Jernej Knific (63160164) **Naloga 1** 11. december 2020

Uvod. Cilj naloge je bil, najti način kako z linearno regresijo na podlagi podatkov voženj avtobusov LPP v letu 2012(brez podatkov za mesec december) čimbolje napovedati prihode avtobusov v mesecu decembru.

Predtekmovanje:

Podatki. Pri predtekmovanju podatki(train-pred.csv.gz) obsegajo podatke o vožnjah avtobusne linije 14 od januarja do novembra leta 2012. Vsak zapis obsega podatke o registerski tablici avtobusa, voznikovem ID-ju, številki proge, začetni postaji, času odhoda, končni postaji in času prihoda. Pri nalogi sem uporabil podatek o številki proge ter podatke o odhodu in prihodu avtobusa.

Napovedni model.

- **Dan v tednu** V stolpec na mestu tega atributa zapišem vrednost(1 - 7) glede na to kateri dan v tednu je.
- **Vikend** V stolpec na mestu tega atributa zapišem vrednost 1 če je dan del vikenda sepravi sobota ali nedelja, in 0 če ni.
- **Ura odhoda** V stolpec na mestu tega atributa zapišem vrednost(0 - 23) glede na uro odhoda.
- **Praznik** V stolpec na mestu tega atributa zapišem vrednost 0 če ta dan ni praznični in 1 če je. Ob praznikih je pričakovan čas vožnje krajši saj naj nebi bilo toliko gneče na cestah.
- **Mesec** V stolpec na mestu tega atributa zapišem vrednost(1 - 12) glede na to kateri mesec je.
- **Deli dneva** Dan sem razdelil na 5 delov in sicer na jutro(6 - 10), dopoldan(10 - 14), popoldan(14 - 18), večer(18 - 21) in noč(21 - 6). Glede na uro odhoda zapišem vrednost 1 v stolpec ki je določen za ta del dneva, v ostale pa vrednost 0. Ta atribut je močno vplival an rezultat saj je bo nekaterih delih dneva več gneče na cestah in so posledično tudi vožnje daljše.

Rezultati. Pri testiranju na lestvici sem najprej doesegel rezultat 162. Program sem nato izboljšal tako da sem dodal nova atributa in sicer atributa mesec in del dneva. Atribut del dneva je zelo vplival na rezultat, kar je tudi razumljivo saj je gneča na cesti različna glede na del dneva. Pomembna se mi zdita tudi atributa ura odhoda ter dan v tednu saj sta močno vplivala na rezultat. Z izboljšavo sem dosegel rezultata 151.

Tabela 1: Rezultati predtekmovanja

Datoteka	**Rezultati (.txt)**	**Ocena na lestvici**
predtekmovanje.py	*linRegRezultatiCeloLeto+deliDneva.txt	151.53239

Tekmovanje:

Podatki. Pri tekmovanju podatki(train.csv.gz) obsegajo podatke o vožnjah vseh avtobusnih linij od januarja do novembra leta 2012. Vsak zapis obsega podatke o registerski tablici avtobusa, voznikovem ID-ju, številki proge, začetni postaji, času odhoda, končni postaji in času prihoda. Pri nalogi sem uporabil podatek o številki proge, začetni in končni postaji proge ter podatke o odhodu in prihodu avtobusa. Najprej sem podatke obdelal tako da sem naredil slovar katerega key je številka proge ter začetna in končna postaja, value pa vsi zapisi te proge iz dane train.csv datoteke. Zapisov, ki so imeli čas odhoda večji od časa prihoda nisem upošteval.

Napovedni model.

- **Ura odhoda** Za uro odhoda sem imel rezerviranih 24 stolpcev(0 - 23), tako sem vrednost 1 zapisal v stolpec, ki je predstavljal uro odhoda v ostale pa vrednost 0.
- **Dan v tednu** Za ta atribut sem imel rezerviranih 7 stolpcev(24 - 30), tako sem vrednost 1 zapisal v stolpec, ki je predstavljal trenutni dan v tednu v ostale pa vrednost 0.
- **Deli dneva** Dan sem razdelil na 5 delov in sicer na jutro(6 - 10), dopoldan(10 - 14), popoldan(14 - 18), večer(18 - 21) in noč(21 - 6). Glede na uro odhoda zapišem vrednost 1 v stolpec ki je določen za ta del dneva, v ostale pa vrednost 0. Ta atribut je močno vplival an rezultat saj je bo nekaterih delih dneva več gneče na cestah in so posledično tudi vožnje daljše.
- **Vikend** V stolpec na mestu tega atributa zapišem vrednost 1 če je dan del vikenda sepravi sobota ali nedelja, in 0 če ni.
- **Praznik** V stolpec na mestu tega atributa zapišem vrednost 0 če ta dan ni praznični in 1 če je. Ob praznikih je pričakovan čas vožnje krajši saj naj nebi bilo toliko gneče na cestah.
- **Mesec** V stolpec na mestu tega atributa zapišem vrednost(1 - 12) glede na to kateri mesec je.

Rezultati. Pri testiranju na lestvici sem najprej doesegel rezultat 193. Program sem nato izboljšal tako da sem spremenil predstavitev atributv ura odhoda in dan v tednu. Za oba atributa sem ustvaril toliko solpcev kolikor vrednosti lahko predstavljata ter nato v stolpec vrednosti, ki jo je vseboval trenutni zapis vpisal 1 v ostale pa 0. Ta način predtstavitve atributov mi je izboljšal rezultat na 185. Menim da je do izboljšave prišlo ker se zaradi take predstavitve spremeni tudi graf pri linearni regresiji(iskanje premice, ki predstavlja najboljše predikcije).

Tabela 2: Rezultati tekmovanja

Datoteka	**Rezultati (.txt)**	**Ocena na lestvici**
tekmovanje.py	*linRegRezultatiTekmovanje.txt	185.54213

Izjava o izdelavi domače naloge. Domačo nalogo in pripadajoče programe sem izdelal sam.