PRETPROCESIRANJE SKUPA PODATAKA

Dobavljanje podataka

Skup podataka koji smo koristili u projektu nalazi se na sajtu kaggle, dostupan je na sledećem linku: https://www.kaggle.com/code/fabiendaniel/predicting-flight-delays-tutorial. Ovaj skup podataka sadrži informacije o letovima, uključujući otkazivanja i pomjeranja letova za 2015 godinu.

Preuzeti skup podataka sadrži 38 kolona. Ciljno obilježje je kolona ARRIVAL_DELAY koja predstavlja razliku u minutama između zakazanog i stvarnog vremena dolaska aviona. Za avione koji nisu kasnili vrijednost ovog obilježja je 0 ili negativna vrednost ako su stigli pre zakazanog vremena. Ukupan broj instanci je 5 819 079 od čega je 56.33% letova imalo kašnjenje.

Ovaj skup podataka se sastoji od 3 csv fajla od kojih su nama bili potrebni flights.csv airports.csv koji sadrže podatke o avionima i aerodromima. Ove fajlove smo spojili u jedan koristeći zajedničku kolonu IATA CODE.

Podaci o avionima (data scraping)

Dodatni podaci o avionima koji bi mogli uticati na predikciju su prikupljani sa sajta https://flightaware.com. To podrazumeva informacije o tipu aviona i starosti aviona. Starost aviona smo računali na osnovu datumu prvog leta aviona kada nam podatak o starosti nije bio dostupan.

Podaci o vremenskim uslovima (API)

Podaci o vremenskim uslovima za polazni i dolazni aerodrom iz 2015. godine su prikupljeni sa sajta https://open-meteo.com. Ovaj sajt smo odabrali jer nam omogućava besplatno dobavljanje istorijskih i budućih vremenskih uslova za polazni i dolazni aerodrom. Dobavljeni su sledeći podaci:

- WCODE - vremenski kod
- TEMPERATURE_2M_MAX - maksimalna temperatura
- TEMPERATURE_2M_MIN - minimalna temperatura
- APPARENT_TEMPERATURE_MAX - prividna maksimalna temperatura
- APPARENT_TEMPERATURE_MIN - prividna minimalna temperatura
- PRECIPITATION_HOURS - sati padavine
- SUNRISE - izlazak sunca
- SUNSET - zalazak sunca
- WINDSPEED_10M_MAX - brzina vetra
- WINDGUSTS_10M_MAX - udari vetra
- WINDDIRECTION_10M_DOMINANT_10M_MAX - pravac vetra

Čišćenje podataka

Nakon dobavljanja podataka, sledeći korak je čišćenje podataka, što podrazumeva:

- Uklanjanje nedostajućih vrednosti
- Smanjivanje skupa podataka
- Kategorizacija ciljnog obeležja
- Uklanjanje outlier-a

Uklanjanje nedostajućih vrednosti

S obzirom, da je preuzeti skup podataka imao preko 5 miliona redova, odlučili smo se da redove koje imaju nedostajuće vrednosti za ciljno obeležje jednostavno uklonimo iz skupa podataka.

Ciljno obeležje nam je ARRIVAL_DELAY, tako da smo prolazili kroz skup podataka i brisali one redove gde je to obeležje nedostajalo.

Smanjivanje skupa podataka

Nakon uklanjanja redova sa nedostajućim vrednostima, skup podataka se nije značajno smanjio, odnosno i dalje je broj redova iznosio blizu 5 miliona. Uzimajući u obzir obradu podataka (data scraping, pozivanje API-ja za dobavljanje podataka o vremenskim uslovima…) bilo bi nam potrebno mnogo vremena za ovu veličinu skupa podataka. Tako da smo se odlučili da skup podataka smanjimo sa 5 miliona na oko 10 000 uzoraka.

Što se tiče strategije za smanjivanje skupa podataka, odlučili smo se da koristimo strategiju uzimanja svakog k-tog uzorka iz skupa podataka. Da bi dobijeni skup podataka bio prava slika većeg skupa od kojeg je dobijen, bilo je potrebno sortirati polazni skup po vrednostima ciljnog obeležja, što je kod nas bilo vreme dolaska aviona (ARRIVAL_DELAY).

Nakon sortiranja skupa podataka, da bi smo dobili 10 000 uzoraka bilo je potrebno da uzimamo svaki 500-ti uzorak iz sortiranog skupa podataka.

Kategorizacija ciljnog obeležja

Nakon smanjivanja skupa podataka, primetili smo da nam ciljno obeležje ima veliki raspon vrednosti (**npr. od -10 do 200**). Ovo bi moglo da utiče na računanje korelacije između ciljnog i ostalih obeležja, a samim tim i na rezultate predikcije, tako da smo se odlučili da izvršimo kategorizaciju ciljnog obeležja zarad dobijanja što boljih rezultata.

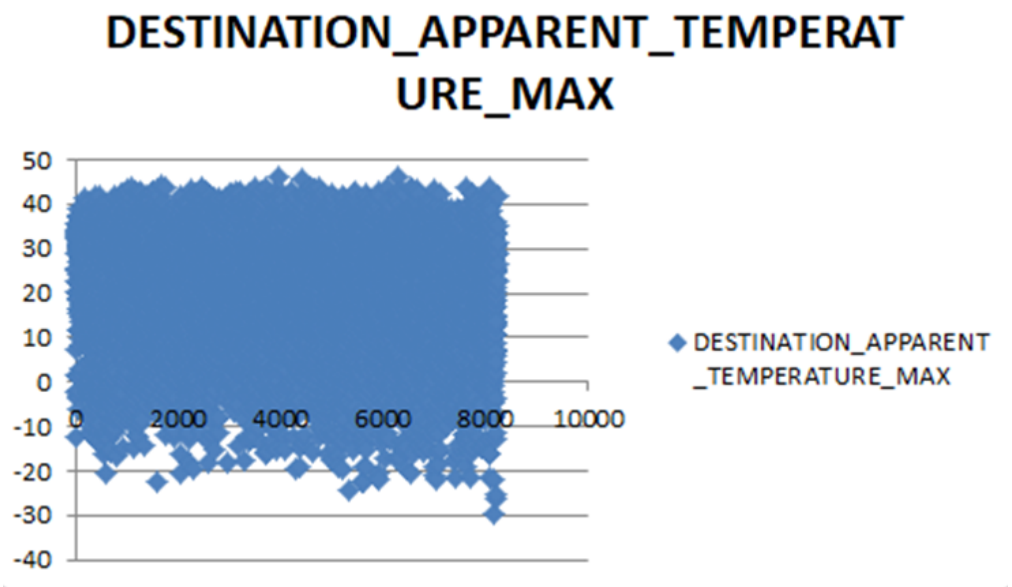
Vrednosti ciljnog obeležja smo podelili u tri kategorije:

