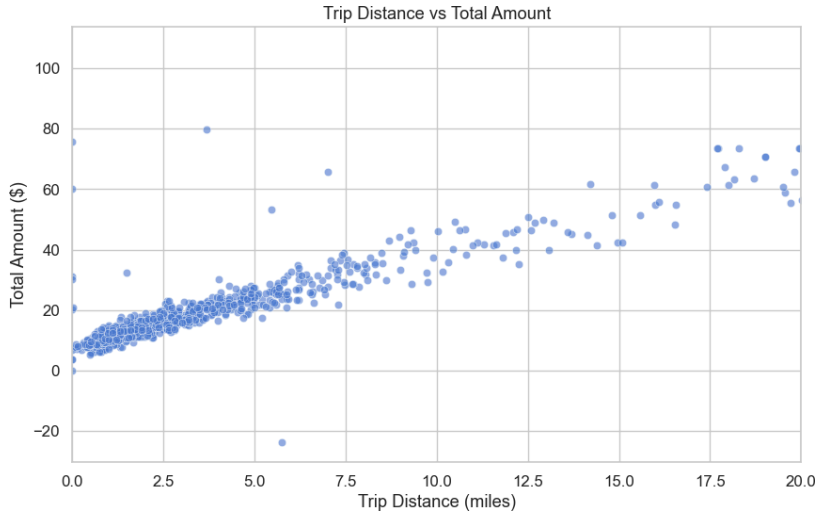

Grafikon pokazuje oštar skok u ukupnim napojnicama u ponoć (preko 2000 USD), nakon čega slijede vrijednosti blizu nule za sve ostale sate. To sugerira da su napojnice snažno koncentrirane u 0. satu, što moguće ukazuje na anomaliju u podacima ili skupni unos na početku dana.

**Fig 4: Fare Amount Distribution by Payment Type**



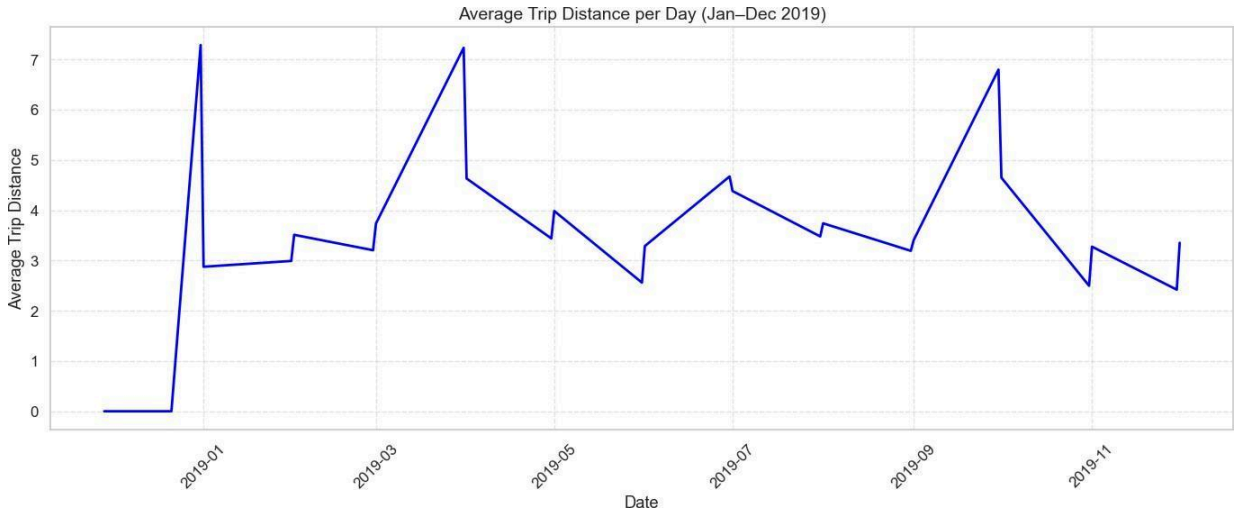

Vrste plaćanja 1.0 i 2.0 imaju slične medijane cijena oko 10–12 USD s nekoliko visokih odstupanja. Tip 3.0 pokazuje viši medijan cijene blizu 15–20 USD i manju varijabilnost. Tip 4.0 ima najniže cijene, uključujući negativne vrijednosti, što vjerojatno ukazuje na povrate novca ili prilagodbe.

**Fig 5: Trip Distance vs Total**




Dijagram raspršenja (scatter plot) pokazuje jaku pozitivnu korelaciju između udaljenosti vožnje i ukupnog iznosa. Većina točaka slijedi uzlazni trend, gdje duže vožnje općenito koštaju više. Pojavljuju se neka odstupanja, uključujući visoke cijene za kratke vožnje i negativan iznos na oko 5,5 milja, što vjerojatno ukazuje na povrat novca.

**Fig 6 : Average Trip Distance per Day**




Grafikon pokazuje fluktuirajuće prosječne udaljenosti vožnji tijekom 2019. godine. Vrhunci oko ožujka-travnja, srpnja i rujna dosežu oko 7 milja, dok padovi u svibnju, kolovozu i na kraju godine padaju na oko 2,5–3 milje, što ukazuje na moguće sezonske ili tjedne trendove.

## Metodologija

Ova faza uključuje višestupanjski pristup predviđanju iznosa taksi vožnji i analizi obrazaca vožnji koristeći mješavinu tradicionalnih algoritama strojnog učenja, ansambl modela, tehnika dubokog učenja i metoda nenadziranog grupiranja. Implementacija kombinira modele izgrađene od nule s Pythonom (NumPy) i uspostavljenim bibliotekama (Scikit-learn, TensorFlow, itd.) kako bi se osiguralo i teorijsko razumijevanje i praktična izvedba.

### K-najbližih susjeda (kNN)

kNN algoritam implementiran je pomoću NumPyja, izračunavajući udaljenosti između točaka podataka metrikom euklidske udaljenosti. Ključne funkcije uključuju računanje parnih udaljenosti i odabir k najbližih susjeda, gdje k varira od 3 do 15 radi optimizacije veličine susjedstva. Značajke kao što su udaljenost vožnje, broj putnika i vrijeme preuzimanja standardizirane su prije izračuna udaljenosti. Predviđanje je napravljeno prosjekom ciljnih vrijednosti (npr. iznos vožnje) najbližih susjeda.

