



# Airline Passenger Satisfaction

U ovom projektu istražuje se skup podataka koji sadrži rezultate ankete o zadovoljstvu putnika avio-kompanije. Cilj je identifikovati faktore koji su visoko povezani sa zadovoljstvom ili nezadovoljstvom putnika i **predvideti zadovoljstvo** putnika na osnovu dostupnih informacija. Analizirajući ovaj skup podataka, dobijamo dublji uvid u ključne determinante zadovoljstva putnika i razvijamo model koji može predvideti zadovoljstvo na osnovu relevantnih faktora. Ova saznanja mogu biti korisna za avio-kompanije u poboljšanju kvaliteta usluga i iskustva putnika. U nastavku dokumentacije, detaljnije će se prikazati analiza, rezultati i zaključci dobijeni iz ovog istraživanja.

## Skup podataka

Podaci iz tabele pružaju raznolike informacije o putnicima i njihovim iskustvima tokom leta. Neki od njih se odnose na demografske informacije (poput pola i starosti), dok se drugi odnose na zadovoljstvo različitim aspektima putovanja (poput usluge, udobnosti, hrane itd.). Kroz analizu ovih atributa, mogu se istražiti veze između različitih faktora i zadovoljstva putnika, kao i predvideti nivo zadovoljstva na osnovu dostupnih informacija.

1. Gender: Pol putnika (Female, Male)
2. Customer Type: Tip putnika (Loyal customer, Disloyal customer)
3. Age: Stvarna starost putnika (integer vrednost)
4. Type of Travel: Vrsta putovanja putnika (Personal Travel, Business Travel)

5. Class: Klasa putovanja u avionu (Business, Eco, Eco Plus)
6. Flight Distance: Dužina leta (integer vrednost)
7. Inflight wifi service: Nivo zadovoljstva uslugom bežičnog interneta tokom leta (ocena od 0 do 5)
8. Departure/Arrival time convenient: Nivo zadovoljstva vremenom polaska/dolaska (ocena od 0 do 5)
9. Ease of Online booking: Nivo zadovoljstva online rezervacijom (ocena od 0 do 5)
10. Gate location: Nivo zadovoljstva lokacijom gejta (ocena od 0 do 5)
11. Food and drink: Nivo zadovoljstva hranom i pićem (ocena od 0 do 5)
12. Online boarding: Nivo zadovoljstva online ukrcavanjem (ocena od 0 do 5)
13. Seat comfort: Nivo zadovoljstva udobnošću sedišta (ocena od 0 do 5)
14. Inflight entertainment: Nivo zadovoljstva zabavom tokom leta (ocena od 0 do 5)
15. On-board service: Nivo zadovoljstva uslugom osoblja tokom leta (ocena od 0 do 5)
16. Leg room service: Nivo zadovoljstva prostorom za noge (ocena od 0 do 5)
17. Baggage handling: Nivo zadovoljstva rukovanjem prtljagom (ocena od 0 do 5)
18. Check-in service: Nivo zadovoljstva uslugom pri prijavi (ocena od 0 do 5)
19. Inflight service: Nivo zadovoljstva uslugom tokom leta (ocena od 0 do 5)
20. Cleanliness: Nivo zadovoljstva čistoćom (ocena od 0 do 5)
21. Departure Delay in Minutes: Kašnjenje u minutima prilikom polaska (integer vrednost)
22. Arrival Delay in Minutes: Kašnjenje u minutima prilikom dolaska (float vrednost)
23. Satisfaction: Nivo zadovoljstva putnika (Satisfaction, neutral or dissatisfaction)

## Preprocesiranje podataka

Za potrebe ovog projekta korišćena su **dva CSV fajla**: trening i test. Trening CSV fajl sadrži podatke na kojima je vršena analiza i modeliranje, dok test CSV fajl sadrži podatke koji su korišćeni za evaluaciju modela.

Nakon učitavanja podataka iz CSV datoteka, pristupa se preprocesiranju podataka kako bi se osiguralo da su podaci spremni za dalju analizu i modelovanje. U ovoj fazi

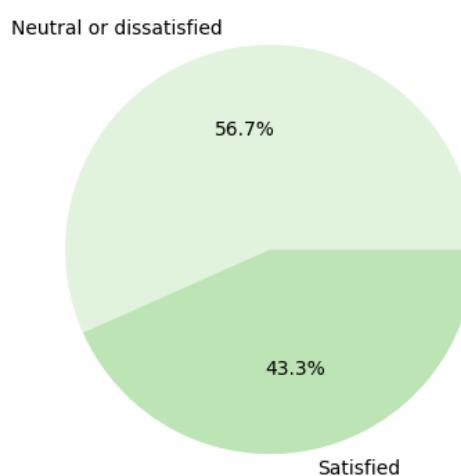
preprocesiranja, obavljeni su sledeći koraci:

## Uklanjanje Bespotrebnih Kolona

	<b>id</b>	Gender	Customer Type	Age	Type of Travel
0	70172	Male	Loyal Customer	13	Personal Travel
1	5047	Male	disloyal Customer	25	Business travel
2	110028	Female	Loyal Customer	26	Business travel
3	24026	Female	Loyal Customer	25	Business travel
4	119299	Male	Loyal Customer	61	Business travel
5	111157	Female	Loyal Customer	26	Personal Travel
6	82113	Male	Loyal Customer	47	Personal Travel
7	96462	Female	Loyal Customer	52	Business travel
8	79485	Female	Loyal Customer	41	Business travel
9	65725	Male	disloyal Customer	20	Business travel
10	34991	Female	disloyal Customer	24	Business travel
11	51412	Female	Loyal Customer	12	Personal Travel
12	98628	Male	Loyal Customer	53	Business travel
13	83502	Male	Loyal Customer	33	Personal Travel
14	95789	Female	Loyal Customer	26	Personal Travel

Prve dve kolone su pre svega uklonjene iz trening i test skupova podataka, budući da su beskorisne za analizu i modeliranje

## Provera Balansiranosti Skupa



Izvršena je analiza raspodele ciljne promenljive 'satisfaction' kako bi se utvrdilo da li je skup podataka izbalansiran ili postoji neravnoteža između zadovoljnih i nezadovoljnih putnika. Grafik prikazuje da je skup **prilično izbalansiran**. : Kada skup podataka nije izbalansiran, model može biti pristran prema većinskoj klasi i prikazivati visoku tačnost samo zato što je uspešno identifikovao većinsku klasu, dok manjinsku klasu ne uspeva tačno predvideti. To može dovesti do iskrivljenih rezultata i pogrešne interpretacije performansi modela.

