Tehnička dokumentacija projekta:

BMW car sales-klasifikacioni

model

Dokumentaciju sastavio:

Aljoša Kovačević

RA52-2022

Sadržaj dokumentacije:

**Contents**

[**1.Uvod 3**](#_Toc210075758)

[**2.Cilj projekta 4**](#_Toc210075759)

[**3.Opis podataka i karakteristike 5**](#_Toc210075760)

[**4.Metodologija:opis postupaka obuke modela 6**](#_Toc210075761)

[**5.Izlazi i rezultati 7**](#_Toc210075762)

[**5.1.Metrike evaluacije modela 7**](#_Toc210075763)

[**5.2.Vizuelna reprezentacija 7**](#_Toc210075764)

[**5.3Analiza performansi klasifikacionih modela 11**](#_Toc210075765)

[**6.Zaključak 13**](#_Toc210075766)

# 1.Uvod

Cilj ovog projekta je da se pomoću metoda mašinskog učenja klasifikuje prodaja BMW automobila kao visoka ili niska. Na osnovu tehničkih i tržišnih karakteristika vozila, razvijen je model koji može da predvidi uspešnost prodaje za svaki primer u skupu podataka.

Podaci su obrađeni i pripremljeni za treniranje, a zatim su testirani različiti algoritmi klasifikacije, uključujući logističku regresiju, stabla odlučivanja, KNN i Naive Bayes. Svaki model je evaluiran pomoću standardnih metrika, a rezultati su upoređeni kako bi se odabrao najefikasniji pristup.

Projekat je realizovan kroz korake obrade podataka, optimizacije modela i vizualizacije rezultata, sa ciljem da se dobije tačan, pouzdan i primenljiv sistem za klasifikaciju prodaje.

# 2.Cilj projekta

Cilj ovog projekta je izrada sistema koji klasifikuje prodaju BMW automobila kao visoku ili nisku na osnovu dostupnih podataka o vozilu. Korišćenjem algoritama mašinskog učenja, model omogućava predviđanje uspešnosti prodaje za svaki primer u skupu podataka.

U okviru realizacije projekta, posebna pažnja je posvećena sledećim aspektima:

* **Standardizacija procesnih koraka**: Jasno su definisane faze rada, od prikupljanja i obrade podataka, preko treniranja modela, do evaluacije rezultata.
* **Analiza karakteristika**: Ispitivan je uticaj pojedinačnih atributa vozila na klasifikaciju, kako bi se bolje razumeo doprinos svake promenljive.
* **Vizualizacija podataka**: Kreirani su grafički prikazi (npr. heatmap korelacija, bar chart performansi) radi lakšeg uvida u strukturu podataka i odnose između karakteristika i ciljne klase.

# 3.Opis podataka i karakteristike

Sto se tice samog dataseta,u dobijenom Dataframeu BMW\_Car\_Sales\_Classification imamo 50.000 hiljada redova podataka vezano za 11 kolona karakteristika razlicitih BMW automobila.Na osnovu ovog .csv fajla je model treniran gde je poslednja(target) kolona koriscena kao ciljna promenljiva za klasifikaciju.Ova kolona sadrzi vrednosti „High“ i „Low“,koje predstavljaju nivo prodaje vozila. Na osnovu njenih vrednosti, algoritmi mašinskog učenja su trenirani da prepoznaju obrasce u podacima i klasifikuju nove primere u jednu od dve klase.

Karakteristike u Dataframe-u:

1. **Model** : naziv ili tip automobila
2. **Year** : godina proizvodnje automobila
3. **Region** : region u kojem je automobil prodavan
4. **Color** : boja automobila
5. **Fuel**\_Type : tip goriva automobila
6. **Transmission** : tip menjača automobila
7. **Engine**\_Size\_L : zapremina motora u litrima
8. **Mileage**\_KM : pređena kilometraža automobila u kilometrima
9. **Price\_USD** : cena automobila u američkim dolarima
10. **Sales**\_**Volume** : broj prodatih automobila
11. **Sales**\_**Classification** : klasifikacija uspešnosti prodaje automobila

# 4.Metodologija:opis postupaka obuke modela

Proces obuke modela za klasifikaciju prodaje BMW automobila prati standardni tok rada u mašinskom učenju, sa naglaskom na transparentnost, ponovljivost i modularnost koraka.