- 0 kategorija - zanemarljivo kašnjenje (< 15 min)
- 1 kategorija - malo kašnjenje (15 - 60 min)
- 2 kategorija - veliko kašnjenje (> 60 min)

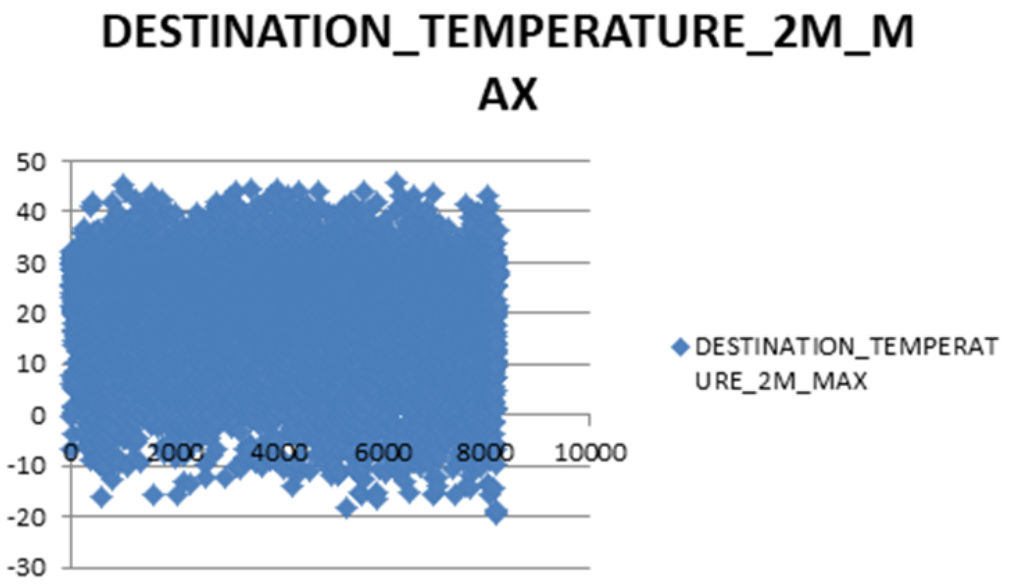
Uklanjanje oullier-a

Nakon kategorizacije ciljnog obeležja, primetili smo da nam ostala obeležja imaju vrednosti koje dosta odskaču od ostalih, odnosno kada smo nacrtali scatter grafika pojedinih obeležja primetili smo da postoje oulier-i. Tako da je naš sledeći korak bio uklanjanje tih outlier-a.

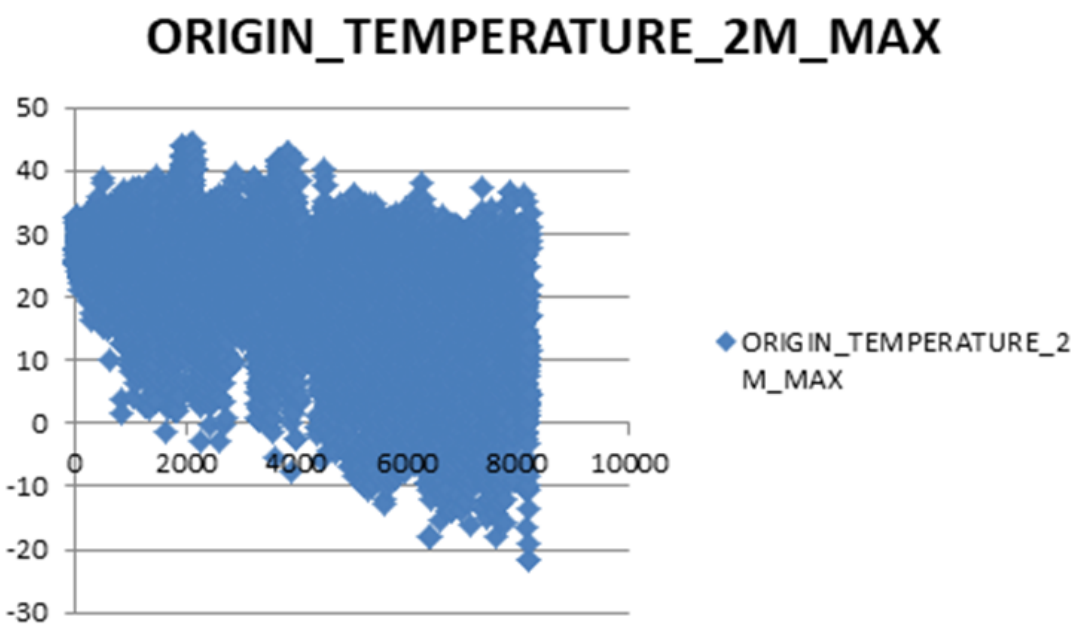
Na sledećim slikama su prikazani scatter grafici obeležja iz kojih smo se odlučili da uklonimo outlier-e.