**Distance Metric (Euclidean Distance)**

For two points x=(x1,x2,...,xn) and y=(y1,y2,...,yn), the Euclidean distance is calculated as:

**Modeli nadziranog strojnog učenja**

Razvijeno je više modela nadziranog učenja pomoću biblioteke Scikit-learn, uključujući linearnu regresiju, regresiju stabla odlučivanja i regresiju potpornih vektora (SVR). Linearna regresija modelirala je odnos između ulaznih značajki i cilja kao linearnu jednadžbu. Stabla odlučivanja rekurzivno su dijelila prostor značajki kako bi se minimizirala pogreška predviđanja, koristeći maksimalnu dubinu postavljenu između 5 i 15. SVR je koristio jezgru radijalne bazne funkcije s hiperparametrima C i gama podešenim unutar raspona [0.1, 10], odnosno [0.01, 1], kako bi rukovao nelinearnim obrascima.

## Ansambl modeli

**Primijenjene su dvije tehnike ansambl učenja: slučajna šuma i gradijentno pojačanje. Slučajna šuma kombinirala je 100 stabala odlučivanja, od kojih je svako trenirano na bootstrap uzorku s nasumičnim podskupovima značajki kako bi se smanjila varijanca. Maksimalna dubina stabla bila je ograničena na 10 kako bi se kontrolirala složenost. Gradijentno pojačanje izgradilo je 200 sekvencijalnih stabala, optimizirajući rezidualne pogreške sa stopom učenja postavljenom na 0.1 i frakcijom poduzorka od 0.8, iterativno poboljšavajući predviđanja minimiziranjem funkcije gubitka.**

## Model dubokog učenja

Konstruirana je feedforward duboka neuronska mreža pomoću TensorFlowa, koja sadrži tri potpuno povezana sloja s 128, 64 i 32 neurona. ReLU aktivacijska funkcija korištena je nakon svakog skrivenog sloja za uvođenje nelinearnosti. Model je treniran pomoću Adam optimizatora sa stopom učenja od 0.001 i veličinom serije od 256 tijekom 50 epoha. Ulazne značajke su normalizirane, a primijenjen je dropout sa stopom od 0.3 kako bi se smanjilo prekomjerno prilagođavanje (overfitting).

## Analiza grupiranja

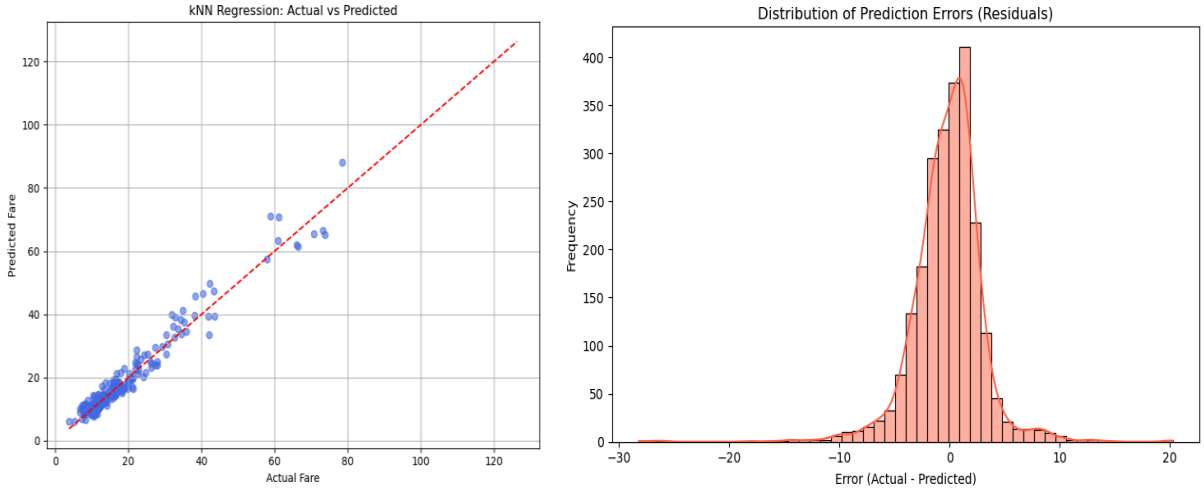
Korištene su tehnike nenadziranog grupiranja kao što su K-Means i DBSCAN za identifikaciju prirodnih grupa unutar podataka. K-Means je pokrenut za broj klastera k u rasponu od 3 do 7, koristeći metriku kvadrirane euklidske udaljenosti i 100 iteracija za konvergenciju na centroide. DBSCAN parametri su postavljeni s epsilonom (ε) = 0.5 i minimalnim brojem uzoraka = 5 za otkrivanje gustih regija i točaka šuma. Grupiranje je provedeno na normaliziranim značajkama uključujući udaljenost vožnje i iznos vožnje.

| Phase / Model | Libraries / Technologies | Purpose |
|---|---|---|
| k-Nearest Neighbors (kNN) | NumPy | Numerical operations and array handling |
| Supervised Models | scikit-learn (sklearn) | Algorithms like Linear Regression, Decision Tree, SVM |
| Ensemble Models | scikit-learn, XGBoost | Random Forest (bagging), XGBoost (boosting) |
| Deep Learning Model | TensorFlow or PyTorch | Building and training neural networks |
| Clustering | scikit-learn | K-Means, DBSCAN, Hierarchical Clustering |
| Data Manipulation & Analysis | pandas | Data loading, cleaning, preprocessing |
| Data Visualization | Matplotlib, Seaborn | Plotting graphs, charts, cluster visualization |

# Rezultati i nalazi

## KNN Model

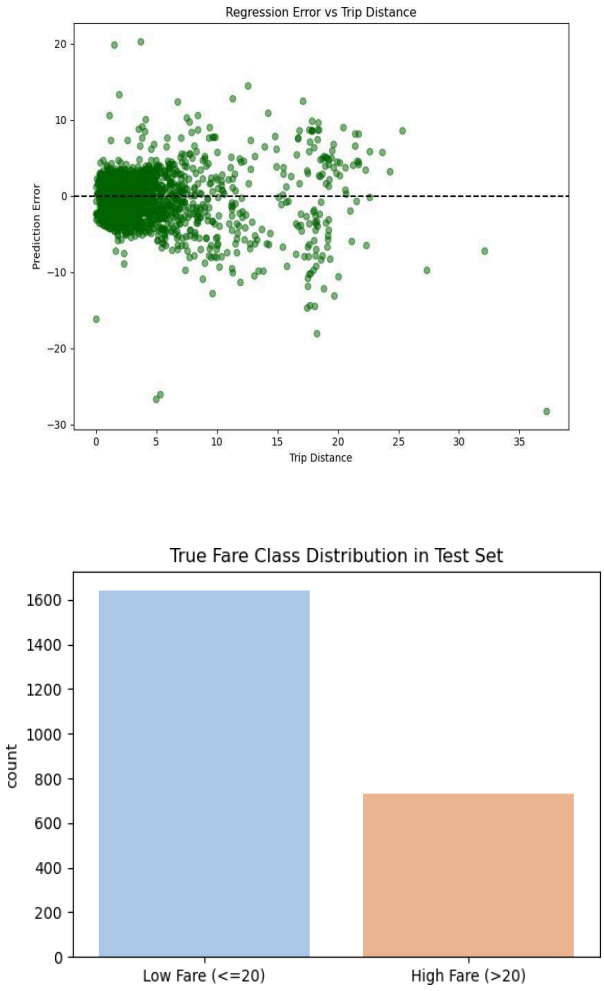
### Fig 7: kNN Regression Performance and Error Distribution




