

Izveštaj o Analizi Podataka i Modeliranju

1. Pregled Skupa Podataka

- Skup podataka sadrži informacije vezane za prodaju, uključujući podatke o datumu narudžbine, načinu isporuke, geografskim segmentima, kategorijama proizvoda i prihodima.
- Nakon čišćenja podataka:
 - Uklonjeni su redovi sa nedostajućim vrednostima.
 - Konvertovani su datumi narudžbine i isporuke u datetime format.

Ključna statistika:

- **"Sales"** (prodaja) je standardizovana sa prosekom 0 i standardnom devijacijom 1.
- Kategorijske promenljive su enkodovane korišćenjem LabelEncoder.

2. Istraživačka Analiza Podataka (EDA)

- **Histogrami:** Distribucije numeričkih i kategorijskih promenljivih prikazane su radi identifikacije obrazaca i outliera.
 - Primer: sales pokazuje normalizovanu distribuciju.
- **Korelaciona analiza:** Prikaz korelacije između numeričkih karakteristika pomoću heatmap-a.
 - Visoka korelacija uočena je između određenih geografskih karakteristika i prodaje.

3. Obrada Podataka

- Redukovane nepotrebne kolone, uključujući ship_date, ship_mode, sub_category, segment, i category, za čišćenje podataka i poboljšanje performansi modela.
- Svi podaci standardizovani su korišćenjem StandardScaler.

4. Klaster Analiza

Klasterizacija je primenjena radi identifikacije obrazaca među podacima.

- **KMeans:** Optimalan broj klastera identifikovan metodom lakta.

- **DBSCAN:** Klasterizacija na osnovu gustine.
- **Hijerarhijsko klasterisanje:** Upotreba Ward metode.
- **GMM:** Gaussian Mixture Model.

Zaključak: KMeans i GMM pokazali su bolje rezultate po metriki "Silhouette Score" u poređenju sa ostalim metodama.

5. Vizualizacija Klastera

- PCA je korišćena za redukciju dimenzionalnosti na tri komponente.
- 3D scatter plotovi prikazuju distribuciju klastera prema PCA komponentama.

6. Analiza Karakteristika

- **Lasso Regresija:**
 - Najznačajnije karakteristike: postal_code, order_date, region.
- **Random Forest:**
 - Najveća značajnost: ship_date, order_date, region.

Vizualizacija: Bar grafici za poređenje značajnosti karakteristika.

7. Zaključak i Preporuke

- Klasterizacija omogućava segmentaciju kupaca ili proizvoda na osnovu zajedničkih karakteristika, čime se otvara prostor za unapređenje marketinških strategija.
- Analizom vizualizacija trendova mogu se identifikovati ključni obrasci, poput onih gde proizvodi iz prvog i drugog klastera beleže povećanu prodaju u specifičnim vremenskim periodima, dok proizvodi iz trećeg klastera ostvaruju stabilnu prodaju tokom čitave godine.
- Takođe, klasterizacija omogućava identifikaciju sezonskih i kontinuiranih obrazaca kupovine, pružajući osnovu za preciznije predviđanje potražnje i optimizaciju poslovnih odluka.