## Provera Nedostajućih Vrednosti

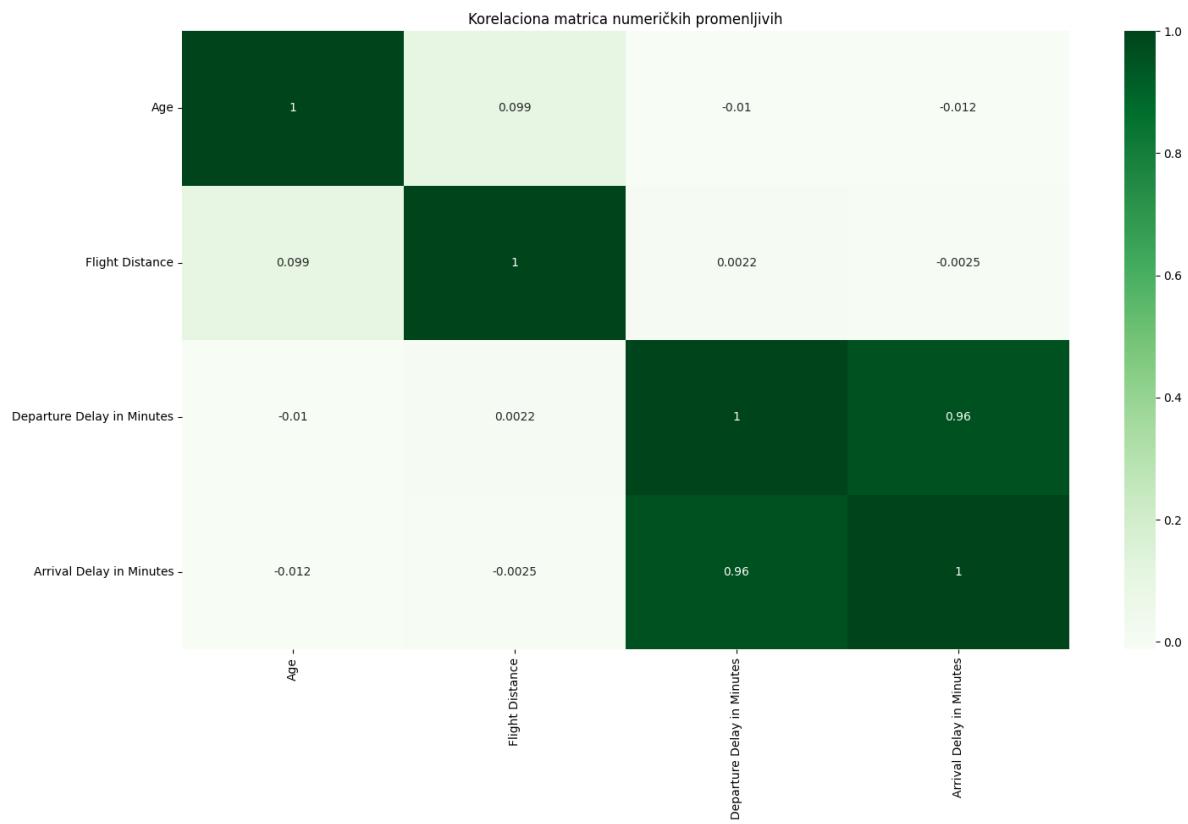
Nakon provere broja nedostajućih vrednosti po koloni, za trening i test skup pomoću funkcije `.isnull().sum()` uočava se da su sve nedostajuće vrednosti u koloni *Arrival Delay in Minutes*. I to za trening skup **310**, i za test skup **83** nedostajuće vrednosti.

Korišćena je metoda popunjavanja vrednosti medijanom. **Medijana** se često koristi za popunjavanje nedostajućih vrednosti jer je otporna na ekstremne vrednosti i neće bitno izmeniti raspodelu podataka.

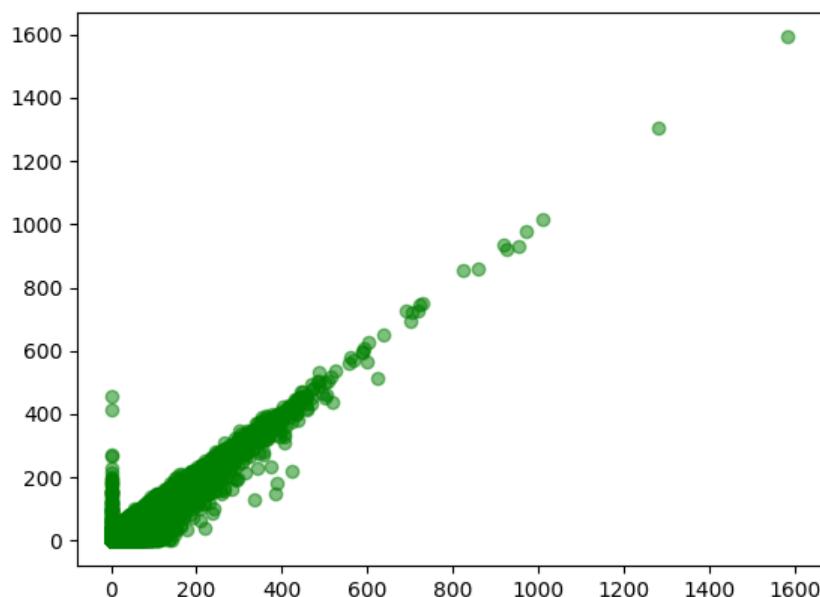
## Provera Korelacije Numeričkih Vrednosti

Korak provere korelacije numeričkih vrednosti ima za cilj da identificuje potencijalne veze između numeričkih atributa u skupu podataka. Koristi se korelaciona matrica kako bi se vizuelno prikazale ove veze.

