Seminarska naloga

pri predmetu Umetna Inteligenca

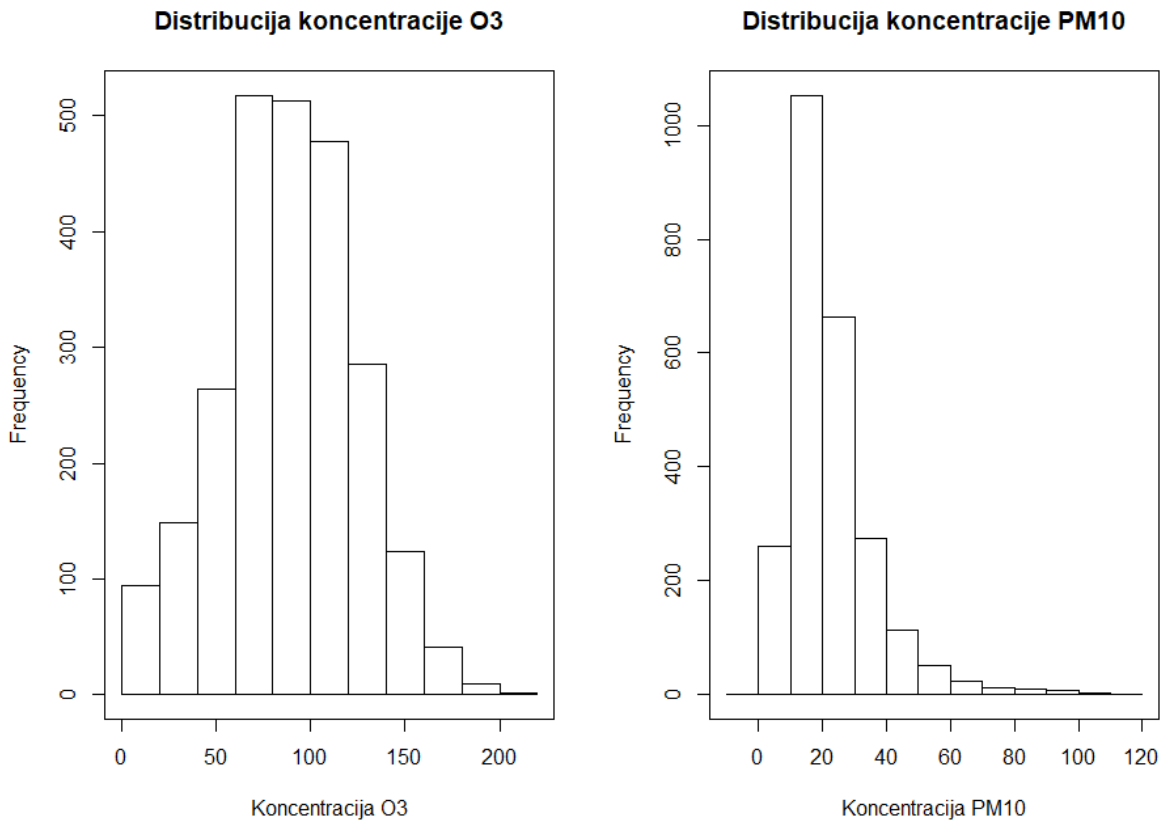
Avtorja: Simon Šegatin (63160294), Anže Košir (63160174)

Ljubljana, 11. 12. 2017

1. Uvod

Seminarsko nalogo sva začela z internetno raziskavo o variaciji koncentracije O3 in PM10 čez različna časovna obdobja. To nama je pokazalo, da tako koncentracija trdih delcev PM10 kot ozona O3 najbolj niha na letni ravni. Nekateri članki so omenjali tudi variacijo ozona glede na to ali je bila meritev opravljena na deloven dan ali med vikendom. Na podlagi teh informacij sva se nato odločila, da atribut Datum odstraniva ter dodava atribute Year, Season, Month in Weekday. Ob pregledu obstoječih atributov sva opazila, da je atribut Glob_sevanje_min vedno enak 0, zato sva ga odstranila. Dodala sva tudi atributa O3Class in PM10Class, ki predstavljata razrede koncentracije O3 in PM10.