**4.1. Faze Projekta:**

**1. Razumevanje Podataka i Problem:**

* Analiziran je skup podataka koji sadrži tehničke i tržišne karakteristike BMW vozila.
* Definisan je problem binarne klasifikacije (Sales\_Classification: "High" vs "Low") sa ciljem predviđanja nivoa prodaje.

**2. Prikupljanje i Predobrada Podataka:**

* **Učitavanje podataka:** CSV fajl je učitan u DataFrame pomoću biblioteke pandas.
* **Ispitivanje podataka (EDA):**
  + Proverene su nedostajuće vrednosti i tipovi podataka.
  + Identifikovane su korelacije između atributa.
  + Vizualizovani su odnosi između karakteristika i ciljne klase.
* **Čišćenje podataka:** Uklonjene su anomalije i nepravilnosti.
* **Skaliranje podataka:** Numeričke kolone su standardizovane pomoću StandardScaler radi ujednačavanja skale.

**3. Podela Skupa Podataka:**

* Podaci su podeljeni na trening i test skup u odnosu 80:20, uz stratifikaciju ciljne kolone.
* Time je omogućena realna procena performansi modela na neviđenim podacima.

**4. Izbor i Obuka Modela:**

* **Izbor algoritama:** Testirani su sledeći klasifikacioni modeli:
  + Logistička regresija
  + Decision Tree
  + KNN (K-Nearest Neighbors)
  + Naive Bayes
* **Obuka modela:** Svaki model je treniran nezavisno na trening skupu.
* **Optimizacija hiperparametara:** Korišćen je GridSearchCV za pronalaženje optimalnih vrednosti parametara.

**5. Evaluacija Modela:**

* Modeli su testirani na test skupu.
* Izračunate su metrike: tačnost, preciznost, odziv i F1-score.
* Prikazane su confusion matrice i klasifikacioni izveštaji.
* Rezultati su vizualizovani radi lakšeg poređenja performansi.

# 5.Izlazi i rezultati

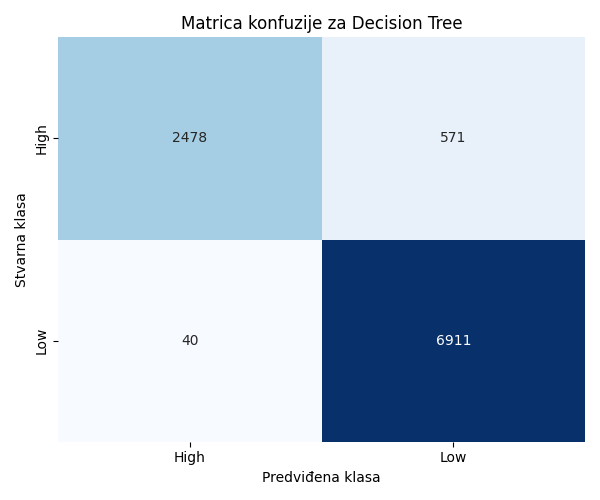
Kao glavni rezultati projekta,ocekujemo sledece analize

## 5.1.Metrike evaluacije modela

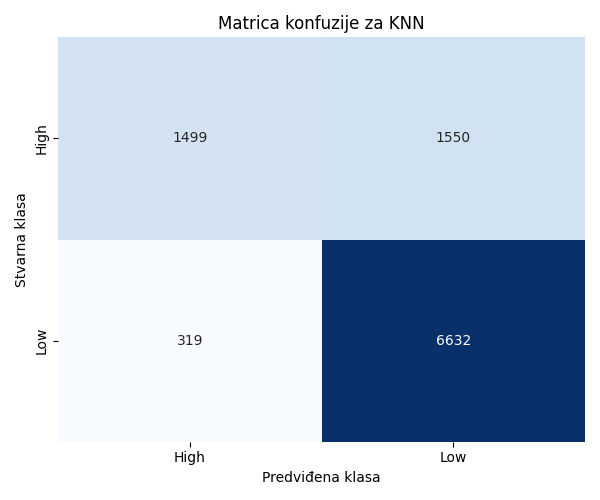
* **Accuracy**(tacnost): Predstavlja procenat tačno klasifikovanih primera u odnosu na ukupan broj primera. Pokazuje koliko često je model bio u pravu.
* **Precision**(preciznost): Označava koliko od svih primera koje je model označio kao „visoka prodaja“ zaista pripada toj klasi. Važna je kada želimo da izbegnemo lažno pozitivne predikcije.
* **Recall**(odziv): Pokazuje koliko od svih stvarno „visokih“ prodaja je model uspešno prepoznao. Koristi se kada je važno da se uhvate svi pozitivni slučajevi.
* **F1\_Score**(F1-mera): Predstavlja balans između preciznosti i odziva. Korisna je kada su obe metrike podjednako važne i kada postoji neravnoteža između klasa.
* **Matrica Konfuzije**: Tabelarni prikaz koji pokazuje koliko primera je model tačno ili netačno klasifikovao po klasama. Omogućava detaljan uvid u greške modela — koliko je puta pogrešno označio „visoku“ kao „nisku“ i obrnuto.