Prvo, izračunavamo korelacionu matricu pomoću funkcije `corr()` nad numeričkim kolonama skupa podataka. Ova matrica prikazuje korelaciju između svakog para numeričkih atributa. Vrednosti korelacije se prikazuju kao brojevi između -1 i 1, gde vrednost bliža 1 ukazuje na pozitivnu korelaciju, vrednost bliža -1 ukazuje na negativnu korelaciju, dok vrednost blizu 0 ukazuje na slabu ili nultu korelaciju.



Primećuje se **visok stepen zavisnosti** između *Arrival Delay* i *Departure Delay* (0.96) sto je i logično. Dodatno, crtanje scatter plot grafa nam omogućava da vizualno prikažemo linearnu zavisnost između ove dve numeričke kolone. Na osi x prikazujemo vrednosti "*Arrival Delay in Minutes*", dok na osi y prikazujemo vrednosti "*Departure Delay in Minutes*".



Nakon utvrđivanja jake korelacije između *arrival* i *departure delay* odbacujemo kolonu *arrival delay*.

## Pretvaranje Kategoričkih Atributa u Numerički Oblik

U ovom koraku korišćene su dve tehnike: **one-hot kodiranje** i **label kodiranje**.

### One-hot kodiranje

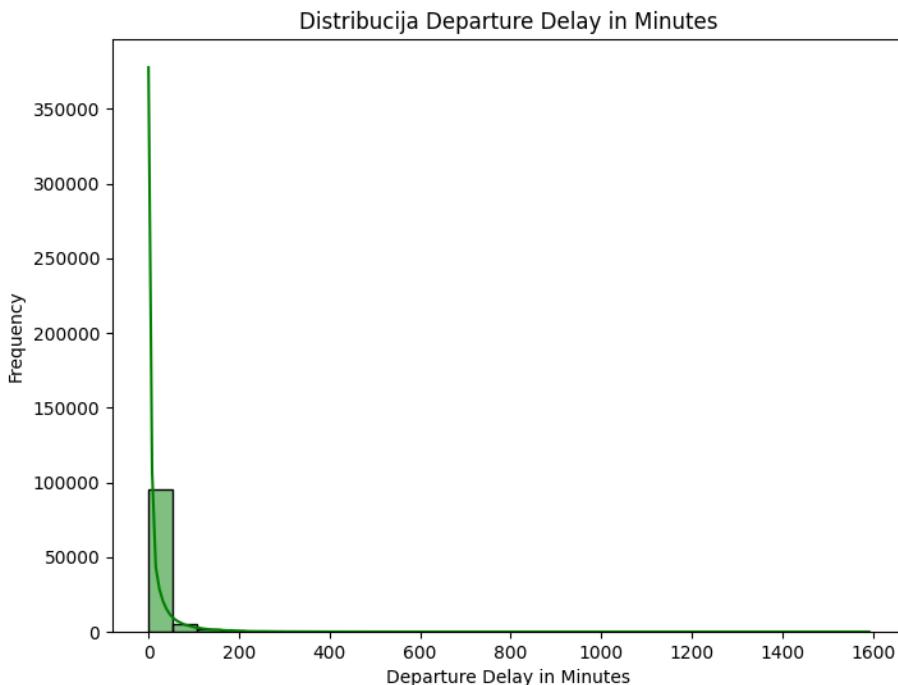
One-hot kodiranje se primenjuje na nezavisne kategoričke kolone u skupu podataka. Ova tehnika pretvara svaku kategoričku vrednost u novu binarnu kolonu. Na primer, za kategoriju "Gender" (pol) sa vrednostima "Female" i "Male", one-hot kodiranje će stvoriti dve nove kolone: "Gender\_Female" i "Gender\_Male". Ako je osoba ženskog pola, vrednost u koloni "Gender\_Female" će biti 1, dok će vrednost u koloni "Gender\_Male" biti 0, i obrnuto. Ovo omogućava da se kategoričke vrednosti predstave numerički, a da se ne unese nikakva inherentna hijerarhija ili numerička vrednost.

### Label kodiranje

Label kodiranje se primenjuje na ciljnu promenljivu, odnosno atribut koji treba da se predvedi. Ova tehnika dodeljuje jedinstven brojčani kod svakoj kategoričkoj vrednosti. Na primer, za ciljnu promenljivu "satisfaction" sa vrednostima "Satisfied" i "Neutral or dissatisfied", label kodiranje će dodeliti brojčane vrednosti 0 i 1, respektivno.

## Otklanjanje Anomalija

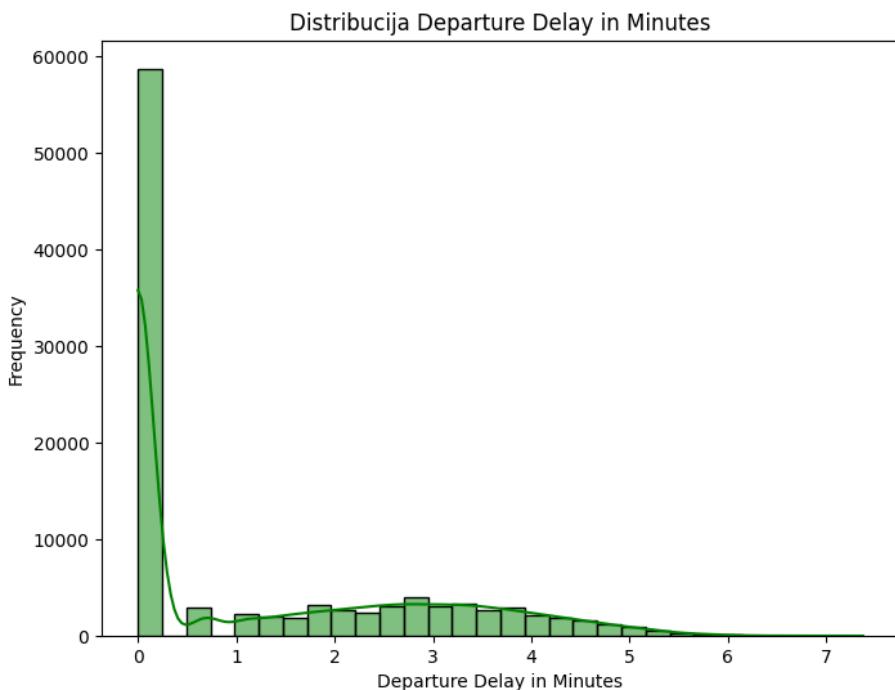
Anomalije su ekstremne vrednosti koje se značajno razlikuju od ostalih vrednosti u skupu podataka. U ovom koraku, koristimo **Z-Score metodu za otkrivanje anomalija** u numeričkim kolonama. Z-Score je mera udaljenosti između svake vrednosti i srednje vrednosti u odnosu na standardnu devijaciju. Vrednosti koje su više od 3 standardne devijacije udaljene od srednje vrednosti smatraju se anomalijama. Nakon što su sve anomalije prebrojane, rezultati su sledeći, `Age : 17, Flight Distance : 58, Departure Delay in Minutes : 2222`. Primeti se da najviše anomalija ima u koloni *Departure Delay in Minutes*. Da bi se odredila tehnika otklanjanja anomalija, iscrtava se grafik **distribucije podataka** u ovoj koloni.