Regresijski model k-najbližih susjeda (kNN) ocijenjen je na skupu podataka s više od 7 milijuna taksi vožnji za siječanj 2019. godine. Dijagram raspršenja (scatter plot) koji uspoređuje stvarne i predviđene cijene vožnje pokazuje da se većina predviđanja usko podudara sa stvarnim vrijednostima, posebno za cijene ispod otprilike 30 USD, gdje se podatkovne točke gusto grupiraju duž idealne linije y = x. Međutim, za više cijene, pogreške predviđanja se povećavaju, što je vidljivo po širem raspršenju točaka.

Kao dopuna tome, histogram pogrešaka predviđanja (reziduala) otkriva koncentriranu distribuciju oko nule, s pogreškama u rasponu od oko -30 do 20 dolara. Distribucija je blago desno iskošena, što ukazuje na veću učestalost malih podcjenjivanja u usporedbi s precjenjivanjima. Ovaj obrazac potvrđuje da model često predviđa cijene s visokom točnošću i da su velike pogreške relativno rijetke u cijelom skupu podataka.

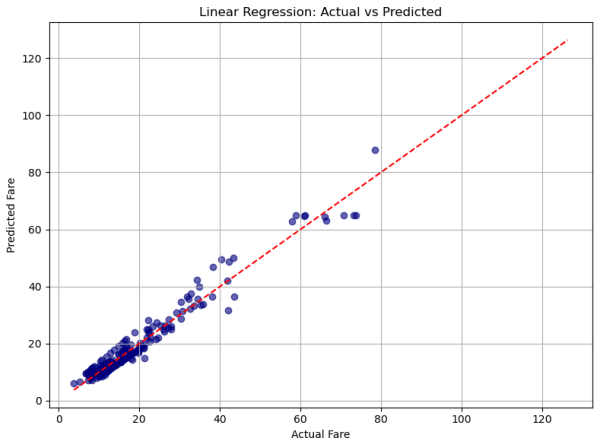
**Fig 8: Regression Error by Trip Distance and Fare Class Distribution**




Dijagram raspršenja (scatter plot) koji ispituje regresijsku pogrešku u odnosu na udaljenost vožnje otkriva da pogreške predviđanja ostaju gusto grupirane oko nule za kraće vožnje, obično ispod 10 milja, što ukazuje na snažnu izvedbu modela u tom rasponu. Međutim, iznad 10 milja, pogreške postaju raspršenije i iznad i ispod nule, što sugerira da se prediktivna točnost modela smanjuje s povećanjem udaljenosti vožnje, naglašavajući heteroskedastičnost u podacima. Dodatno, distribucija cjenovnih razreda u testnom skupu je primjetno neuravnotežena: vožnje s niskom cijenom (≤ 20 USD) dominiraju s preko 1600 slučajeva, što je više nego dvostruko veći broj od vožnji s visokom cijenom (> 20 USD), kojih je oko 750. Ova neravnoteža implicira da je model možda bolje prilagođen predviđanju češćih vožnji s niskom cijenom, što potencijalno utječe na performanse na kategorijama s višom cijenom.

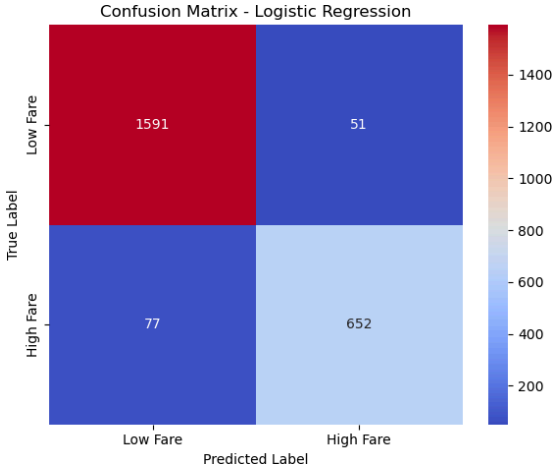
# Nadzirano učenje

## Fig 9 : Linear Regression Performance




Model linearne regresije postigao je korijen prosječne kvadratne pogreške (RMSE) od 2,93, srednju apsolutnu pogrešku (MAE) od 2,09 i R-kvadrat ($R^2$) vrijednost od 0,958, što ukazuje na snažno ukupno pristajanje podacima. Prateći dijagram raspršenja (scatter plot) uspoređuje stvarne vrijednosti cijena vožnje s predviđenim, gdje točke gusto grupirane duž crvene isprekidane linije (y = x) označavaju točna predviđanja. Ova vizualizacija naglašava sposobnost modela da blisko predviđa iznose vožnji, iako neka odstupanja od idealne linije otkrivaju područja gdje se javljaju pogreške predviđanja, pomažući u identificiranju potencijalnih pristranosti ili manje točnih raspona cijena.