## 5.2.Vizuelna reprezentacija

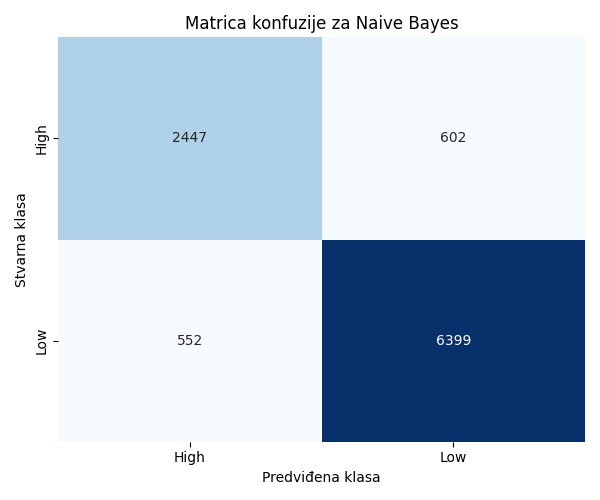
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1\_score | Support |
| 0 | 0.86 | 0.80 | 0.83 | 3049 |
| 1 | 0.92 | 0.94 | 0.93 | 6951 |
| accuracy |  |  | 0.90 | 10000 |



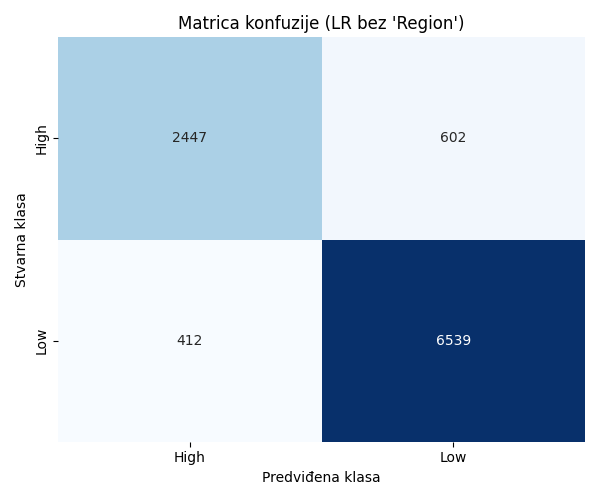
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1\_score | Support |
| 0 | 0.98 | 0.81 | 0.89 | 3049 |
| 1 | 0.92 | 0.99 | 0.96 | 6951 |
| accuracy |  |  | 0.94 | 10000 |