Rezultat ovog plota pokazuje da raspodela vrednosti kolone "Departure Delay in Minutes" ima visok stepen asimetrije i prisustvo nekoliko ekstremnih vrednosti sa visokim kašnjenjem pri polasku. Ove ekstremne vrednosti mogu imati značajan uticaj na analizu ili modeliranje.

Sada, da bi se smanjio uticaj ovih ekstremnih vrednosti, primenjuje se **logaritamska transformacija**. To se radi kako bi se "ravnomerno" raspodelile vrednosti i smanjio njihov uticaj na analizu. Kada se primeni logaritamska transformacija, ekstremne vrednosti će biti smanjene, dok će manje vrednosti biti blago povećane.

Nakon izvršene transformacije broj anomalija za kolonu *Departure Delay in Minutes* jeste **54**. A distribucija sada izgleda ovako:



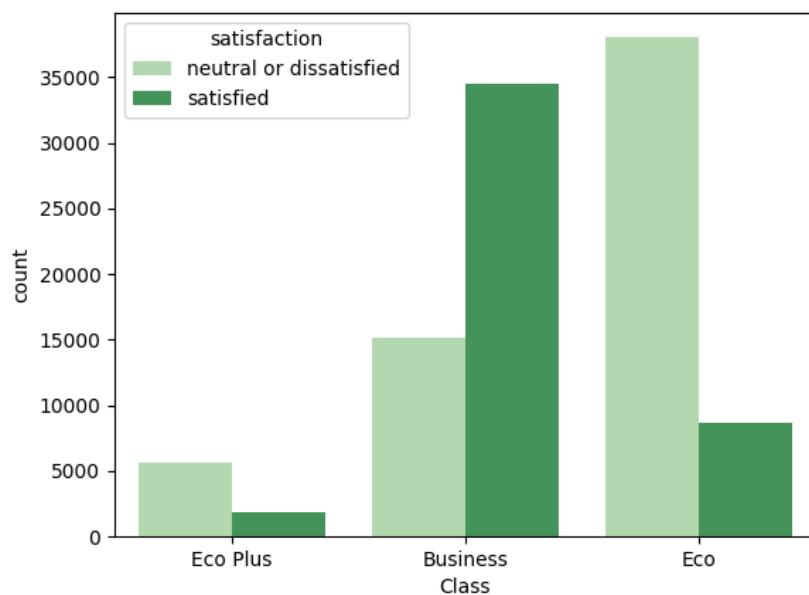
## Kolone Nakon Preprocesiranja

Age, Flight Distance, Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, Cleanliness, Departure Delay in Minutes, satisfaction, Gender\_Female, Gender\_Male, Customer Type\_Loyal Customer, Customer Type\_disloyal Customer, Type of Travel\_Business travel, Type of Travel\_Personal Travel, Class\_Business, Class\_Eco, Class\_Eco Plus

## Eksplorativna Analiza Skupa

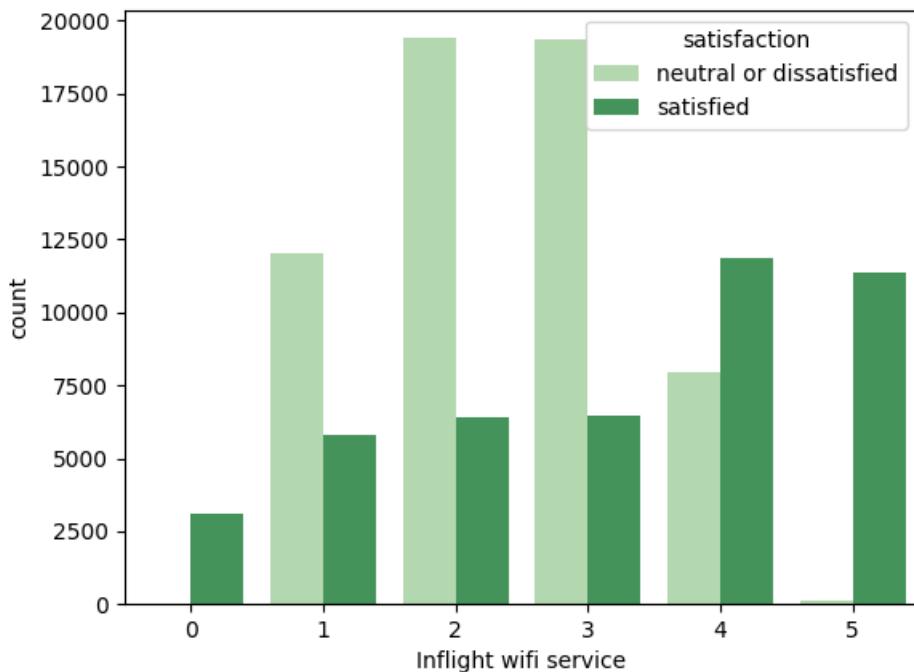
U eksplorativnoj analizi skupa podataka istražuju se i vizualizuju različite karakteristike i obrasci u podacima kako bi se steklo dublje razumevanje skupa podataka. Ova analiza omogućava da se otkriju veze, trendovi i nepravilnosti u podacima pre nego što se upusti u dalju obradu ili primenu modela.