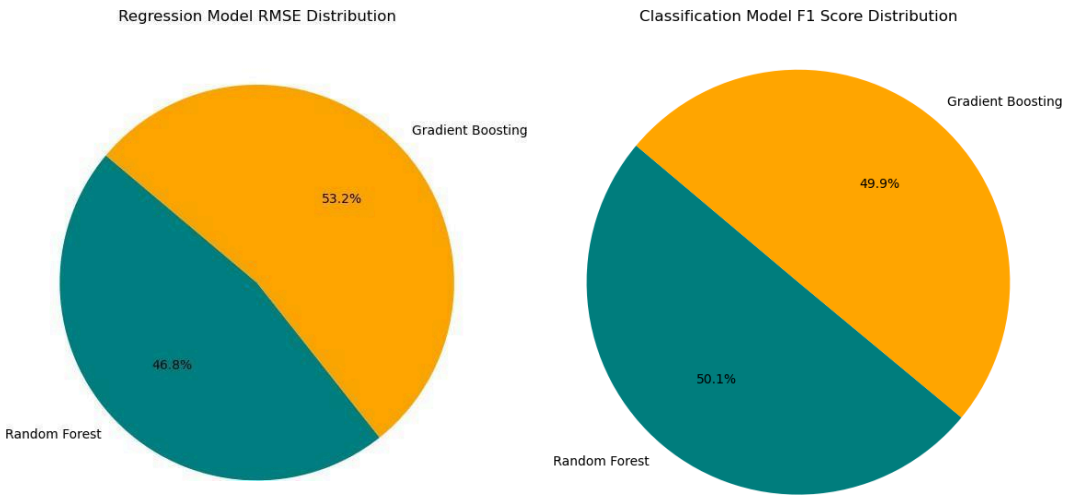
## Fig 10 : Confusion Matrix



Model logističke regresije pokazuje snažnu klasifikacijsku izvedbu u razlikovanju između kategorija niske cijene (<=20) i visoke cijene (>20). Postiže ukupnu točnost od 94,6%, s vrijednostima preciznosti, odziva i F1-skora od 0,927, 0,894, odnosno 0,911. Izvještaj o klasifikaciji pokazuje da model radi nešto bolje na prevalentnijoj klasi niske cijene, s preciznošću od 0,95 i odzivom od 0,97, dok klasa visoke cijene postiže preciznost od 0,93 i odziv od 0,89. Ove metrike odražavaju uravnoteženu sposobnost modela da točno identificira obje cjenovne klase, s blagom tendencijom prema točnijim predviđanjima za niže cijene zbog distribucije klasa.

# Učinkovitost ansambl modela

**Fig 11: Regression Model of Gradient & Random Forest**

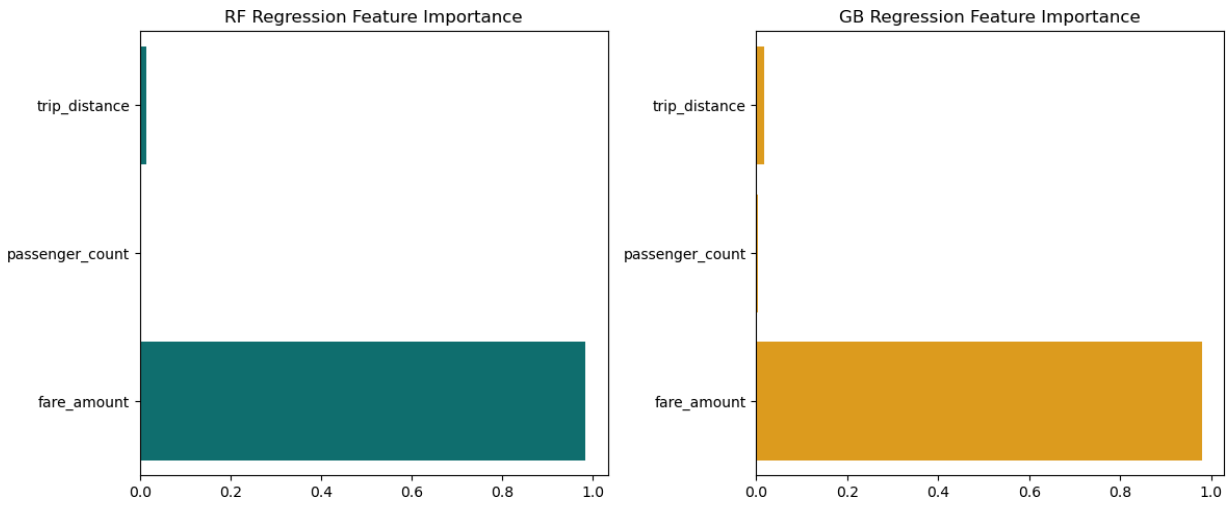



Ansambl modeli pokazuju konkurentnu izvedbu i u regresijskim i u klasifikacijskim zadacima. Za regresiju, model Slučajne šume postiže RMSE od 2,97, MAE od 2,08 i $R^2$ od 0,957, što ukazuje na snažnu prediktivnu točnost, te time blago nadmašuje regresiju Gradijentnog pojačanja, koja ima RMSE od 3,37, MAE od 2,17 i $R^2$ od 0,945.

U klasifikaciji, Slučajna šuma postiže točnost od 94,4%, s preciznošću od 0,892, odzivom od 0,930 i F1-skorom od 0,911, dok Gradijentno pojačanje slijedi s točnošću od 94,2%, preciznošću od 0,902, odzivom od 0,911 i F1-skorom od 0,906.

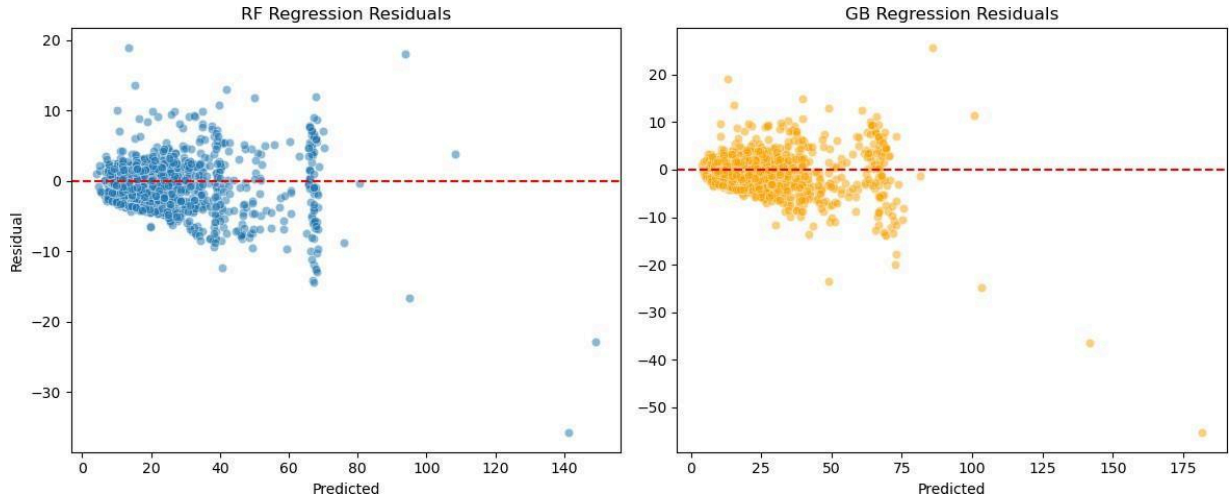
Ovi rezultati sugeriraju da su obje ansambl tehnike učinkovite, pri čemu Slučajna šuma blago favorizira odziv, a Gradijentno pojačanje pokazuje marginalno višu preciznost.