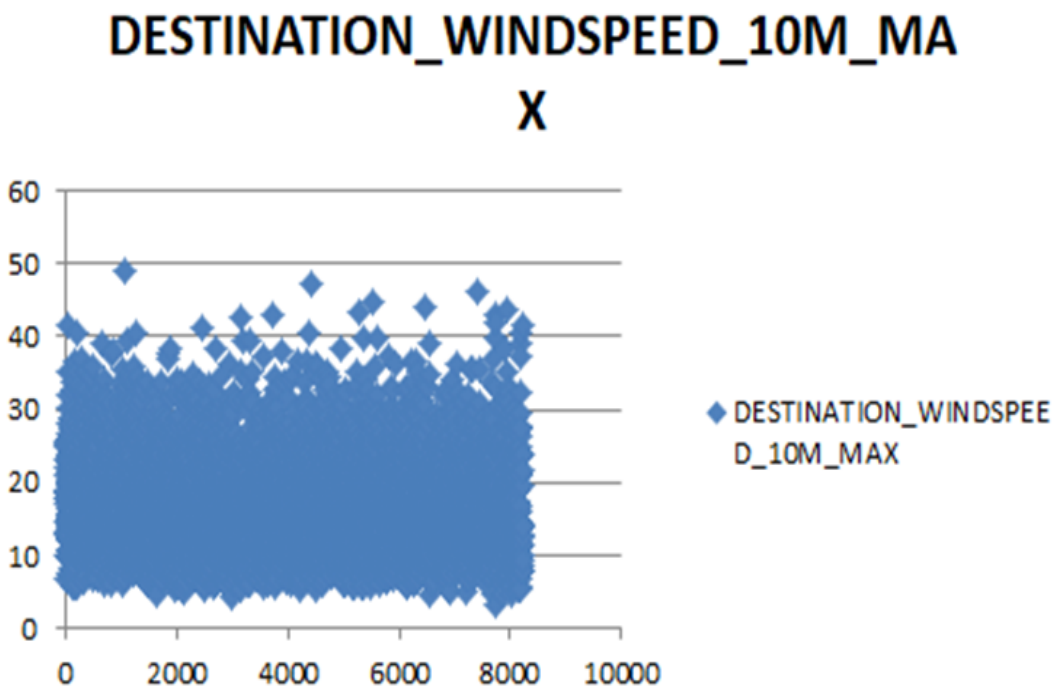
Slika 1



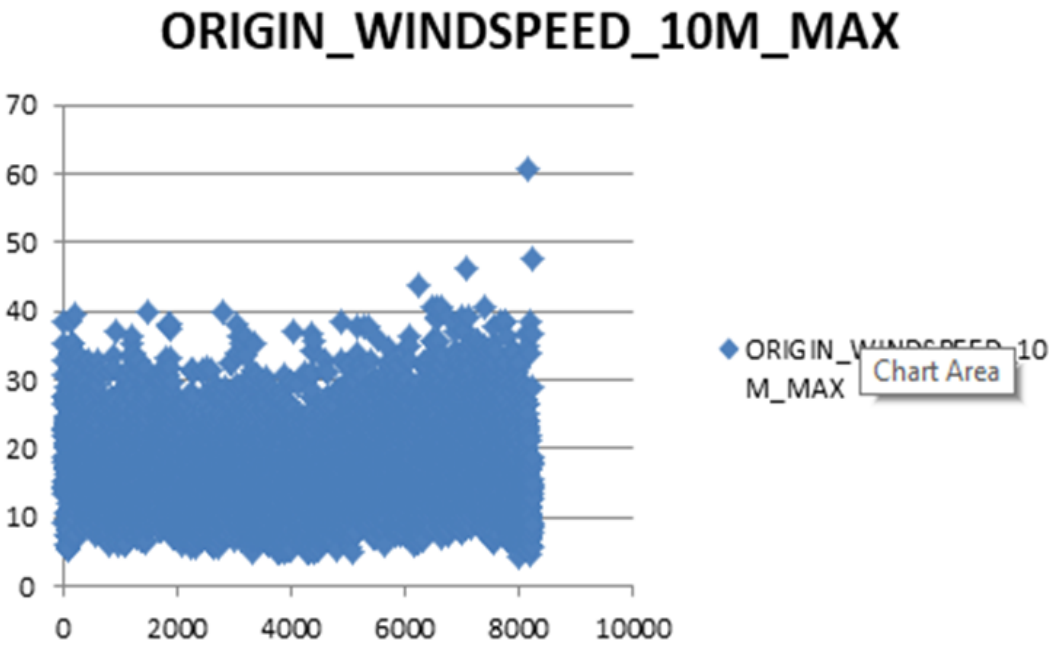
Slika 2



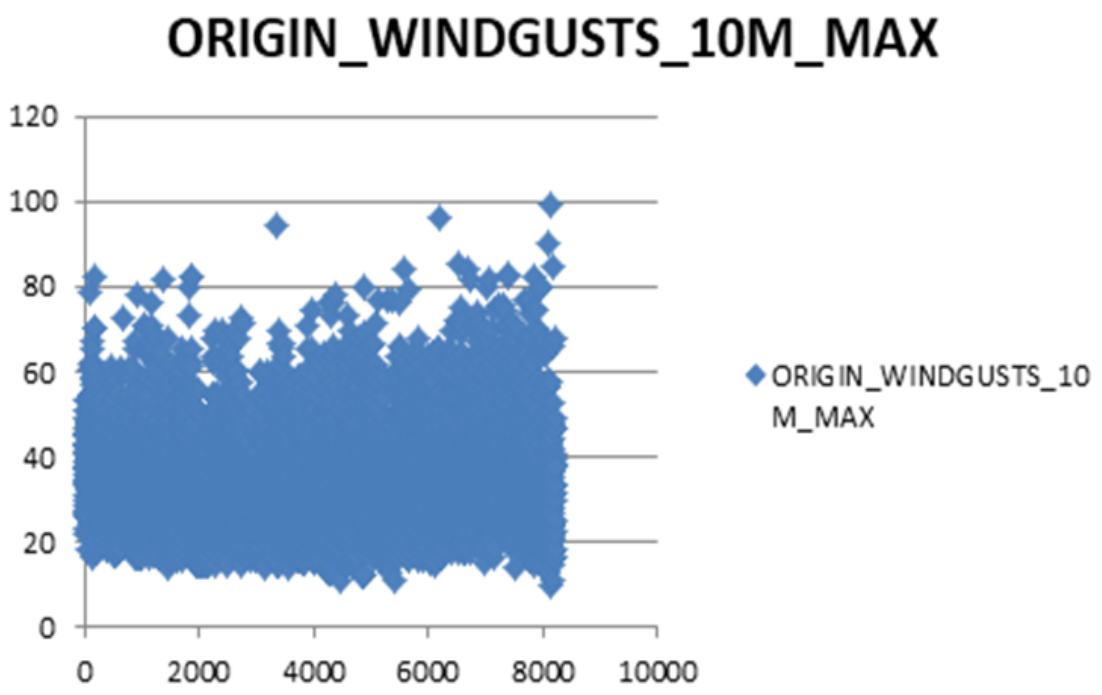
Slika 3



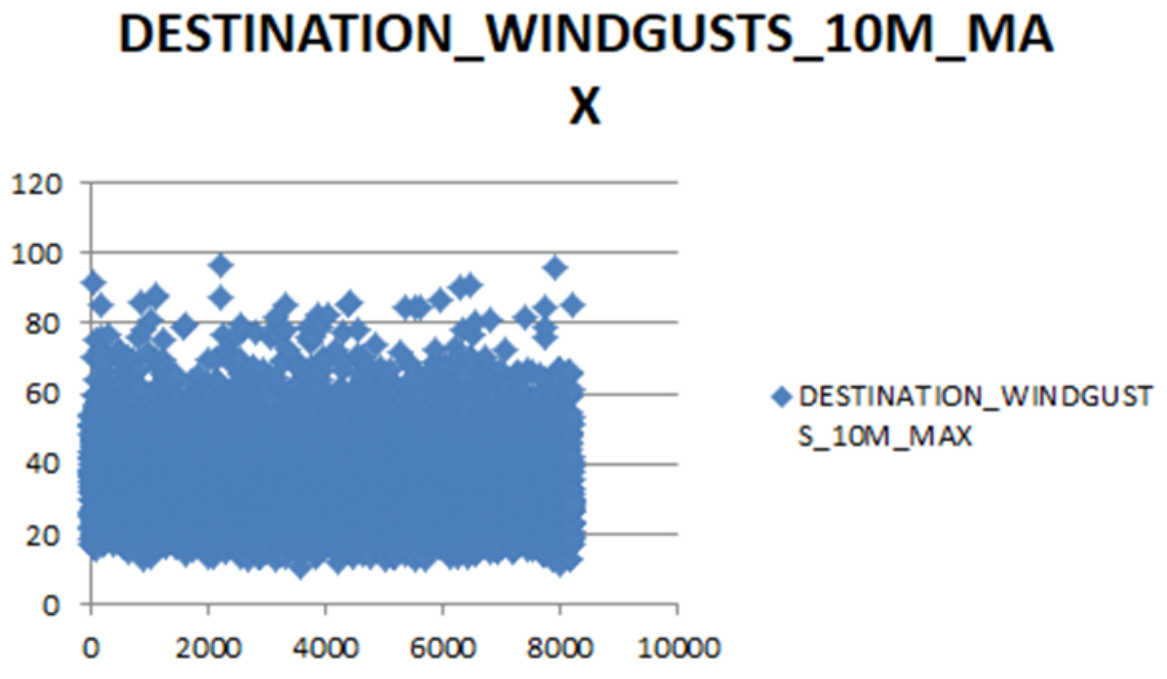
Slika 4



Slika 5



Slika 6

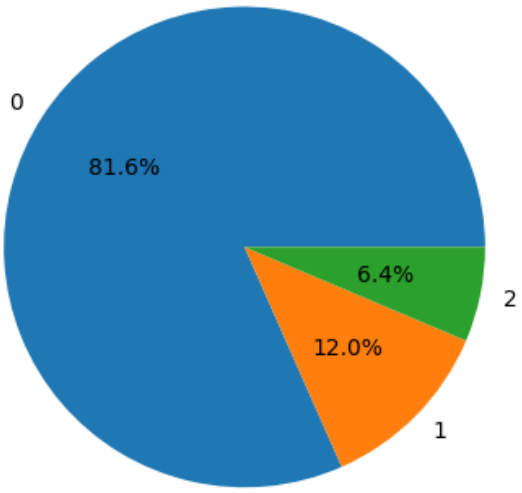


Slika 7

EKSPLORATIVNA ANALIZA

Prvi korak u sređivanju skupa podataka bila je kategorizacija ciljnog obeležja. Na sledećem grafiku (slika 8) nalazi se procentualna raspodela letova prema kategorijama kašnjenja. Svi letovi su podeljeni u tri kategorije:

- Zanemarljivo kašnjenje - kategorija 0 (< 15 min)
- Malo kašnjenje - kategorija 1 (15 min - 60 min)
- Veliko kašnjenje - kategorija 2 (> 60 min)