### Zadovoljstvo putnika u odnosu na klasu letenja



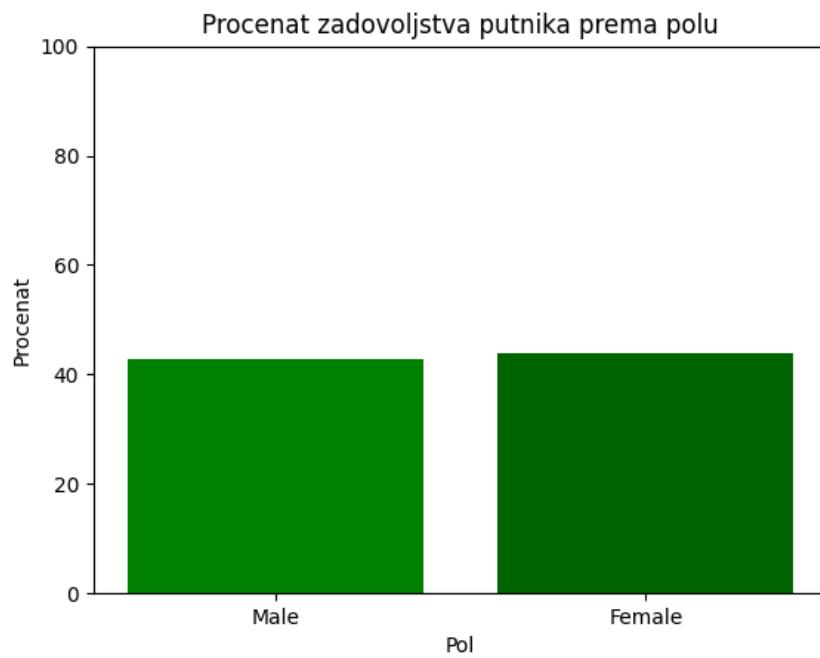
Većina putnika koji su leteli economy ili economy plus klasom izražavaju nezadovoljstvo, dok su putnici biznis klase uglavnom zadovoljni.

## Zadovoljstvo putnika u odnosu na ocenu wifi-ja



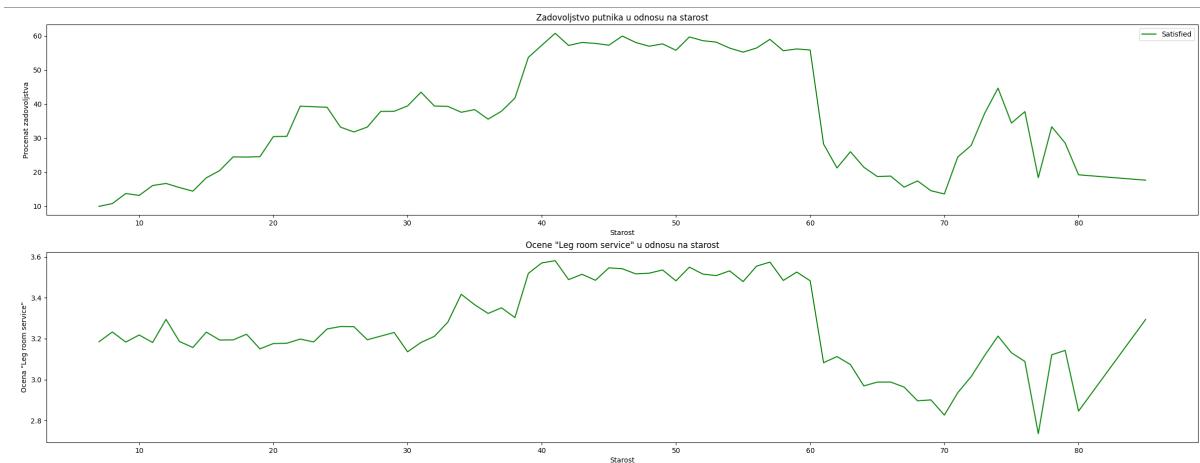
Primećuje se da su svi putnici koji su dali ocenu za internet 5 zadovoljni, ali je takođe interesantno da niko nije nezadovoljan od putnika koji su dali ocenu 0.

## Zadovoljstvo putnika prema polu



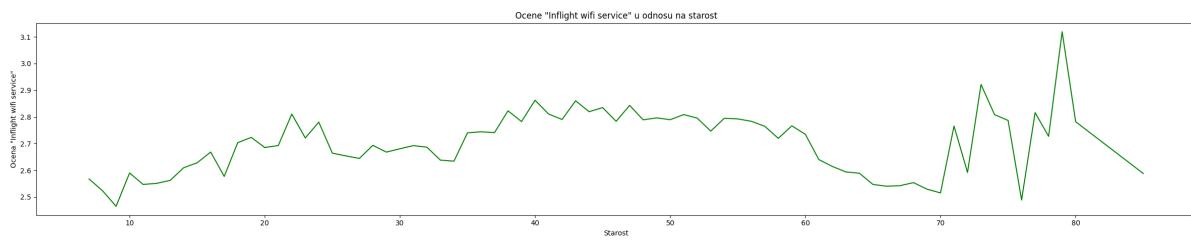
Grafik pokazuje da pol nema skoro nikakve veze sa zadovoljstvom putnika, na šta ćemo se vratiti kod selekcije atributa.

