04 Domaci SVM

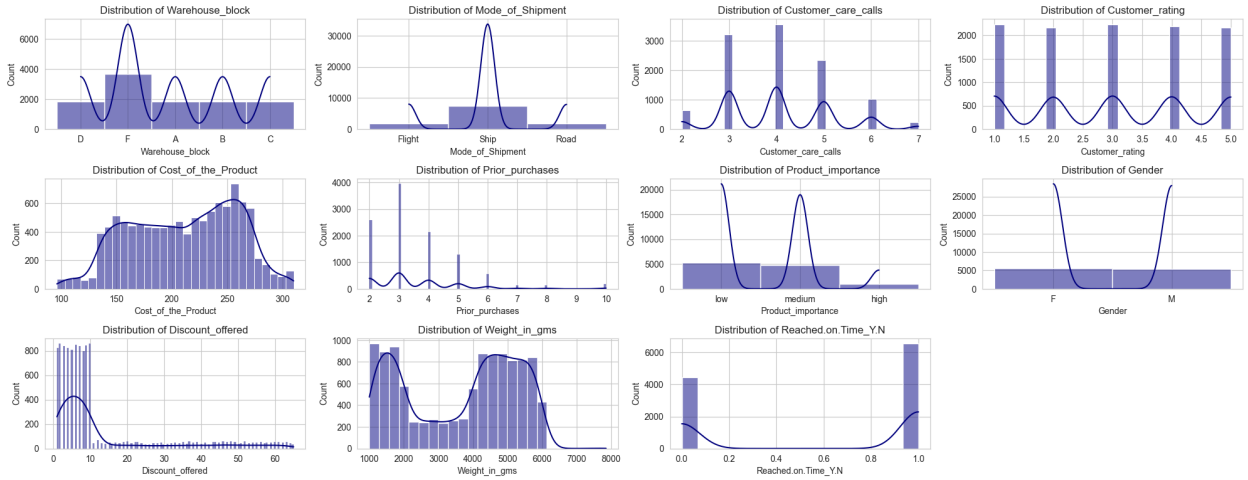
Predmet: Tehnike i metode analize podataka

Student 1636 Milica Jovanovic

Za potrebe ovog domaćeg zadatka korišćen je dataset: https://www.kaggle.com/datasets/prachi13/customer-analytics, namenjen je za analizu podataka o klijentima i istraživanje njihovog ponašanja.
Dataset sadrži 10.999 podataka sa 12 atributa.
Podaci sadrže sledeće informacije:

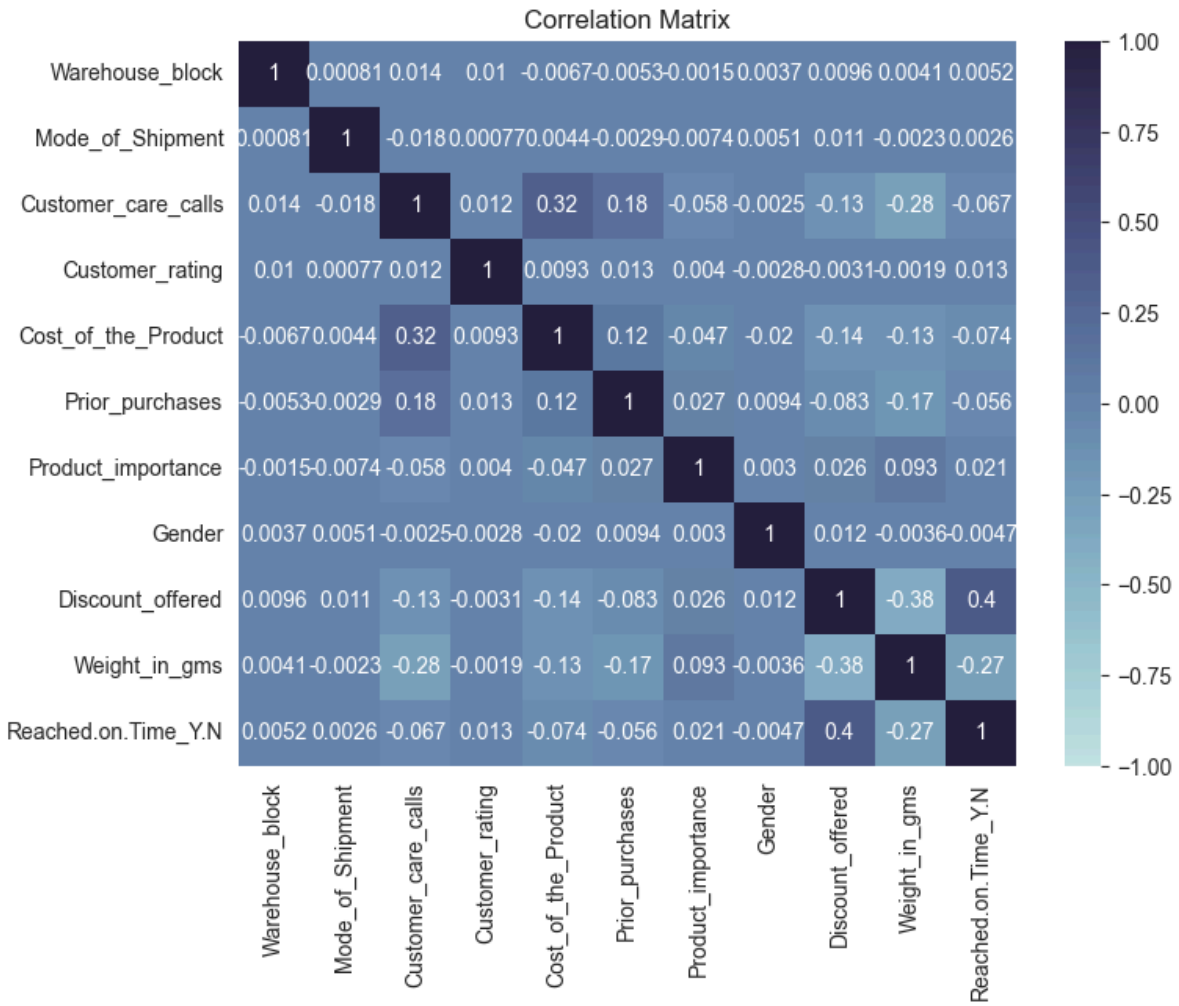
- **ID:** ID broj kupca.
- **Warehouse block (blok magacina):** Kompanija ima veliki magacin koji je podeljen na blokove kao što su A, B, C, D, E.
- **Mode of shipment (način dostave):** Kompanija šalje proizvode na više načina, kao što su brod, avion i put.
- **Customer care calls (pozivi za korisničku podršku):** Broj poziva napravljenih zbog upita o isporuci.
- **Customer rating (ocena kupca):** Kompanija je ocenila svakog kupca. 1 je najniža ocena (najgore), a 5 je najviša ocena (najbolje).
- **Cost of the product (cena proizvoda):** Cena proizvoda u američkim dolarima.
- **Prior purchases (prethodne kupovine):** Broj prethodnih kupovina kupca.
- **Product importance (važnost proizvoda):** Kompanija je kategorizovala proizvode prema različitim parametrima kao što su nizak, srednji i visok.
- **Gender (pol):** Muški i ženski.
- **Discount offered (ponuđeni popust):** Popust ponuđen na određeni proizvod.
- **Weight in gms (težina u gramima):** Težina proizvoda u gramima.
- **Reached on time (isporučeno na vreme):** Ciljna promenljiva, gde 1 označava da proizvod NEMA isporuku na vreme, a 0 označava da je proizvod isporučen na vreme.

Dataset nema null vrednosti, duplikate niti anomalije i ima relativno dobru raspodelu vrednosti. Izbacena je samo kolona ID jer nike relevantna za dalje istrazivanje.