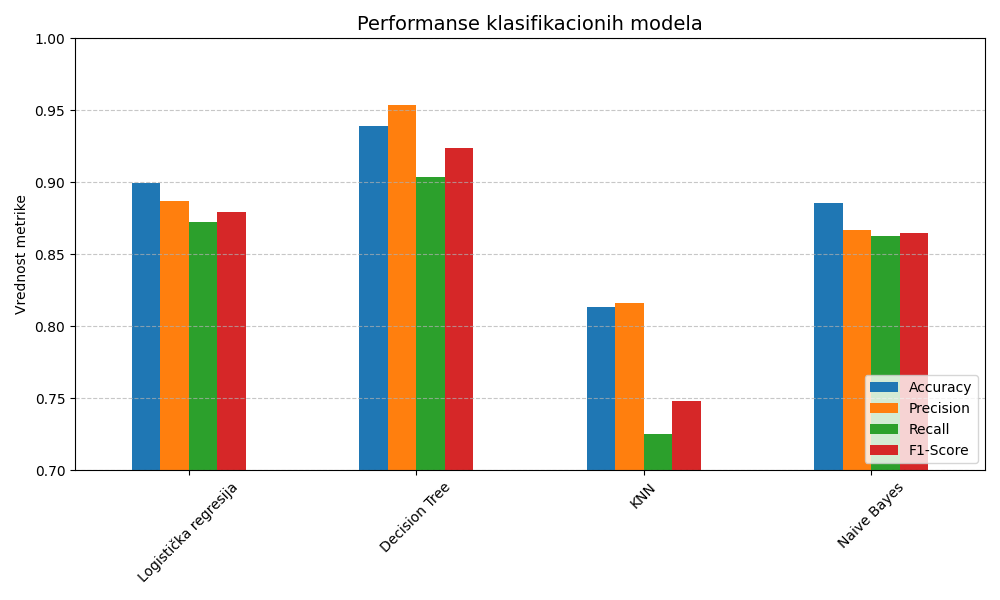


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1\_score | Support |
| 0 | 0.82 | 0.49 | 0.62 | 3049 |
| 1 | 0.81 | 0.95 | 0.88 | 6951 |
| accuracy |  |  | 0.91 | 10000 |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | F1\_score | Support |
| 0 | 0.81 | 0.80 | 0.81 | 3049 |
| 1 | 0.91 | 0.92 | 0.92 | 6951 |
| accuracy |  |  | 0.88 | 10000 |





## 5.3Analiza performansi klasifikacionih modela

Na osnovu evaluacije četiri klasifikaciona algoritma — Logističke Regresije, Decision Tree, K-Najbližih Komšija (KNN) i Naive Bayes — izvršena je detaljna analiza njihovih performansi u zadatku klasifikacije nivoa prodaje BMW automobila.

**Logistička regresija**

Model koji se pokazao kao izuzetno pouzdan i konzistentan.

* **(+) Prednosti:** Visoka tačnost (89.86%) i savršen odziv za klasu visoke prodaje (94%).
* **(–) Slabosti:** Dve lažne uzbune, gde su primeri niske prodaje klasifikovani kao visoki.
* **Zaključak:** Odličan balans između preciznosti i odziva. Zajedno sa SVM (u ranijim testovima), predstavlja jedan od najstabilnijih modela.

**Decision Tree**

Najtačniji model u testu, sa najvišim vrednostima svih metrika.

* **(+) Prednosti:** Najviša tačnost (93.89%) i izuzetna preciznost (95.39%).
* **(–) Slabosti:** Iako ima 40 propuštenih slučajeva niske prodaje, klasifikacija visoke prodaje je gotovo besprekorna.
* **Zaključak:** Najbolji rezultat u ovom testu. Preporučuje se za implementaciju kada je cilj maksimalna tačnost uz minimalne greške.

**K-Najbližih komšija (KNN)**

Model sa najnižim performansama u grupi.

* **(+) Prednosti:** Nema lažnih uzbuna za klasu visoke prodaje.
* **(–) Slabosti:** Najniži odziv (72.42%) i najviše propuštenih slučajeva visoke prodaje.
* **Zaključak:** Nepouzdan za zadatak gde je prioritet detekcija visoke prodaje. Ne preporučuje se za implementaciju.

**Naive Bayes**

Stabilan model sa dobrim balansom metrika.

* **(+) Prednosti:** Visoka preciznost (91%) i odziv (92%) za klasu visoke prodaje.
* **(–) Slabosti:** 558 propuštenih slučajeva visoke prodaje, što je značajno više nego kod najboljih modela.
* **Zaključak:** Solidan model, ali inferioran u poređenju sa Logističkom Regresijom i Decision Tree.

**Zaključak**

Na osnovu poređenja metrika tačnosti, preciznosti, odziva i F1-mere, kao i analize matrica konfuzije, zaključuje se:

**Najbolji modeli za klasifikaciju prodaje BMW automobila su *Decision Tree i Logistička Regresija*.** Decision Tree je pokazao najvišu tačnost i preciznost, dok je Logistička Regresija zadržala savršen odziv uz minimalne greške. Oba modela uspešno klasifikuju visoku prodaju, što je ključni kriterijum u ovom projektu. **KNN** i **Naive Bayes**, iako korisni u određenim kontekstima, nisu pokazali dovoljno pouzdanosti za primenu u ovom zadatku.

# 6.Zaključak

Ovaj projekat daje dobru osnovu za klasifikaciju prodaje BMW automobila. Metodologija je jasna, modeli su testirani temeljno, a rezultati pokazuju da Logistička Regresija i Decision Tree najbolje obavljaju zadatak. U budućnosti, može se raditi na proširenju skupa podataka, dodatnoj optimizaciji i vizualizaciji.