## Leg room vs Starost vs Zadovoljstvo putnika



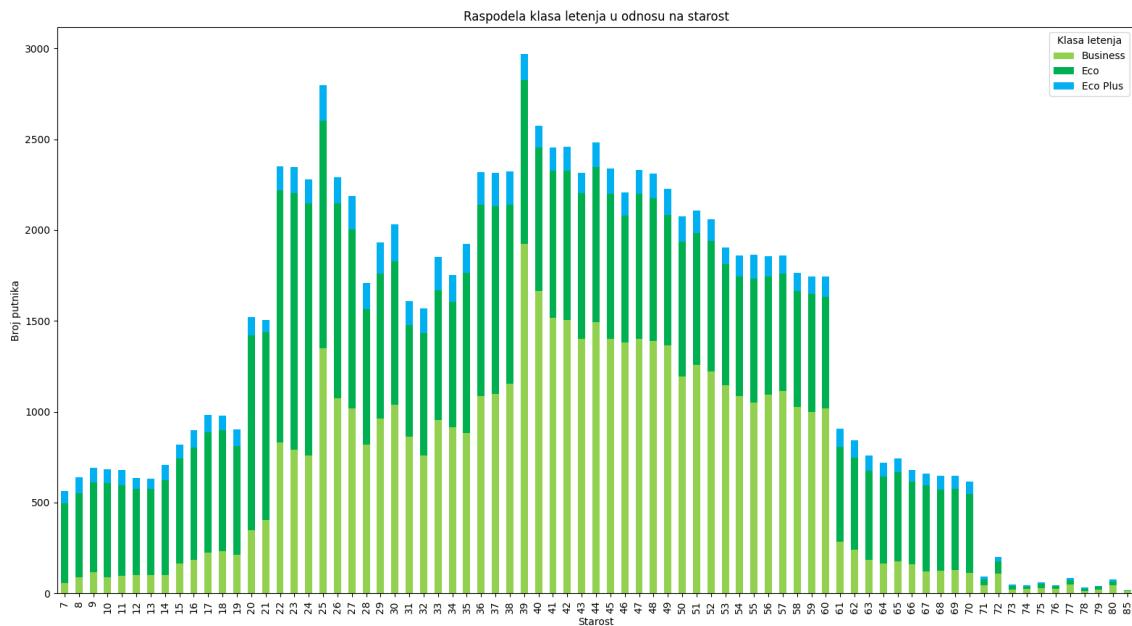
Na prvom grafiku prikazano je zadovoljstvo putnika u odnosu na godine, dok je na drugom grafiku prikazano zadovoljstvo putnika prostorom za noge u odnosu na godine. Primećuje se da je najmlađim putnicima prostor za noge nešto što najmanje utiče na samo zadovoljstvo letenjem.

## Zadovoljstvo wifi-jem u odnosu na godine



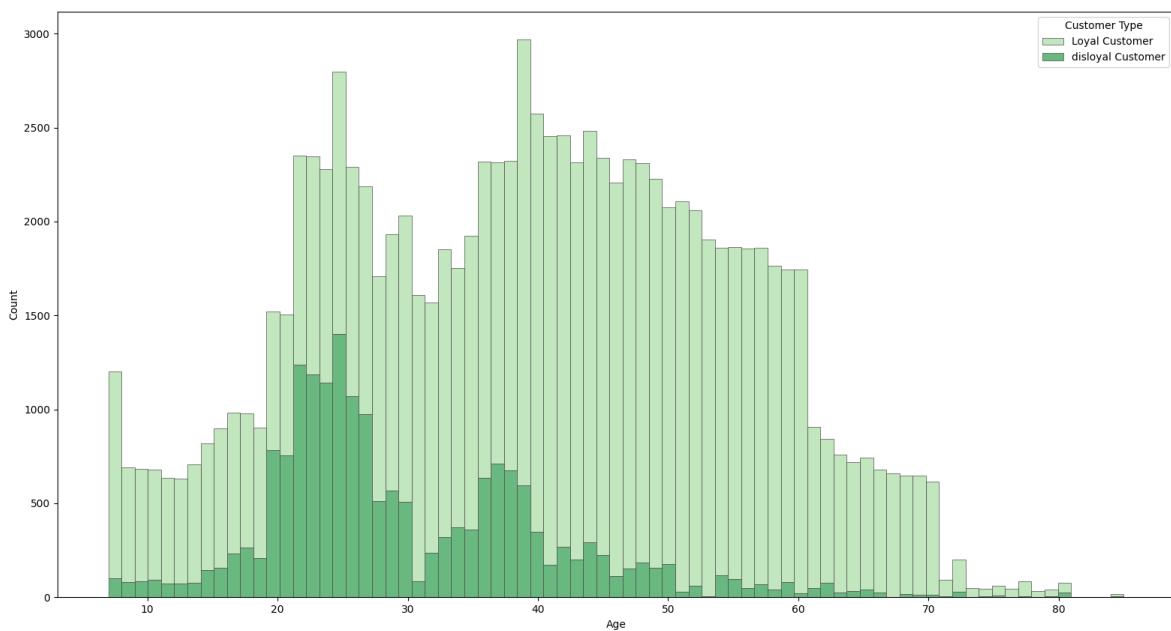
Primećuje se da su najmlađi putnici najmanje zadovoljni sa internetom dok su najstariji najviše zadovoljni. Odnosno da internet najviše koriste najmlađi putnici, što i ima smisla.

## Raspodela klasa letenja u odnosu na starost



Zaključak je da ljudi srednjih godina najviše lete biznis klasom, dok mlađi ljudi najviše lete eco klasom.

## Lojalnost mušterija naspram godina



Najlojalnije mušterije su ljudi srednjih godina, isti ljudi koji najviše lete biznis klasom. Takođe je interesantno da su sve mušterije koje imaju preko 50 godina uglavnom lojalne. Najviše nelojalnih mušterija ima među mladim ljudima.

## Modelovanje

### Izbor Modela

U ovom delu za model uzeto je četiri opcije u razmatranje. **LogisticRegression**, **KNeighborsClassifier**, **RandomForestClassifier**, **DecisionTreeClassifier**. Za svaki od ovih modela izvršeno je testiranje i unakrsna validacija. Rezultati su sledeći:

#### Logistička Regresija

Tacnost: 0.8332306744687403  
 Preciznost: 0.8005099872503187  
 Odziv: 0.8259230027185829  
 F1 score: 0.813017955801105

#### Unakrsna validacija

- Rezultat za fold 1:  
0.8376401520619797
- Rezultat za fold 2:  
0.8453876136855781

#### K Najbližih Suseda

Tacnost: 0.8083615645210964  
 Preciznost: 0.8030374492972361  
 Odziv: 0.7465579233535035  
 F1 score: 0.7737684057444101

#### Unakrsna validacija

- Rezultat za fold 1:  
0.8076608440402291
- Rezultat za fold 2:  
0.8059284923728406

- Rezultat za fold 3:  
0.8386025696549733
- Rezultat za fold 4:  
0.8429815697030941
- Rezultat za fold 5:  
0.8376323387872955
- Prosečna tačnost:  
0.8404488487785841