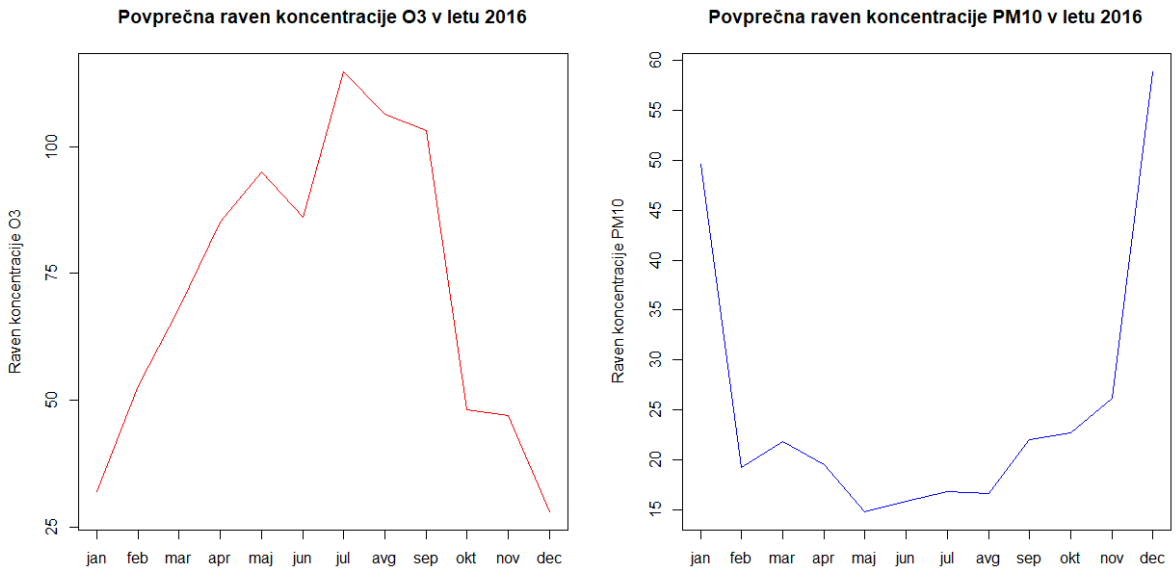
2. Vizualizacija podatkov



Slika 1 - Histograma koncentracij O3 in PM10

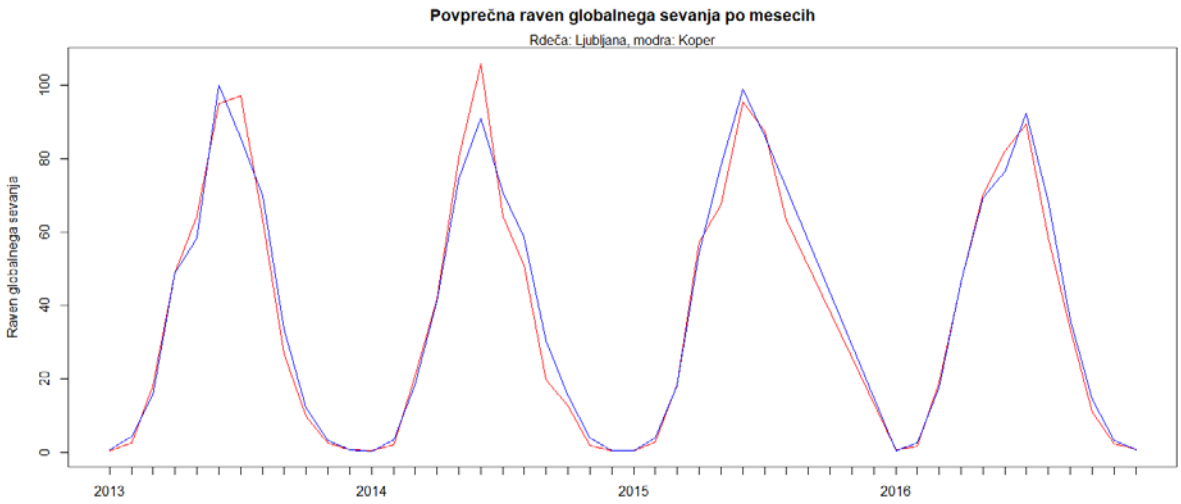
Histograma koncentracij O3 in PM10 na sliki 1 pokažeta, da večina meritev koncentracij O3 spada v srednji razred (med 60 in 120), koncentracij PM10 pa v nizki razred (pod 35).

Med vizualizacijo podatkov sva opazila, da koncentracija O3 in PM10, kot je pokazala najina raziskava, res najbolj niha glede na letni čas (slika 2). Izkazalo se je tudi, da je koncentracija O3 višja v toplejših mesecih, koncentracija PM10 pa v hladnejših.