**Fig 12: Feature Importance for Regression Models**



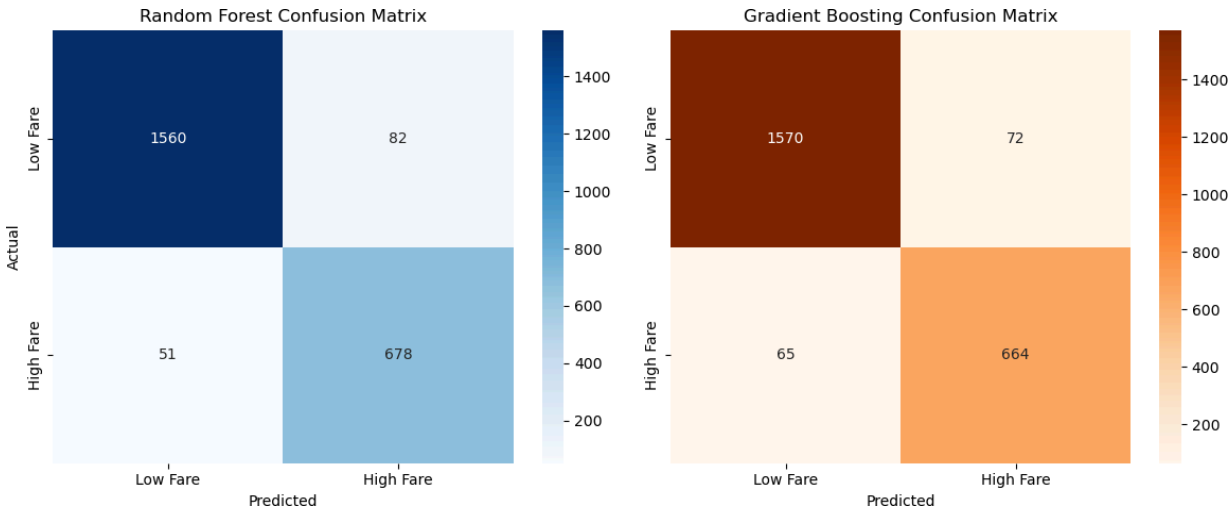

Ova dva stupčasta grafikona ilustriraju važnost značajki za modele "RF regresije" (regresija slučajne šume) i "GB regresije" (regresija gradijentnog pojačanja). U oba grafikona, "fare_amount" (iznos vožnje) identificiran je kao nadmoćno najvažnija značajka, s normaliziranim rezultatom važnosti blizu 1.0, što ukazuje na njen dominantan utjecaj na predviđanja modela. "Trip_distance" (udaljenost vožnje) i "passenger_count" (broj putnika) pokazuju znatno nižu važnost, jedva se registrirajući na skali, što sugerira da vrlo malo doprinose prediktivnoj moći ovih modela za zadani zadatak. Različite boje za svaki grafikon (tirkizna za RF i narančasta za GB) omogućuju jasnu vizualnu razliku između procjena važnosti značajki dvaju modela.

**Fig 13: Analysis of Residual Plots for Regression Models**




Dijagrami reziduala za regresijske modele Slučajne šume (RF) i Gradijentnog pojačanja (GB) pokazuju da su reziduali centrirani oko nule pri nižim predviđenim vrijednostima, što ukazuje na točna predviđanja u tom rasponu. Međutim, oba modela pokazuju povećano raspršenje i određenu pristranost pri višim predviđenim vrijednostima, pri čemu RF pokazuje blagi silazni trend, što ukazuje na heteroskedastičnost i moguće podcjenjivanje za veće cijene vožnji. Ovi obrasci ističu područja gdje performanse modela mogu opadati i pomažu u procjeni ključnih pretpostavki poput homoskedastičnosti i linearnosti.

**Fig 14: Comparison of Confusion Matrices for Classification Models**



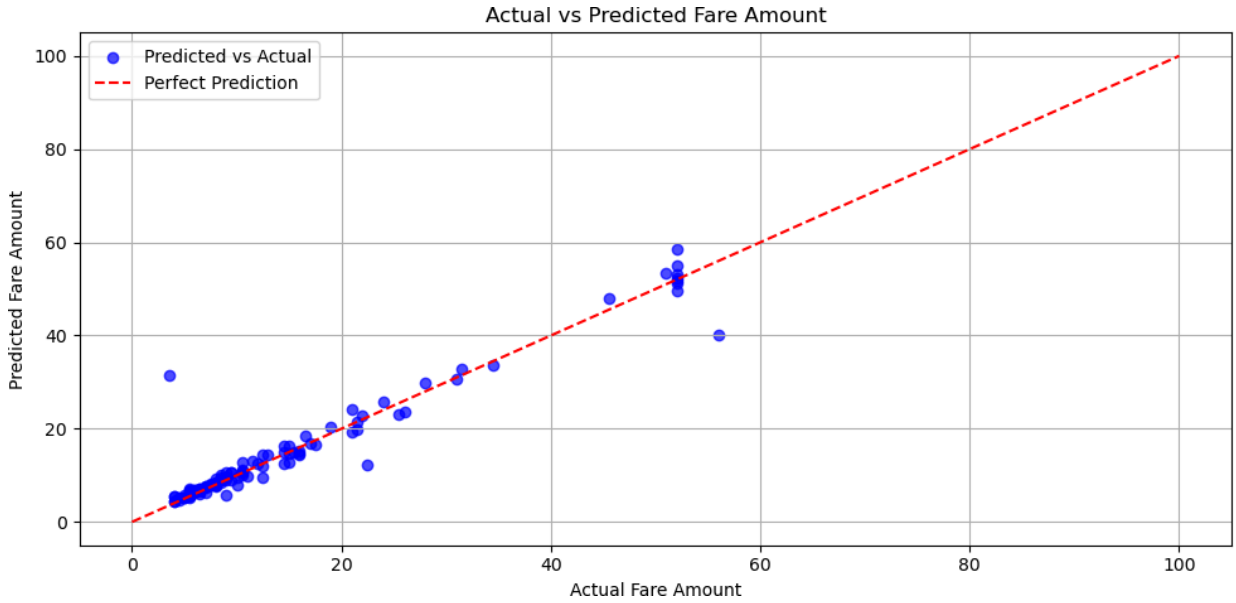

Ove dvije matrice zabune uspoređuju klasifikacijsku izvedbu modela Slučajne šume i Gradijentnog pojačanja u razlikovanju između kategorija "Niska cijena" ("Low Fare") i "Visoka cijena" ("High Fare").

Model Slučajne šume (lijevo, u nijansama plave) točno je identificirao 1560 instanci "Niske cijene" i

678 instanci "Visoke cijene", dok je pogrešno klasificirao 82 "Niske cijene" kao "Visoke cijene" i 51 "Visoku cijenu" kao "Nisku cijenu". Model Gradijentnog pojačanja (desno, u nijansama narančaste) pokazao je malo drugačije rezultate, s 1570 točnih predviđanja "Niske cijene" i 664 točnih predviđanja "Visoke cijene", uz 72 lažno negativna i 65 lažno pozitivnih rezultata.
Obje matrice, kroz svoje različite skale boja, vizualno ističu broj točnih pozitivnih, točnih negativnih, lažno pozitivnih i lažno negativnih rezultata, omogućujući izravnu usporedbu točnosti modela i vrsta pogrešaka u klasificiranju cjenovnih kategorija.