## Random Forest

Tacnost: 0.9637357560825377  
 Preciznost: 0.973734263200797  
 Odziv: 0.9428220643690257  
 F1 score: 0.9580288718588488

### Unakrsna validacija

- Rezultat za fold 1:  
0.9618882633174535
- Rezultat za fold 2:  
0.962850680910447
- Rezultat za fold 3:  
0.9618882633174535
- Rezultat za fold 4:  
0.964198065540638
- Rezultat za fold 5:  
0.9645332050048123
- Prosečna tačnost:  
0.9630716956181609

## Matrice konfuzije

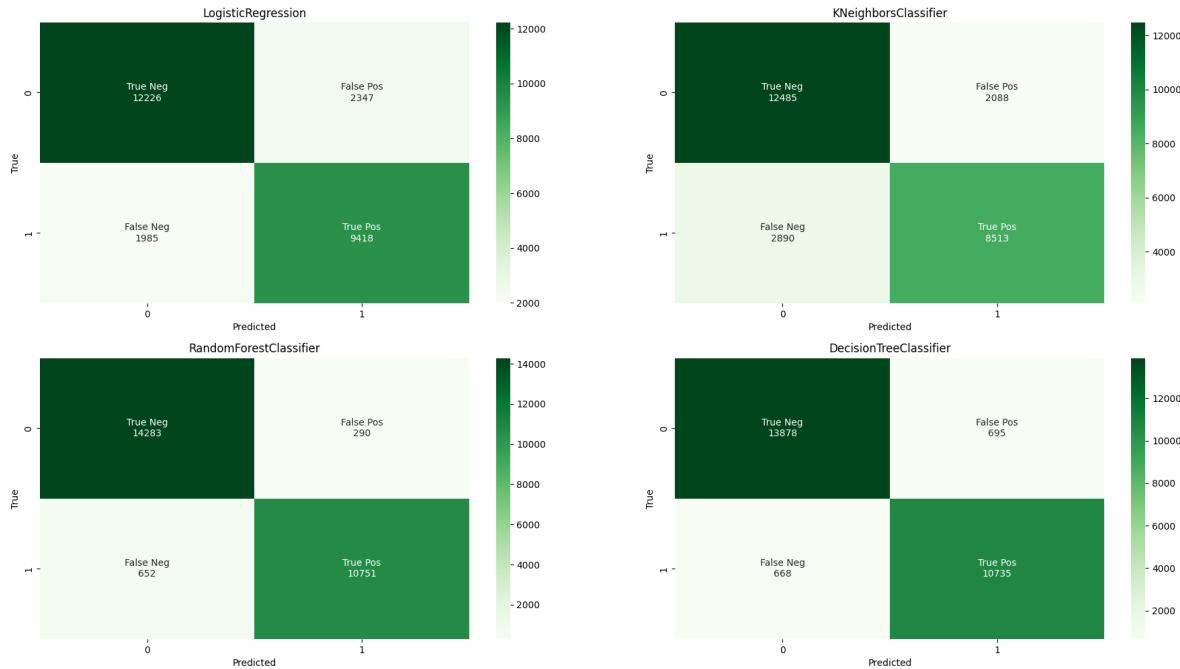
- Rezultat za fold 3:  
0.8023194263991146
- Rezultat za fold 4:  
0.810836822097108
- Rezultat za fold 5:  
0.813618864292589
- Prosečna tačnost:  
0.8080728898403763

## Stablo Odluke

Tacnost: 0.8083615645210964  
 Preciznost: 0.8030374492972361  
 Odziv: 0.7465579233535035  
 F1 score: 0.7737684057444101

### Unakrsna validacija

- Rezultat za fold 1:  
0.9415331312256388
- Rezultat za fold 2:  
0.9464895818295559
- Rezultat za fold 3:  
0.9461046147923584
- Rezultat za fold 4:  
0.9449015928011164
- Rezultat za fold 5:  
0.9463907603464871
- Prosečna tačnost:  
0.9450839361990313



Očito je da *Random Forest* i *Decision Tree* daju najbolje rezultate tako da su oni uzeti u dalje razmatranje i analizu.

## Podešavanje Hiperparametara

U ovom delu projekta, porede se performanse Random Forest i Decision Tree modela i vrši se podešavanje hiperparametara kako bi se postigli bolji rezultati.

Prvo su definisani parametri koje želimo da istražimo za svaki model. Za Random Forest model, definisani su parametri `'n_estimators'` (broj estimatora), `'max_depth'` (maksimalna dubina stabla) i `'min_samples_leaf'` (minimalan broj instanci u listu čvora). Za Decision Tree model, definisani su parametri `'criterion'` (kriterijum za merenje kvaliteta podela), `'ccp_alpha'` (parametar za postpruning) i `'max_depth'` (maksimalna dubina stabla).

Zatim je korišćena GridSearchCV funkcija da bi se pretražio prostor parametara i pronaše najbolje kombinacije za svaki model. GridSearchCV se koristii sa 5-fold unakrsnom validacijom.