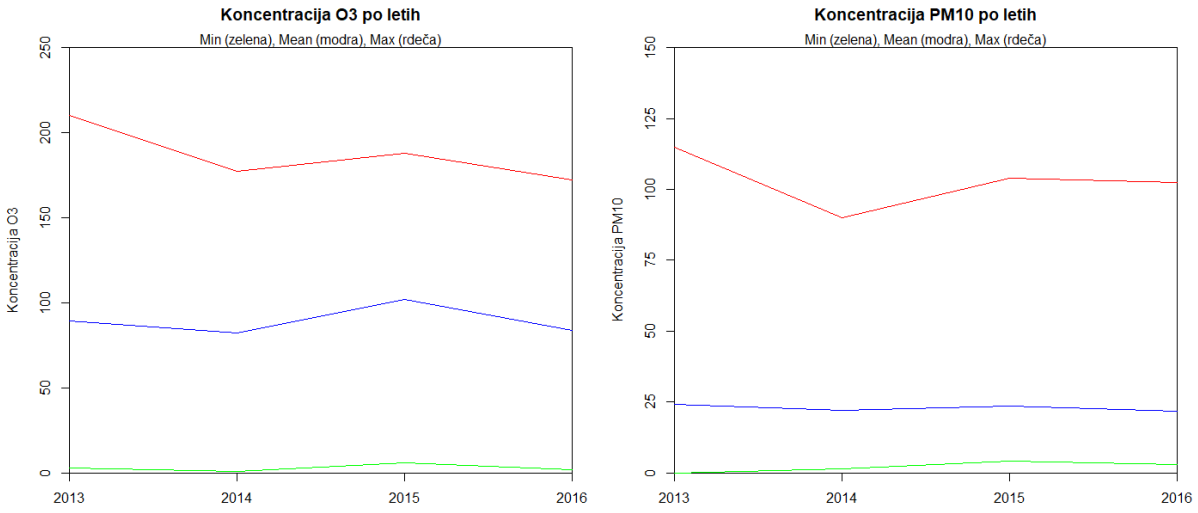
Slika 2 - Povprečni ravni koncentracije O3 in PM10 v letu 2016

Najin sum, da je globalno sevanje najbolj odgovorno za koncentracijo ozona je potrdil graf na sliki 3, iz katerega je lepo razviden vzorec nihanja sevanja, ki je zelo podoben tistemu na sliki 2. Opazna je tudi rahlo višja raven globalnega sevanja na merilni postaji v Kopru. Edina izjema je junij 2014, ko je bila na postaji v Ljubljani izmerjena višja vrednost.



Slika 3 - Raven globalnega sevanja skozi leta (Ljubljana in Koper)

Iz rezultatov zadnjega grafa je razvidno, da koncentracija O3 in PM10 skozi leta ne niha preveč, z izjemo rahlega upada najvišje izmerjene letne koncentracije tako O3 kot PM10 v letu 2014 (slika 4).



Slika 4 - Najmanjša, najvišja in povprečna koncentracija O3 ter PM10 v letih 2013-2016

3. Klasifikacija

Za klasifikacijo sva izbrala modele Odločitveno drevo, naivni Bayes in K-najbližjih sosedov, za izbiro atributov pa sva uporabila funkcijo wrapper z učilnice. Množico podatkov sva razdelila v učno, z meritvami v letih 2013 in 2014, testno, z meritvami v letu 2015, in validacijsko množico, z meritvami v letu 2016. Tako je imela učna množica 1340 vrstic, testna 439, validacijska pa 699.

3.1. Odločitveno drevo

Za izgradnjo modela s pomočjo odločitvenega drevesa sva uporabila knjižnico rpart in atribute Glob_sevanje_max, Temperatura_lokacija_max, Temperatura_Krvavec_mean, Pritisk_min, Sunki_vetra_max, Pritisk_max, Hitrost_vetra_max, Vlaga_max, Padavine_mean, Temperatura_Krvavec_min, Padavine_sum ter Pritisk_mean za koncentracijo O3 in Temperatura_lokacija_max, Hitrost_vetra_min, O3Class, Month, Year, Padavine_sum, Temperatura_lokacija_mean, Postaja, Sunki_vetra_min, Pritisk_max, Temperatura_Krvavec_mean, Pritisk_min, Sunki_vetra_mean ter Padavine_mean za koncentracijo PM10, katere je predlagala funkcija wrapper.

