

S&R5

Srednja elektro-računalniška šola Maribor

NAPOVEDOVANJE DELCEV PM10

Raziskovalno področje: računalništvo

Raziskovalna naloga

Šola: Srednja elektro-računalniška šola Maribor

Avtorji: Maša Šulc, Tara Pučnik, Samuel Perovšek

Mentorja: Bojan Ploj in Andrej Korošec

Maribor, 2025

1. Kazalo vsebine

Kazalo:

1. Kazalo vsebine.....	2
2. Kazalo slik.....	3
3. Seznam kratic.....	4
4. Povzetek.....	5
5. <i>English summary</i>	6
6. Zahvala.....	7
7. Uvod.....	8
Zgodba.....	8
Hipoteze.....	9
8. <i>Pregled stanja tehnike</i>	9
Vpliv vremena na nivo PM10 delcev.....	10
Strojno učenje in napovedovanje onesnaženosti zraka.....	10
Strojno učenje.....	11
Načini ugotavljanja napak.....	11
9. Metodologija.....	12
Razvojno okolje in naše delo.....	12
Priprava podatkov.....	12
Prilagajanje podatkov za analizo.....	13
Prilagoditev časovnih podatkov.....	13
Standardizacija kategoričnih vrednosti.....	14
Proces treniranja.....	14
WEKA.....	14
Orange.....	16
Pripomočki programa Orange.....	17
<i>Podatkovna skupina</i>	17
<i>Pretvorna skupina</i>	19
<i>Evaluacijska skupina</i>	20
<i>Skupina modelov</i>	21
Uporabljeni modeli strojnega učenja.....	21
Nevronska mreža.....	22
Linearna regresija.....	22

Naključni gozd	23
10. Rezultati	23
11. <i>Primerjava modelov</i>	23
Linearna regresija	24
Orange.....	24
Naključni gozd (Naključni gozd).....	25
Metoda podpornih vektorjev (SVM).....	25
Nevronska mreža.....	26
Vizualizacija rezultatov.....	27
12. Diskusija.....	31
Potrjevanje hipotez	31
Potrditve hipotez	32
Razprava o rezultatih	32
Analiza dela	33
Iskanje najboljšega modela za napoved PM10.....	33
Družbena odgovornost in trajnostni razvoj.....	34
Trajnost.....	35
13. Viri.....	35
14. Viri slik.....	36

2. Kazalo slik

Slika 1: PM10 delci	9
Slika 2: Predstavitev povezave v Orange-u (lasten vir)	20
Slika 3: Widget-i, ki obdelujejo delovanje datoteke (lasten vir)	20
Slika 4: Widget za vnos CSV datoteke (lasten vir)	21
Slika 5: Prikaz wigeta, ki nasplošno uvaža datoteke (lasten vir)	21
Slika 6: Tabela podatkov (lasten vir)	22
Slika 7: Widgeti pretvorne skupine (lasten vir)	22

Slika 8: Evaluacijska skupina (lasten vir)	23
Slika 9: Slika strojnih modelov (lasten vir)	24
Slika 10: Rezultati Linearne regresije v WEKI (lasten vir)	27
Slika 11: Rezulta Linearne regresije v Orange-u (lasten vir)	27
Slika 12: Random fores v WEKI (lasten vir)	28
Slika 13: Naključni gozd v Orange (lasten vir)	28
Slika 14: SVM v WEKI (lasten vir)	29
Slika 15: SVM v Orangeu (lasten vir)	29
Slika 16: Nevronska mreža v WEKI (lasten vir)	30
Slika 17: Nevronska mreža v Orangeu (lasten vir)	30
Slika 18: Vizualizacija delovanja linearne regresije (lasten vir)	31
Slika 19: Vizualizacija delovanja SVM-a (lasten vir)	32
Slika 20: Vizualizacija delovanja Naključni gozd (lasten vir)	33
Slika 21: Vizualizacija delovanja Nevronske mreže (lasten vir)	34
Slika 22: Primerjava modelov v Orange-u (lasten vir)	36

3. Seznam kratic

- NLZOH- Nacionalna laboratorija za zdravje, okolje in hrano (ang. National laboratory for health, environment and food)
- ML- Strojno učenje (ang. Machine learning)
- PM10- Trdni delci od 10 μm (ang. Particulate matter of 10 μm)
- MAE- Povprečna absolutna napaka (Mean Absolute Error)
- RMSE- Koren povprečne absolutne napake (Root Mean Squared Error)
- MSE- Povprečje kvadratov napak (Mean Square Error)
- ANNs- Artificial neural networks
- IoT- Internet of Things

4. Povzetek

Raziskava se osredotoča na napovedovanje ravni PM10 [1] v kakovosti zraka ob ustvarjanju modela strojnega učenja na podlagi podatkov Nacionalne laboratorije za dravje, okolje in hrano (NLZOH). Naš glavni cilj je zmanjšati napake pri napovedovanju z manipulacijo podatkov, prehodom s platforme strojnega učenja WEKA na Orange in konstruiranjem modela z uporabo različnih podatkov in orodij. Opisujemo postopek natančnega prilagajanja modela in ocenjujemo uspešnost različnih učnih metod z uporabo metrik, kot sta »Povprečna Absolutna Napaka« (MAE) in »Povprečna Kvadratna Napaka« (RMSE). Z združevanjem teh meritev z različnimi knjižnicami nam je to pomagalo razglasiti najboljšo metodo.