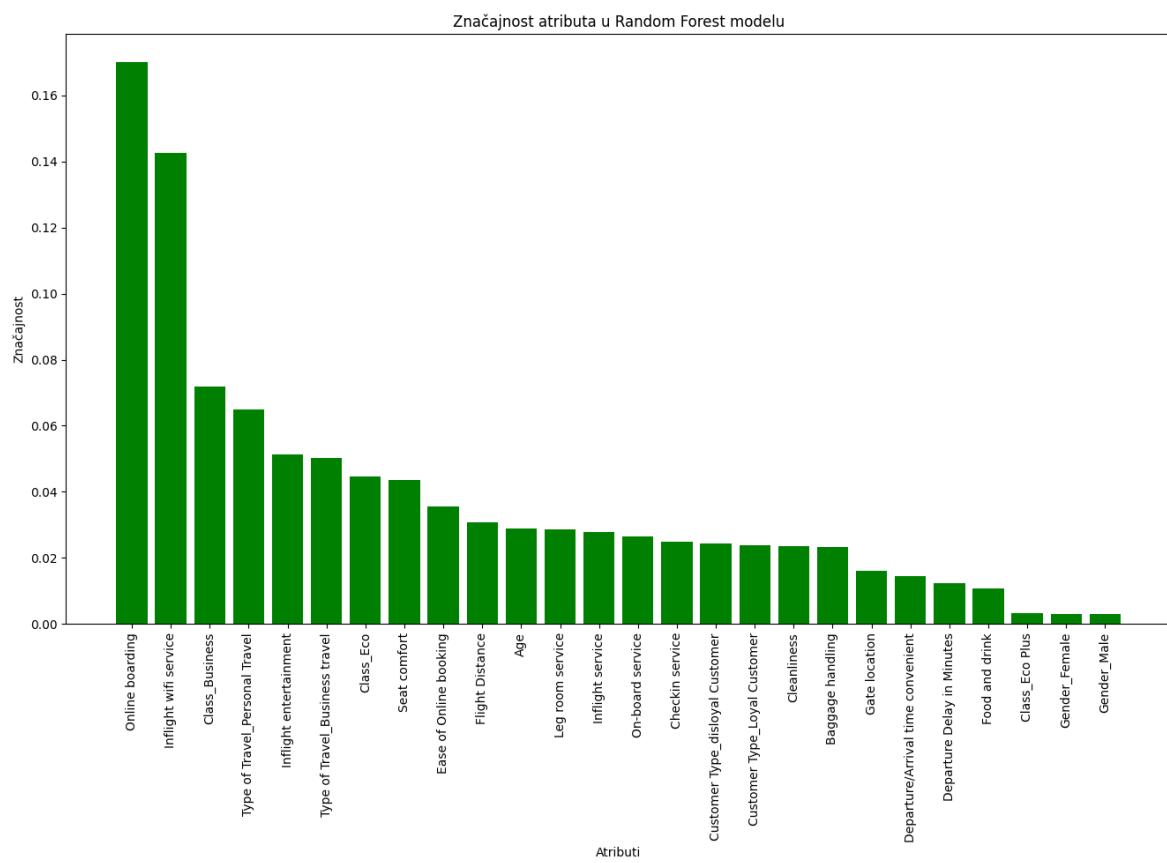
Nakon pretrage parametara, upoređena je tačnost modela pre i posle podešavanja hiperparametara. Ako je tačnost posle podešavanja veća od početne tačnosti, smatramo da je podešavanje bilo uspešno.

Nažalost, u ovom slučaju podešavanje parametara za Decision Tree nije bilo uspešno. Međutim, za Random Forest model smo dobili rezultate. Mada i posle podešenih parametara **Random Forest** je i dalje imao veću tačnost, tako da je on

uzet kao **pobednički model**. Finalni rezultat je bio, Decision Tree: 0.95 vs Random Forest: 0.96.

## Odabir Značajnih Atributa

U ovom delu projekta vrši se izračunavanje značajnosti atributa u Random Forest modelu kako bi se identifikovali najvažniji atributi za predviđanje zadovoljstva putnika. Koristii se `feature_importances_` atribut Random Forest modela koji daje značajnost svakog atributa.



Sa grafika se primećuje da *Online Boarding* i *Inflight Wifi Service* najviše utiču na zadovoljstvo putnika, dok pol putnika i hrana tokom leta najmanje utiču. Na ovaj način izvršena je selekcija atributa, gde su *Gender Male*, *Gender Female*, *Food and drink* izbačeni.

Nakon toga tačnost modela je testirana još jednom, i rezultat je da nije došlo do značajne promene. Tačnije 0.9636202648598706 (**posle**) vs 0.9636202648598706 (**pre**)

# Zaključak

U okviru ovog projekta, uspešno smo istražili i analizirali skup podataka o zadovoljstvu putnika avio-kompanije. Primenili smo različite tehnike obrade podataka, uključujući kodiranje kategoričkih atributa, otklanjanje anomalija i selekciju značajnih atributa. Takođe smo eksplorativno analizirali podatke kako bismo dobili uvid u odnose između različitih atributa i zadovoljstva putnika.

Nakon toga, trenirali smo četiri različita modela mašinskog učenja - LogisticRegression, KNeighborsClassifier, RandomForestClassifier i DecisionTreeClassifier. Evaluirali smo performanse tih modela koristeći različite metrike kao što su tačnost, preciznost, odziv i F1 mera.

Na osnovu rezultata, zaključili smo da je model RandomForestClassifier najbolji za naš problem klasifikacije zadovoljstva putnika. Model je pokazao visoku tačnost i preciznost, kao i solidan odziv. Ovo ukazuje na to da je model sposoban da tačno identificuje zadovoljstvo putnika na osnovu dostupnih atributa.

## Finalne performanse

**Tačnost od 96.36%** ukazuje na to da je model RandomForestClassifier veoma precisan u predviđanju zadovoljstva putnika. Ovo znači da je veliki broj predikcija tačno klasifikovan.

**Preciznost od 97.34%** predstavlja udeo pravilno predviđenih pozitivnih rezultata (zadovoljstvo putnika) u ukupno predviđenim pozitivnim rezultatima. Ovo znači da je model vrlo dobar u identifikovanju stvarno zadovoljnih putnika.

**Odziv od 94.29%** predstavlja udeo pravilno predviđenih pozitivnih rezultata u stvarno pozitivnim rezultatima. Ovo ukazuje na sposobnost modela da identificuje zadovoljstvo putnika sa visokom pouzdanošću.

**F1 score od 0.9579** predstavlja harmoničnu sredinu između preciznosti i odziva. Ova mera uzima u obzir i lažno pozitivne i lažno negativne rezultate. Visok F1 score ukazuje na dobar balans između tačnosti i pouzdanosti modela.

**Matrica konfuzije** za model RandomForestClassifier izgleda ovako:

- True Negatives (TN): 14279
- False Positives (FP): 294
- False Negatives (FN): 651
- True Positives (TP): 10752

Ova matrica prikazuje broj tačno predviđenih negativnih i pozitivnih rezultata. U ovom slučaju, imamo 14279 tačno predviđenih negativnih rezultata (putnici nisu zadovoljni) i 10752 tačno predviđenih pozitivnih rezultata (putnici su zadovoljni). Takođe imamo 294 lažno pozitivna rezultata (putnici su pogrešno predviđeni kao zadovoljni) i 651 lažno negativnih rezultata (putnici su pogrešno predviđeni kao nezadovoljni).