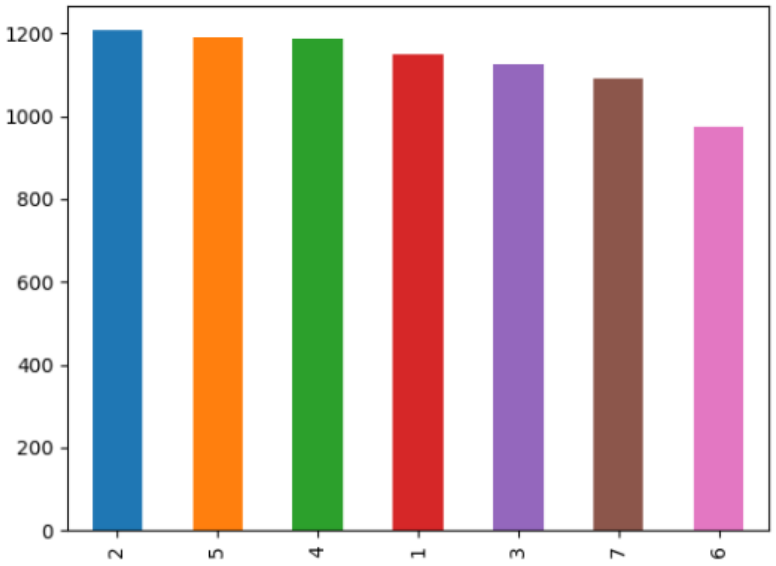


Slika 8

Na grafiku se vidi da je veći procenat letova koji su imali kašnjenje manje od 15 minuta. Što znači da je skup podataka neizbalansiran i da može uticati na metode klasifikacije.

Zavisnost kašnjenja u odnosu na vremensku komponentu

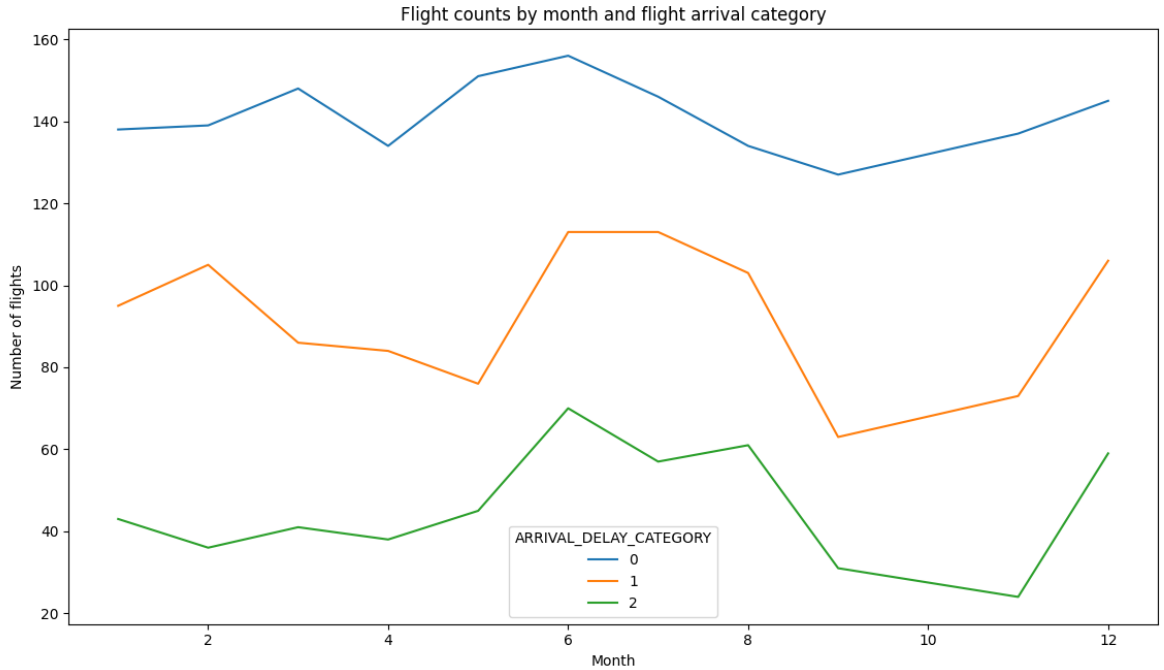
Na slici 9 nalazi se grafik koji pokazuje koliko je bilo kašnjenja za svaki dan u nedelji.
Na osnovu ovog grafika se može zaključiti da obilježje DAY_OF_WEEK ne utiče mnogo na predikciju kašnjenja, s obzirom da je broj kašnjenja približan za svaki dan u nedelji.



Slika 9

Legenda: 1 - ponedeljak, 2 - utorak, 3 - sreda, 4 - četvrtak, 5 - petak, 6 - subota, 7 - nedelja

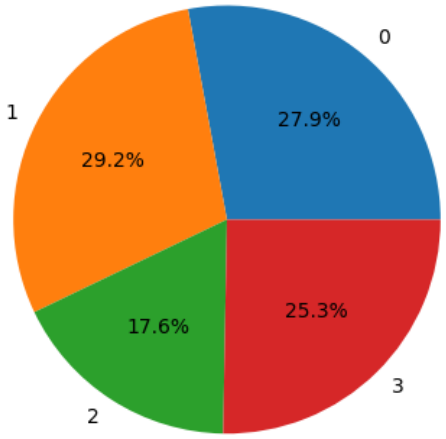
Na slici 10 je prikazan grafik na kome se vidi zavisnost između meseca u godini i broja kašnjenja letova u svakoj od kategorija kašnjenja. Možemo zaključiti da je u letnjim mesecima (6 - 9 meseca) najveći broj kašnjenje po svim kategorijama, pa smo iz tog razloga odlučili da obeležje MONTH bude uključeno u predikciju.