Model je pravilen razred koncentracije O3 in PM10 napovedal z naslednjo natančnostjo:

DT	Testna množica	Validacijska množica
O3 (rpart)	0.7312073	0.7639485
PM10 (rpart)	0.8769932	0.9399142
O3 (CORElearn)	0.715262	0.7410587
PM10 (CORElearn)	0.8792711	0.9298999

Za primerjavo sva zgradila tudi model s knjižnico CORElearn.

3.2. Naivni Bayes

Model z algoritmom naivni Bayes sva prav tako zgradila s knjižnico CORElearn, atribute pa ponovno izbrala s funkcijo wrapper, ki je predlagala Glob_sevanje_max, Pritisk_max, Month, Postaja, PM10Class, Year, Sunki_vetra_mean, Temperatura_Krvavec_min ter Vlaga_min za koncentracijo O3 in Month, Postaja, Sunki_vetra_max, Padavine_sum, Vlaga_max, Padavine_mean, Year ter Pritisk_max za koncentracijo PM10.

Model je pravilen razred koncentracije O3 in PM10 napovedal z naslednjo natančnostjo:

Bayes (CORElearn)	Testna množica	Validacijska množica
O3	0.7539863	0.7482117
PM10	0.8610478	0.9012876

3.3. K-najbližjih sosedov

Za izgradnjo modela z algoritmom KNN sva uporabila knjižnico CORElearn in atribute Month + Temperatura_lokacija_max, Vlaga_max, Sunki_vetra_max, Pritisk_max, Sunki_vetra_min, Padavine_mean, Glob_sevanje_mean, Weekday, Temperatura_Krvavec_min, Sunki_vetra_mean + Temperatura_Krvavec_mean, Hitrost_vetra_max, Vlaga_min, Padavine_sum, PM10Class, Temperatura_Krvavec_max, Glob_sevanje_max ter Season za koncentracijo O3 ter Temperatura_lokacija_max, Temperatura_Krvavec_max, Sunki_vetra_min, Hitrost_vetra_min, Temperatura_lokacija_min, Temperatura_Krvavec_mean, Glob_sevanje_mean, Postaja, Padavine_mean, Glob_sevanje_max, Padavine_sum, Temperatura_lokacija_mean, Temperatura_Krvavec_min, Hitrost_vetra_mean, Hitrost_vetra_max, O3Class, Pritisk_mean, Year, Month, Vlaga_min, Vlaga_max, Pritisk_min, Sunki_vetra_max ter Sunki_vetra_mean za koncentracijo PM10. Te sva ponovno izbrala s pomočjo funkcije wrapper. Nato sva za zanko for(k in 1:100) poiskala najbolj ustrezen k, v primeru koncentracije O3 19, koncentracije PM10 pa 30.

Model je pravilen razred koncentracije O3 in PM10 napovedal z naslednjo natančnostjo:

KNN (CORElearn)	Testna množica	Validacijska množica
O3 (k=19)	0.7790433	0.788269
PM10 (k=30)	0.8861048	0.9327611

4. Regresija

Za regresijske modele sva se odločila za uporabo Odločitvenega drevesa (rpart in CORElearn) ter K-najbližjih sosedov, atribute pa sva dobila s funkcijo wrapperReg z učilnice. Podatke sva razdelila na učno množico z meritvami v letih 2013 in 2014, testno z meritvami v letu 2015 in validacijsko z meritvami v letu 2016. S tem sva dobila učno množico z 1340 vnosi, testno z 439 in validacijsko z 699.

4.1. Regresijsko drevo (rpart)

Za prvega od modelov z algoritmom Regresijsko drevo sva uporabila knjižnico rpart. Atribute sva izbrala s funkcijo wrapperReg, ki je predlagala Month, Temperatura_lokacija_max, Vlaga_max, Sunki_vetra_max, Pritisk_max, Sunki_vetra_min, Padavine_mean, Glob_sevanje_mean, Weekday, Temperatura_Krvavec_min, Sunki_vetra_mean, Temperatura_Krvavec_mean, Hitrost_vetra_max, Vlaga_min, Padavine_sum, Temperatura_Krvavec_max, Glob_sevanje_max in Season za koncentracijo O3 ter Temperatura_lokacija_max, Hitrost_vetra_min, Month, Year, Padavine_sum,

Temperatura_lokacija_mean, Postaja, Sunki_vetra_min, Pritisk_max, Temperatura_Krvavec_mean, Pritisk_min, Sunki_vetra_mean, Padavine_mean za koncentracijo PM10.