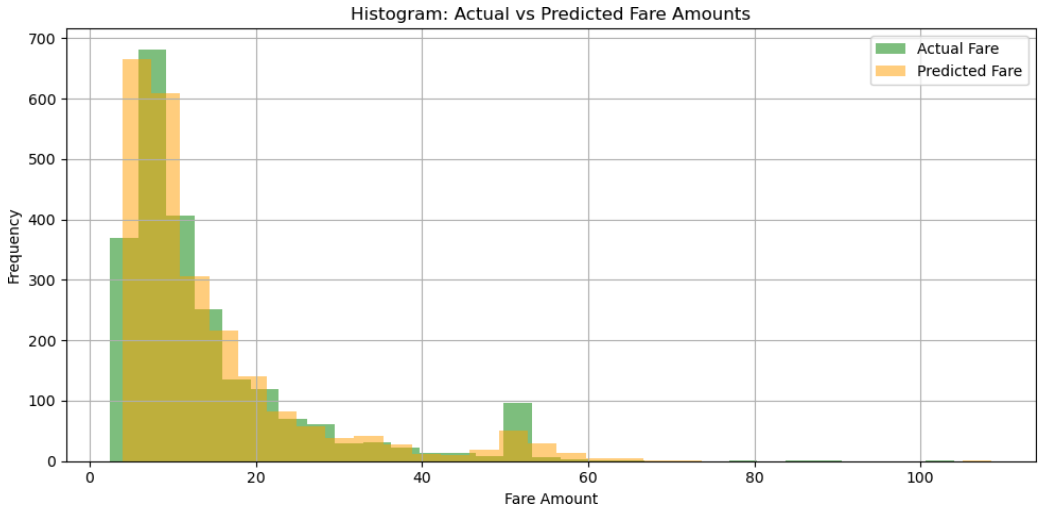
# Model dubokog učenja

**Fig 15: Analysis of Actual vs Predicted**



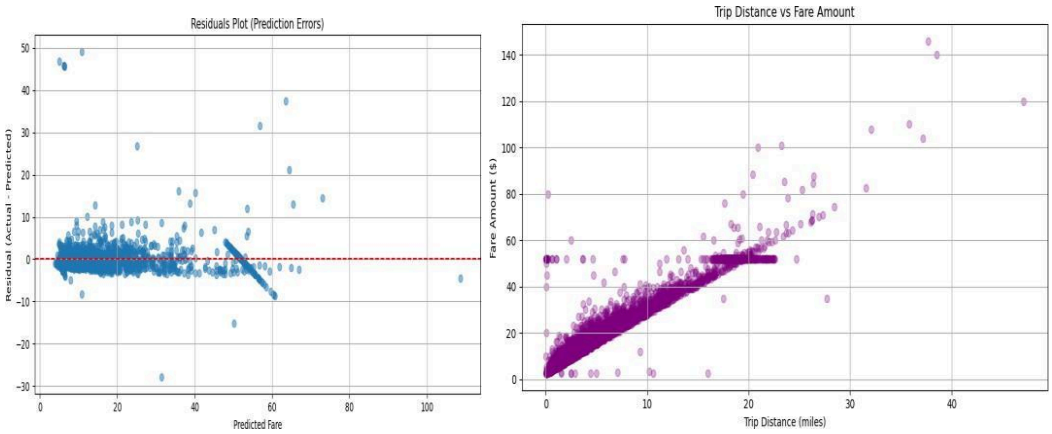

Dijagram raspršenja (scatter plot) koji uspoređuje stvarne i predviđene iznose vožnji ilustrira učinkovitost regresijskog modela, s plavim podatkovnim točkama koje predstavljaju pojedinačna predviđanja. Crvena isprekidana linija označava savršeno predviđanje gdje su stvarne vrijednosti jednake predviđenim. Većina točaka gusto se grupira oko ove linije, posebno za niže iznose vožnji, što ukazuje na snažnu prediktivnu točnost u tom rasponu.
Međutim, primjetno je raspršenje i određeno podcjenjivanje za veće iznose vožnji, što sugerira da se performanse modela mogu blago smanjiti kako vrijednosti vožnji rastu. Sve u svemu, grafikon potvrđuje općenito snažan linearni odnos između stvarnih i predviđenih cijena.