Slika 10

Legenda: 0 - zanemarljivo kašnjenje , 1 - malo kašnjenje , 2 - veliko kašnjenje

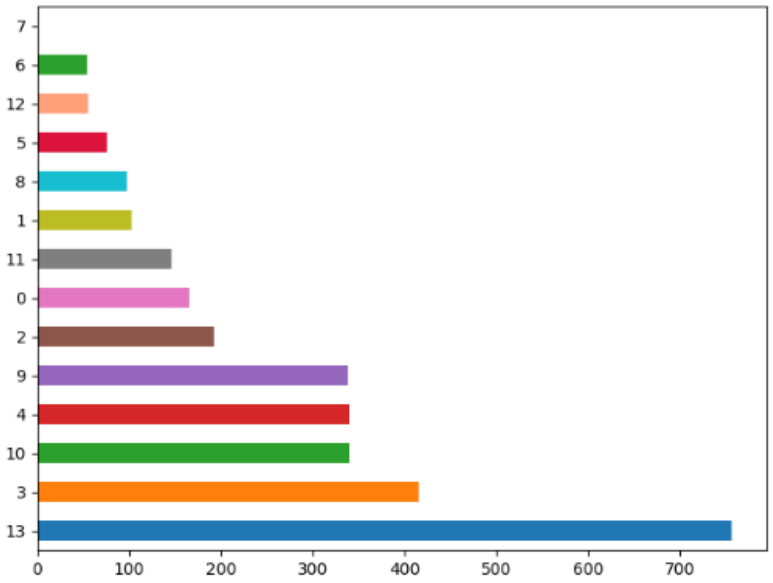
Na slici 11 je prikazan grafik koji pokazuje uticaj godišnjeg doba na kašnjenje letova. Na ovom grafiku se može videti da je najveći broj kašnjenja leti, što potvrđuje našu prethodnu tvrdnju.



Slika 11
Legenda: 0 - proljeće, 1 - ljeto, 2 - jesen, 3 - zima

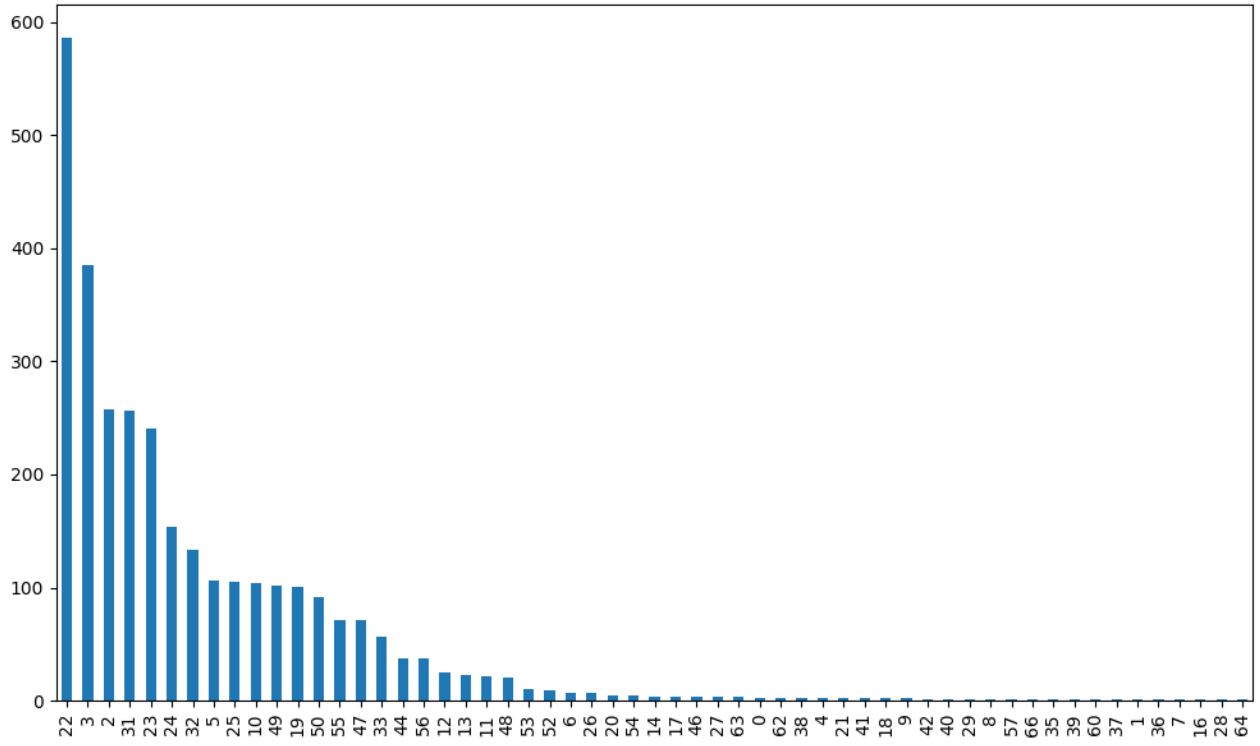
Uticaj avio-kompanija na kašnjenje letova

Na slici 12 prikazan je uticaj aviokompanije na kašnjenje aviona. Na grafiku se vidi da jedna kompanija ima ubjedljivo najviše kašnjenja, samim tim smo odlučili da obeležje AIRLINE uzmemo u obzir prilikom predikcije kašnjenja aviona.