Model je koncentraciji O3 in PM10 napovedal z naslednjimi vrednostmi:

RT (rpart)	mae	rmae	mse	rmse
O3 (testna)	14.31028	0.5127426	315.9297	0.2607591
O3 (validacijska)	7.513557	0.7826471	100.4068	0.4867152
PM10 (testna)	7.289967	0.7346053	89.39416	0.4878
PM10 (validacijska)	7.513557	0.7826471	100.4068	0.4867152

4.2. Regresijsko drevo (CORElearn)

Drugi model z algoritmom Regresijsko drevo sva ustvarila s pomočjo knjižnice CORElearn, atribute pa ponovno zbrala s pomočjo funkcije wrapperReg. Ta je predlagala Month, Temperatura_lokacija_max, Vlaga_max, Sunki_vetra_max, Pritisk_max, Sunki_vetra_min, Padavine_mean, Glob_sevanje_mean, Weekday, Temperatura_Krvavec_min, Sunki_vetra_mean, Temperatura_Krvavec_mean, Hitrost_vetra_max, Vlaga_min, Padavine_sum, Temperatura_Krvavec_max, Glob_sevanje_max in Season za koncentracijo O3 ter Temperatura_lokacija_max, Hitrost_vetra_min, Month, Year, Padavine_sum, Temperatura_lokacija_mean, Postaja, Sunki_vetra_min, Pritisk_max, Temperatura_Krvavec_mean, Pritisk_min, Sunki_vetra_mean in Padavine_mean za koncentracijo PM10.

Model je koncentraciji O3 in PM10 napovedal z naslednjimi vrednostmi:

RT (CORElearn)	mae	rmae	mse	rmse
O3 (testna)	14.3683	0.5148212	380.823	0.3143202
O3 (validacijska)	14.97406	0.530427	366.0923	0.3038365
PM10 (testna)	8.657001	0.8723604	151.1523	0.8247977
PM10 (validacijska)	7.642403	0.7960683	112.9832	0.5476788

4.3. K-najbližih sosedov

Model z algoritmom K-najbližjih sosedov sva zgradila s knjižnico CORElearn, za atribute pa izbrala Month, Temperatura_lokacija_max, Vlaga_max, Sunki_vetra_max, Pritisk_max, Sunki_vetra_min, Padavine_mean + Glob_sevanje_mean + Weekday + Temperatura_Krvavec_min + Sunki_vetra_mean, Temperatura_Krvavec_mean, Hitrost_vetra_max, Vlaga_min, Padavine_sum, Temperatura_Krvavec_max, Glob_sevanje_max in Season s k-jem 70 za koncentracijo O3 ter Temperatura_lokacija_max, Hitrost_vetra_min, Month, Year, Padavine_sum, Temperatura_lokacija_mean, Postaja, Sunki_vetra_min, Pritisk_max, Temperatura_Krvavec_mean, Pritisk_min, Sunki_vetra_mean in Padavine_mean s k-jem 100 za koncentracijo PM10. Določila sva jih s funkcijo wrapperReg.

Model je koncentraciji O3 in PM10 napovedal z naslednjimi vrednostmi:

KNN (CORElearn)	mae	rmae	mse	rmse
O3 (testna)	27.54603	0.8964438	1450.68	0.9819594
O3 (validacijska)	33.13026	1.173575	1830.351	1.519091
PM10 (testna)	8.164687	0.8316049	118.6658	0.6464419
PM10 (validacijska)	7.6427	0.7960992	123.5775	0.5990339

5. Zaključek

Naloga je pokazala, da je za ta primer od testiranih modelov za klasifikacijo najbolj natančen algoritem K-najbližjih sosedov, saj je pravilno napovedal v povprečju 4,56 odstotne točke več primerov v testni množici in 3,72 odstotne točke več primerov v validacijski množici. Pri klasifikaciji koncentracije PM10 je bil prav tako najboljši K-najbližjih sosedov in sicer za v povprečju 1,37 odstotne točke pri testni ter 2,72 odstotne točke pri validacijski množici. Pri regresiji je bil najbolj natančen model z algoritmom Regresijsko drevo iz knjižnice rpart.