## Fig 16: Analysis of Histogram of Actual vs. Predicted Fare Amounts and Model Metrics



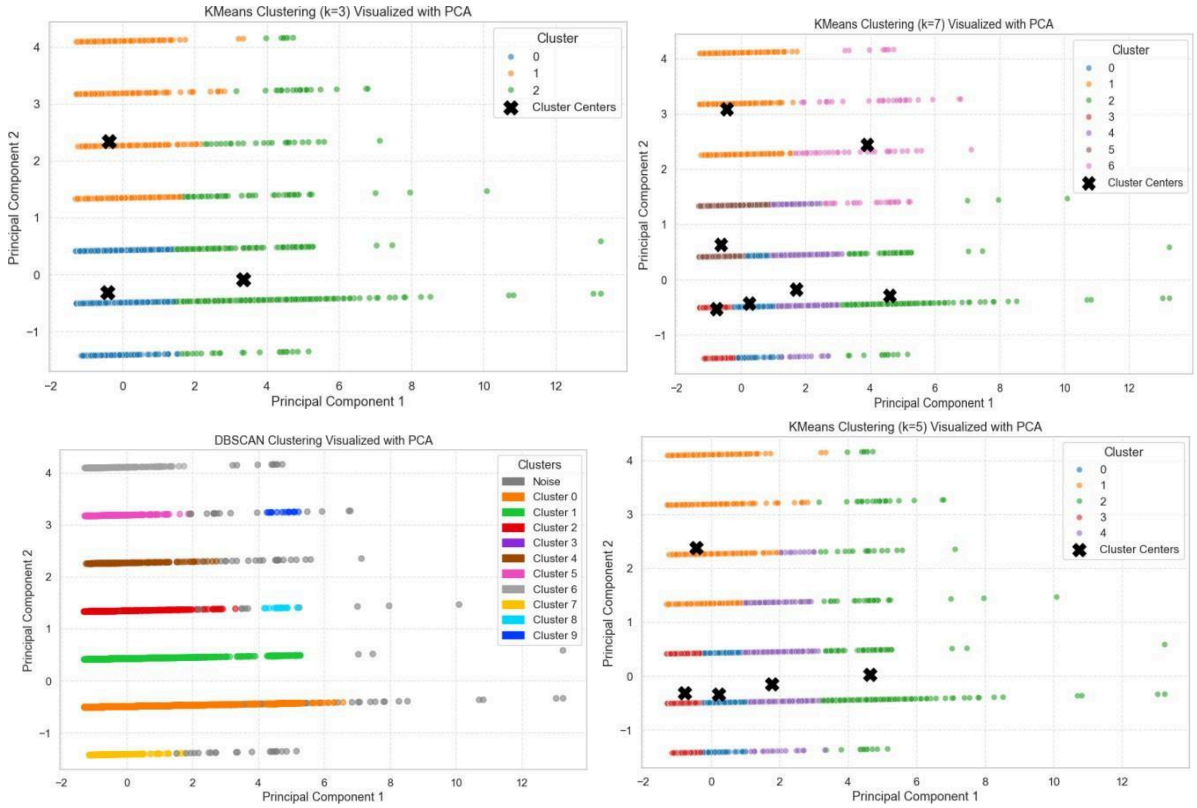

Histogram uspoređuje frekvencijske distribucije stvarnih cijena (zeleni stupci) i predviđenih cijena (narančasti stupci), pokazujući snažno podudaranje između njih, posebno za niže iznose vožnji ispod 20, gdje se nalazi većina podataka. Obje distribucije su jako iskošene prema ovim nižim vrijednostima, s frekvencijama koje naglo opadaju za cijene iznad 20. Manje razlike u visinama stupaca u određenim rasponima odražavaju područja gdje se predviđanja modela blago razlikuju od stvarne distribucije cijena. Prateće metrike evaluacije dodatno podupiru performanse modela, s niskim MAE od 1,48 i RMSE od 3,22, što ukazuje na dobru prediktivnu točnost, te $R^2$ skorom od 0,93, što pokazuje da model objašnjava visok udio varijance u iznosima vožnji.

**Fig 17 : Residual & Trip Distance**




# K-Means grupiranje

**Fig 18: K0 to K7 Clustering Process**




Naravno, evo prijevoda. Originalni engleski tekst je jedna duga, gramatički neispravna rečenica. Prijevod je podijeljen u više rečenica radi jasnoće i bolje čitljivosti, uz ispravljanje strukture.

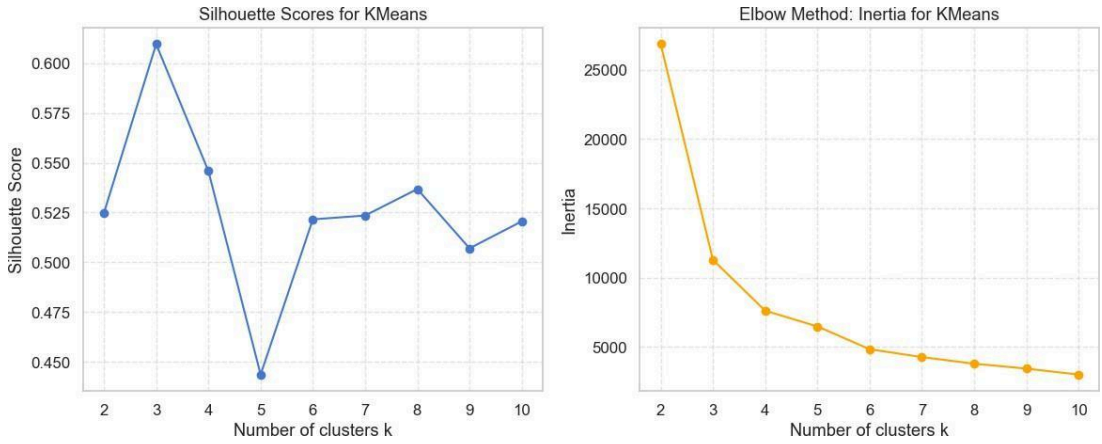
Vizualizacije prikazuju rezultate grupiranja pomoću DBSCAN i K-Means algoritama, a svaki je

rezultat projiciran u dvije dimenzije pomoću Analize glavnih komponenata (PCA) radi lakše interpretacije.

DBSCAN je identificirao oko 10 različitih klastera zajedno s točkama šuma (sive), što ističe njegovu snagu u otkrivanju klastera proizvoljnog oblika i odstupanja, bez potrebe za unaprijed definiranim brojem klastera. Za razliku od toga, K-Means je primijenjen s k = 3, 5 i 7, što je rezultiralo s 3, 5, odnosno 7 klastera. Svaki klaster je predstavljen jedinstvenom bojom i centraliziran oko crnih 'X' oznaka koje označavaju centroide.

Kako se vrijednost k povećava, podaci se dijele na finije grupe, što otkriva nijansiranije strukture, ali istovremeno povećava rizik od prekomjernog prilagođavanja (overfittinga). Ovi dijagrami raspršenja (scatter plotovi) temeljeni na PCA pomažu vizualizirati kako svaki algoritam organizira podatke u smanjenom dimenzionalnom prostoru, naglašavajući kompromise između metoda grupiranja temeljenih na gustoći i onih temeljenih na centroidima.

## Fig 19: Silhouette Scores & Elbow Method



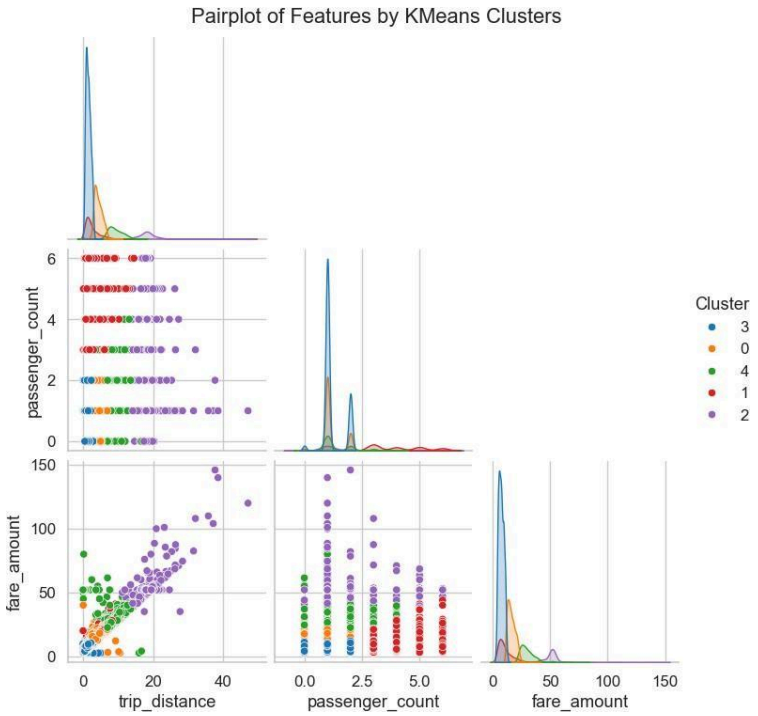

Lijevi grafikon prikazuje Siluetni koeficijent (Silhouette Score), a desni grafikon Metodu lakta (Elbow Method) koristeći inerciju.