Slika 12

Uticaj tipa aviona na kašnjenje letova

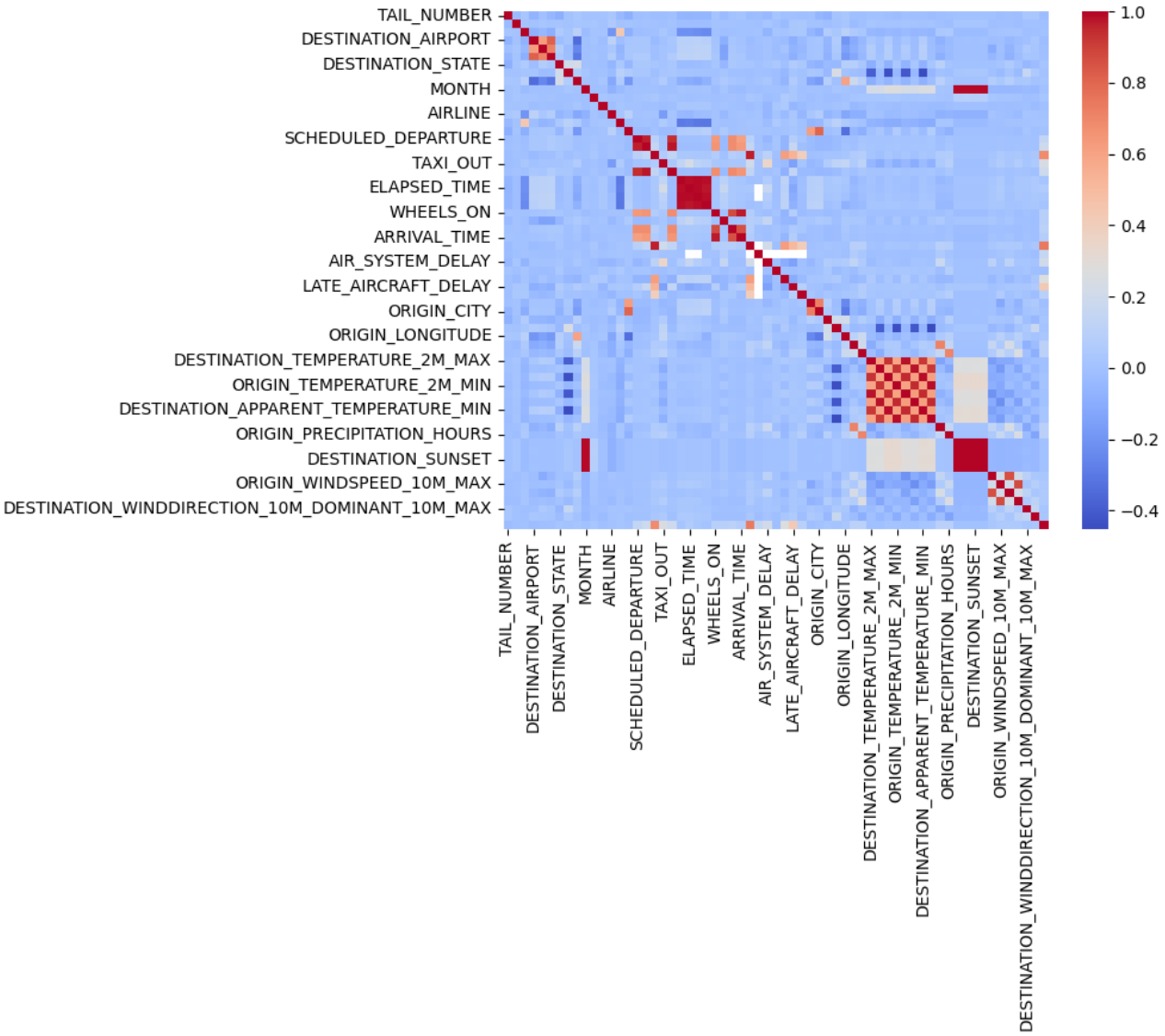


Slika 13 - Prikaz kašnjenja letova po tipu aviona

Na slici 13 se može videti grafik koji prikazuje broj letova koji kasne za svaki tip aviona, gde su na x osi prikazane numeričke oznake tipa aviona, a na y osi broj letova koji kasne. Na grafiku se jasno može videti da se nekoliko tipova aviona dosta izdvajaju od ostalih, odnosno da imaju mnogo veći broj letova koji kasne. Zbog toga smo se odlučili da uzmemo TYPE obeležje kao jedno od obeležja za treniranje modela, jer smatramo da može imati uticaja na predikciju kašnjenja letova.

Korelacije

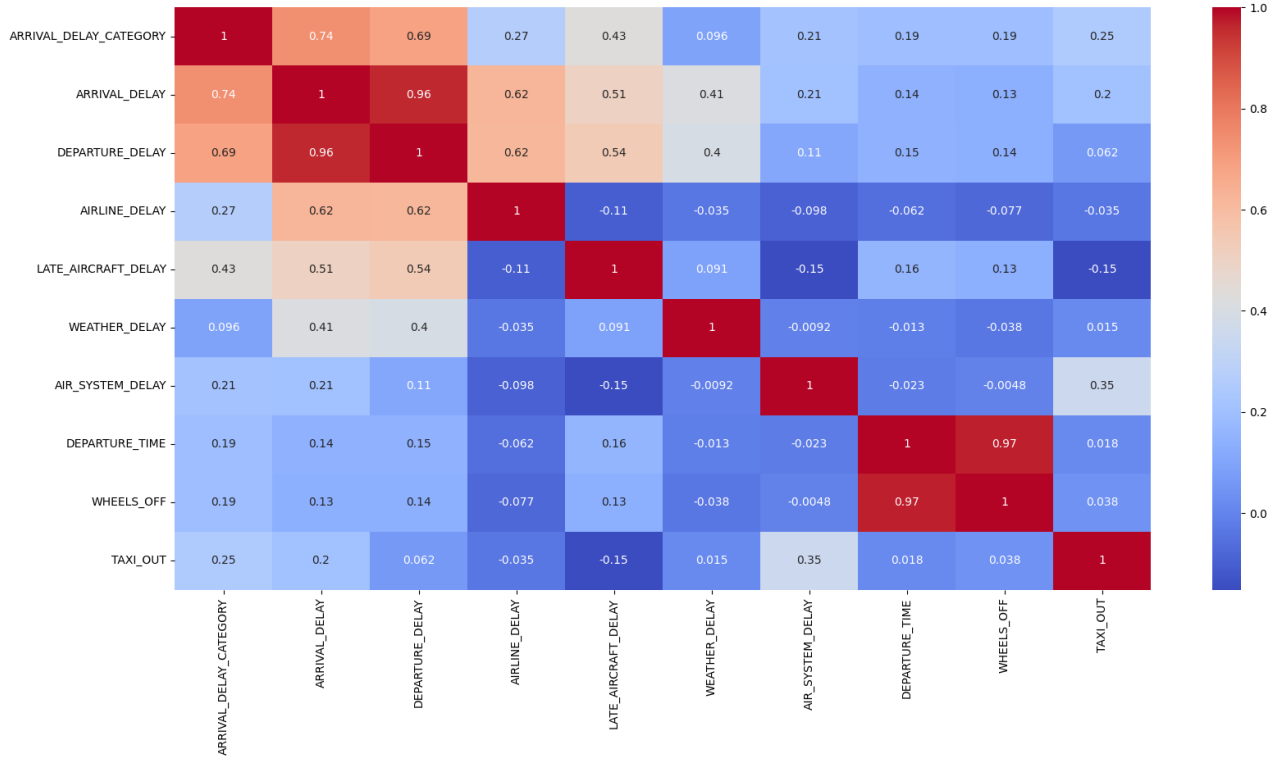
Kao sledeći korak u eksplorativnoj analizi, posmatramo zavisnost obeležja, kako bi utvrdili koja obeležja možemo iskoristiti za obučavanje modela. Zavisnost obeležja se može utvrditi posmatranjem vrednosti korelacije, odnosno analizom matrice korelacije početnog skupa obeležja koja je prikazana na slici 14.