Čeprav je raziskava osredotočena na tehnične vidike, upamo, da bomo ozaveščali o kakovosti zraka, izobraževali in pokazali svoj potencial kot skupina raziskovalcev.

Ključne besede: umetna inteligenca, strojno učenje, kakovost zraka, manipulacija podatkov, napake.

5. English summary

This research focuses on predicting PM10 levels in air quality while creating a machine learning model based on data provided by the National laboratory for health, environment and food (NLZOH). Our main goal is to reduce prediction errors by manipulating data, transitioning from the WEKA machine learning platform to Orange, and constructing a model using various data and tools. We describe the process of finetuning the model and evaluate the performance of different learning methods using metrics such as “Mean Absolute Error” (MAE) and “Root Mean Square Error” (RMSE). By combining these metrics with various libraries, that helped us declare the best method.

Although the research is centered on technical aspects, we hope to raise awareness about air quality, educate, and show our potential as a group of researchers.

Keywords: artificial intelligence, machine learning, air quality, data manipulation, errors.

6. Zahvala

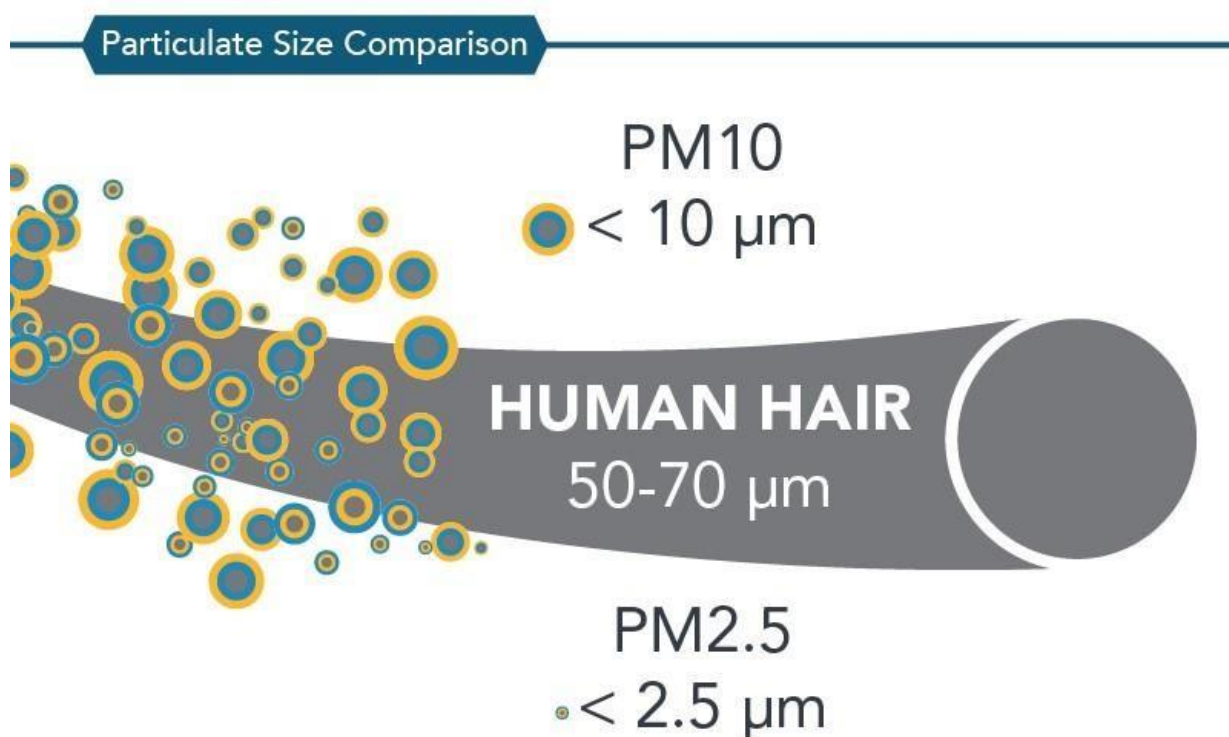
Zelo smo hvaležni našemu mentorju, Bojan Ploju, ki nam je pomagal in dal dovolj tekmovalnega duha, da smo se prepustili takšnim raziskavam, da smo premikali lastne meje in pridobili nove izkušnje na področju raziskovanja in timskega dela.