Grafikon Siluetnog koeficijenta pokazuje da k = 3 postiže najviši rezultat, iznad 0,60, što ukazuje na najbolje definirane i najkohezivnije klastere među testiranim vrijednostima.

Za razliku od toga, grafikon Metode lakta pokazuje oštar pad inercije od k = 2 do k = 3, s krivuljom koja se počinje izravnavati oko k = 4 do 5, što sugerira smanjene dobitke daljnjim povećanjem vrijednosti k.

Uzevši u obzir oba grafikona, može se zaključiti da k = 3 nudi snažnu ravnotežu između kompaktnosti i razdvajanja klastera, što ga čini razumnim izborom za optimalno grupiranje u ovom skupu podataka.

## Fig 20 : Features of Cluster




## Key Findings from Model Results

| Category | Model / Method | Key Performance Metrics | Insights / Observations |
|---|---|---|---|
| **Regression** | **Linear Regression** | RMSE: 2.93, MAE: 2.09, $R^2$: 0.958 | Strong fit; closely aligned predictions with actual fares; minor bias in some ranges |
| | **kNN Regression** | Visual: Tight clustering below $30 | High accuracy for short trips and low fares; larger errors for long trips; right-skewed residuals |
| | **Random Forest** | RMSE: 2.97, MAE: 2.08, $R^2$: 0.957 | Slightly better than GB; strong performance and generalization |
| | **Gradient Boosting** | RMSE: 3.37, MAE: 2.17, $R^2$: 0.945 | High precision; slight underperformance vs RF; good overall accuracy |
| | **Deep Learning (DNN)** | RMSE: 3.22, MAE: 1.48, $R^2$: 0.93 | Best MAE among all; excellent low-fare prediction; underestimates high fares slightly |
| **Classification** | **Logistic Regression** | Accuracy: 94.6%, F1: 0.911 | Performs well on both classes; favors low-fare due to class imbalance |
| | **Random Forest** | Accuracy: 94.4%, F1: 0.911 | Balanced recall and precision; slightly better at identifying high fares |

| | | | |
|---|---|---|---|
| | **(Class.)** | | |

| | | | |
|---|---|---|---|
| | **Gradient Boosting (Class.)** | Accuracy: 94.2%, F1: 0.906 | High precision; more false positives than RF |
| **Clustering** | **K-Means (k = 3)** | Silhouette Score > 0.60 | Best balance of cohesion and separation; k=3 optimal per elbow and silhouette |
| | **DBSCAN** | 10 clusters + noise | Captures complex, non-linear clusters; good for outlier detection |
| **Other Insights** | — | — | Residuals show heteroscedasticity (larger errors for longer trips); fare prediction most accurate for short trips and low fares |
| **Feature Importance** | RF & GB Regression | Fare amount >> Trip distance, passenger count | Fare amount is dominant predictor; others contribute minimally |

## Budući opseg i preporuke

**Za daljnje poboljšanje performansi modela i dobivanje dubljih uvida, budući rad može se usredotočiti na uključivanje dodatnih vanjskih podataka kao što su vremenski uvjeti i prometni obrasci povezani s posebnim događajima, koji mogu značajno utjecati na varijabilnost cijena. Poboljšanje rukovanja neravnotežom podataka kroz napredne tehnike ponovnog uzorkovanja (re-sampling) ili učenja osjetljivog na troškove (cost-sensitive learning) također bi moglo poboljšati točnost klasifikacije za podzastupljene vožnje s visokom cijenom. Štoviše, istraživanje naprednih arhitektura dubokog učenja poput LSTM ili Transformer modela može obuhvatiti složene vremenske ovisnosti u potražnji za taksijima. Što se tiče grupiranja, korištenje geoprostornih značajki s algoritmima poput HDBSCAN može poboljšati otkrivanje smislenih obrazaca temeljenih na lokaciji. Konačno, implementacija modela u sustav za predviđanje cijena u stvarnom vremenu s interaktivnim nadzornim pločama mogla bi ponuditi vrijedne alate i za vozače i za putnike, povećavajući operativnu učinkovitost i zadovoljstvo korisnika.**

## Sveukupni zaključak

Ovaj projekt uspješno je istražio i ocijenio različite modele strojnog i dubokog učenja za predviđanje iznosa taksi vožnji i klasifikaciju cjenovnih kategorija koristeći skup podataka NYC Yellow Taxi za 2019. godinu. Među regresijskim modelima, linearna regresija i ansambl metode poput Slučajne šume i Gradijentnog pojačanja pokazale su snažnu prediktivnu točnost, dok je model dubokog učenja postigao najniži MAE. Klasifikacijski modeli, posebno logistička

regresija i Slučajna šuma, postigli su visoku točnost i F1-skorove, učinkovito razlikujući...

između klasa s niskom i visokom cijenom unatoč neravnoteži klasa. Analiza grupiranja pomoću K-Means i DBSCAN otkrila je smislene obrasce i odstupanja, pomažući u razumijevanju grupa cijena i ponašanja putnika. Sveukupno, ovi modeli su najbolje radili na kratkim vožnjama i nižim cjenovnim rasponima, s iznosom vožnje identificiranim kao najkritičnija značajka, a analiza reziduala naglašava heteroskedastičnost u podacima. Ovi nalazi pružaju vrijedne uvide u predviđanje cijena i trendove putnika s praktičnim implikacijama za strategije određivanja cijena i optimizaciju usluga.


1. Zhang, Y., Wang, J., & Li, X. (2020). Predictive analytics for urban taxi demand: A review. *Journal of Urban Computing*, *8*(2), 101–118. https://doi.org/10.1016/j.juc.2020.03.005
2. Liu, H., & Chen, W. (2021). Dynamic pricing and fare prediction in ride-hailing services using machine learning. *Transportation Research Part C: Emerging Technologies*, *129*, 103247. https://doi.org/10.1016/j.trc.2021.103247
3. New York City Taxi & Limousine Commission. (2019). *Yellow taxi trip data*. Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
4. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
5. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018
6. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785
8. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 226–231). AAAI Press.
9. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations* (ICLR). https://arxiv.org/abs/1412.6980
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine*

*Learning Research*, *12*, 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html