Prilikom prebacivanja kategorickih vrednosti, spopjene su vrednosti za kolonu Mode_of_Shipment, tj, Flight i Road su objedinjene radi boljeg balansa.

```
data['Mode_of_Shipment'] =
data['Mode_of_Shipment'].map({'Flight':1,'Ship':2, "Road": 1})
```

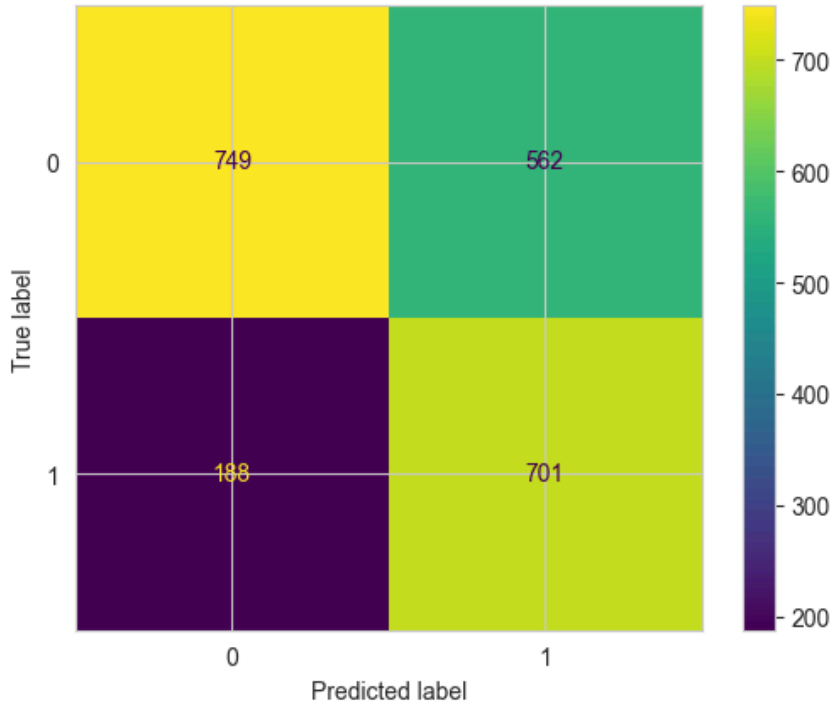


Dataset je podeljen na test i train skupove u odnosu 80:20, i podaci su normalizovani koriscenjem formule date na racunskim vezbama.

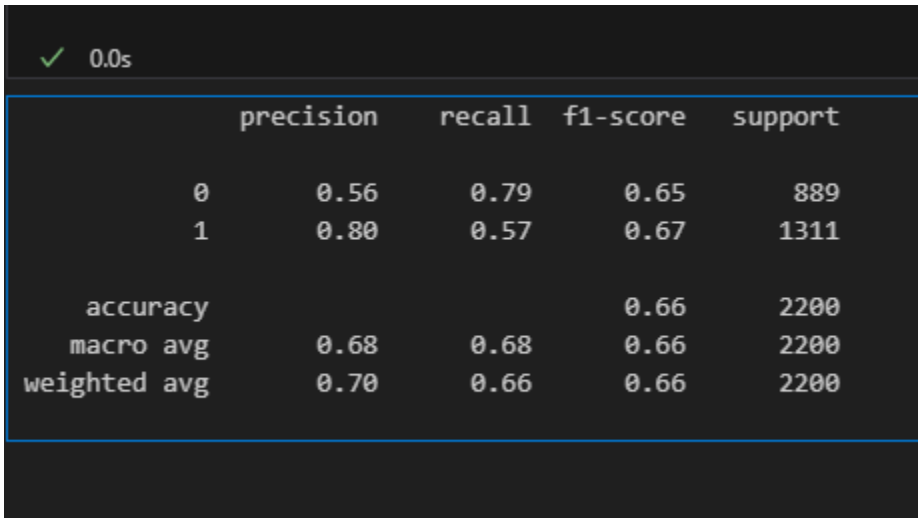
```
# Normalizacija trening skupa (slično radi i MinMaxScaler)
X_train_min = X_train.min()
X_train_max = X_train.max()
X_train_range = (X_train_max- X_train_min)
X_train_scaled = (X_train - X_train_min)/(X_train_range)
print(X_train_scaled.head())
```