Slika 14

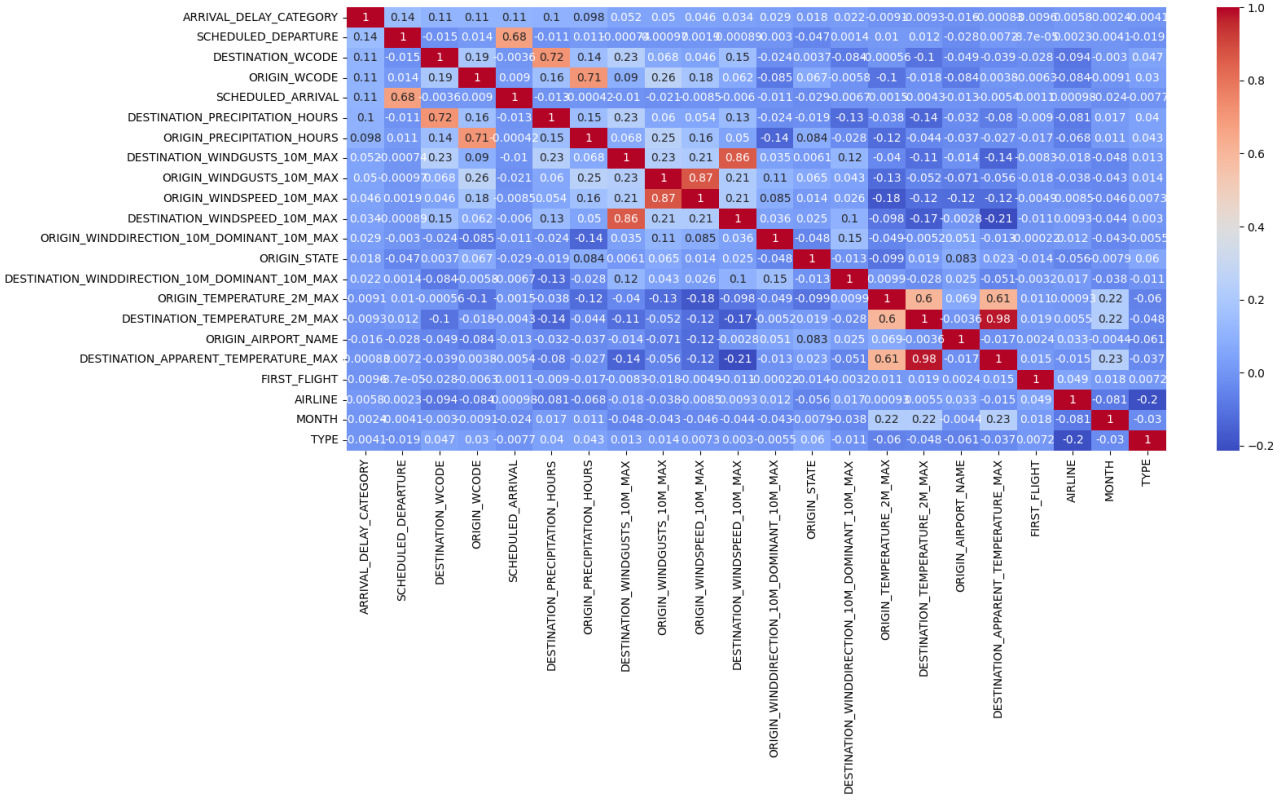
Analizom matrice korelacije se može zaključiti međusobna zavisnost između pojedinih obeležja. Na primer, obeležja koja u svom nazivu imaju reč DELAY su međusobno zavisna, što je logično jer su u pitanju različite vrste kašnjenja. Tako kašnjenje pri polasku utiče na kašnjenje pri dolasku, kašnjenje…

Na slici 15 vidi se matrica korelacije obilježja koja koja predstavljaju različite tipove kašnjenja, odnosno obeležja koja u svom nazivu imaju reč DELAY (kašnjenje). S obzirom na to da vrednosti većine tih obeležja nisu dostupne pri predikciji ne možemo ih koristiti za istu.



Slika 15

Na slici 16 je prikazana matrica korelacije za obilježja koja smo izabrali za predikciju. Obeležja smo odabrali na osnovu eksplorativne analize (MONTH, AIRLINE, TYPE,,..) i analizom numeričkih vrednosti korelacija sa ciljnim obeležjem.



Slika 16