✓ 0.0s

Matrica konfuzije modela as default vrednostima:



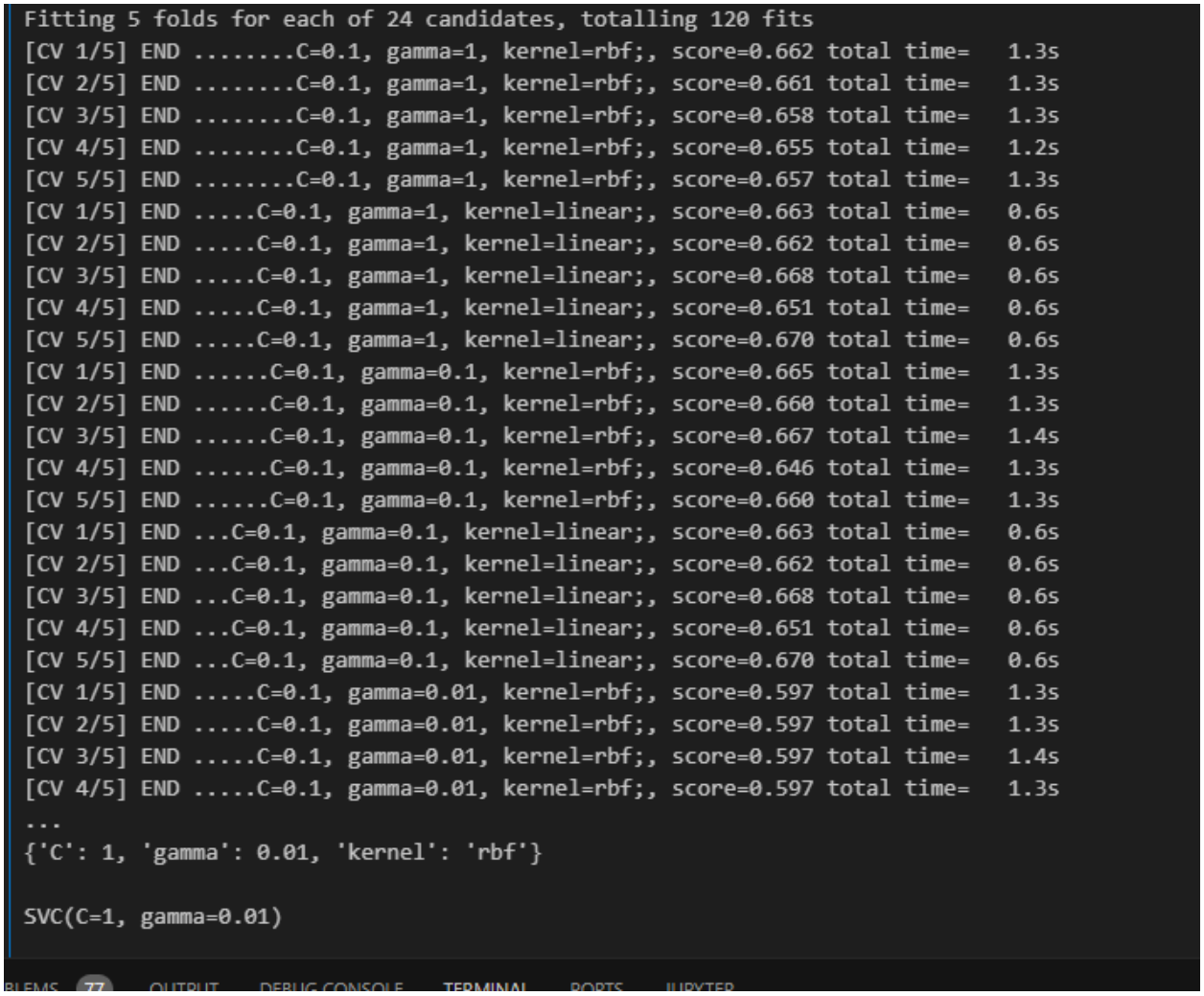
Izvestaj klasifikacije:



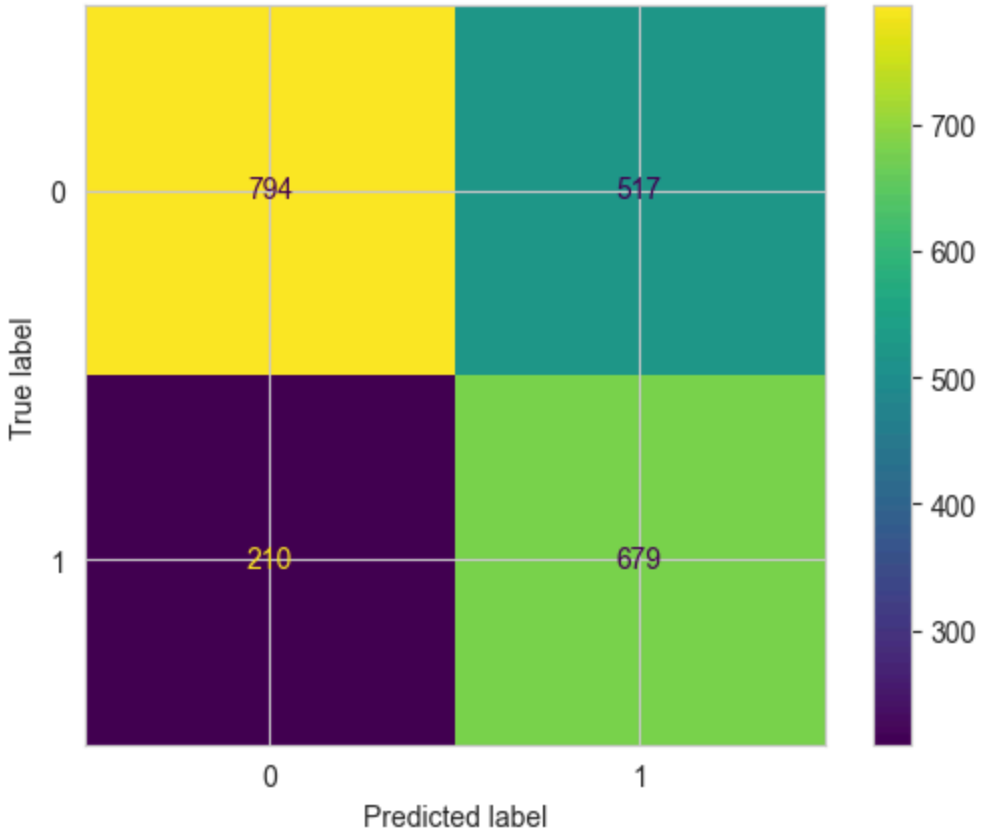
✓ 0.0s

	precision	recall	f1-score	support
0	0.56	0.79	0.65	889
1	0.80	0.57	0.67	1311
accuracy			0.66	2200
macro avg	0.68	0.68	0.66	2200
weighted avg	0.70	0.66	0.66	2200

Zatim je odradjen GridSearchCV radi pronalazenja najboljeg modela. Ovo su dobijeni parametri:



Zatim je napravljen model sa tim paramentrima i ovo je rezultat predikcije tog modela:




```
print(classification_report(y_test,y_predict))
```

✓ 0.0s

	precision	recall	f1-score	support
0	0.57	0.76	0.65	889
1	0.79	0.61	0.69	1311
accuracy			0.67	2200
macro avg	0.68	0.68	0.67	2200
weighted avg	0.70	0.67	0.67	2200