Hvaležni smo tudi Maji Hleb, ki nam je s svojo radodarnostjo in delavnostjo kljub temu, da ni ena od avtoric tega raziskovalnega poročila, zelo pomagala pri tem projektu.

7. Uvod

Kvaliteta zraka je eden izmed najpomembnejših dejavnikov, ki vplivajo na zdravje ljudi. Vsa živa bitja ga rabimo, pa če si to želimo ali ne. Pa vendar smo se »potrudili«, da tudi to uničimo.

Zadnjih 300 let, od izuma parnega stroja, onesnaženost zraka drastično raste zaradi industrijskih obratov, prevoznih sredstev, ter dosti drugih stvari. Prašni delci, PM10 nastanejo pri vsakem segrevanju snovi. Veliko količino delcev povzroča onesnaženost zraka, ki vsa živa bitja izpostavlja nevarnostim. Tako rečeno, varne količine PM10 delcev ni.



Slika 1: PM10 delci

Zgodba

Najprej želimo pojasniti, kako smo pridobili naše učne podatke. Stopili smo v stik z NLZOH (Nacionalni inštitut za javno zdravje), ki je tesno povezan z ARSO (Agencija Republike Slovenije za okolje) [4]. Naš mentor je vzpostavil stik z raziskovalcem, ki se ukvarja s kakovostjo zraka v Mariborski regiji. Iz tega stika smo dobili idejo za naš

raziskovalni projekt, kjer smo kot osnovo za našo nalogo vzeli že obstoječe rezultate z FERI, ki že dolgo sodeluje z NLZOH na podobnih področjih.

V ta namen smo uporabili iste učne podatke kot FERI, ki vključujejo informacije o dejavnikih, ki vplivajo na kakovost zraka (kot so jakost vetra, padavine, temperatura in drugi parametri). Algoritmi strojnega učenja so namreč postali učinkovito orodje za napovedovanje, še posebej pri uporabi kompleksnejših naborov podatkov. Na podlagi ene od datotek, ki smo jo pridobili z NLZOH, smo se odločili razviti algoritem strojnega učenja za napovedovanje onesnaženosti zraka v prihodnjih letih.

Cilj našega raziskovalnega dela je razviti najnatančnejši model za napovedovanje kakovosti zraka v okolici Maribora. Ta model želimo uporabiti kot orodje za vsakodnevno obveščanje javnosti o trenutni ravni prašnih delcev v zraku. Prav tako upamo, da bo naše delo koristilo zavodu, pri zagotavljanju informacij o ravni PM10.

Hipoteze

- Ugotavljali bomo ali podatki vsebujejo nerelavantne attribute. To so atributi za katere predvidevamo da nima ključnega vpliva na natančnost napovedovanja.
- Napako napovedovanja onesnaženosti zraka z delci PM10 bomo poskušali zmanjšati na vrednost manj kot $8 \mu\text{g}/\text{m}^3$ (natančnost FERI-ja).
- Ugotavljali bomo katera vrsta napovednega modela dosega najboljše rezultate.
- Ugotavljali bomo s katerim orodjem (WEKA ali Orange) lahko dosežemo boljše rezultate.

8. Pregled stanja tehnike

Kakovost zraka ne vpliva le na naše zdravje, ampak tudi na okolje. Eden izmed pomembnejših delov onesnaženosti zraka so delci PM10, ki izvirajo iz različnih virov: promet, industrija, nekateri naravnimi procesi itd. Njihova koncentracija v zraku se lahko spreminja glede na vremenske razmere, letni čas in druge dejavnike okolja. Ker so se v preteklih desetletjih za tovrstne napovedi uporabljali modele strojnega učenja, smo se odločili, da svojo raziskavo posvetimo izboljšavi modela, ki lahko na podlagi podatkov, ki smo jih pridobili z NLZOH, napove ravni PM10 delcev za vsak naslednji dan. Naš cilj je narediti model strojnega učenja, ki bo zanesljiv za morebitno varovanje

zdravja ljudi, tako da lahko v nadaljevanju projekta lahko povežemo umetno inteligenco s programom, ki bi ljudem lahko pošiljal obvestila o ravni PM10 v zraku ter ali naj se rekreativno ukvarjajo zunaj ali ne.

Kako se trenutno napoveduje kvaliteta zraka?

Znanstveniki in inženirji že leta delajo na modelih za napovedovanje kakovosti zraka. Tradicionalne metode, kot so statistični modeli (ARIMA, večkratna linearna regresija), so bile uporabljene za napovedovanje in čeprav delujejo razmeroma dobro, imajo omejitve. Ti modeli težko zajamejo zapletena razmerja med različnimi okoljskimi dejavniki in so lahko manj natančni pri obravnavanju hitro spreminjajočih se razmer.

Nedavno je bilo strojno učenje uvedeno kot učinkovitejša alternativa. Napredne tehnike, kot so ANN, odločitvena drevesa in metode globokega učenja, lahko analizirajo ogromne količine podatkov in prepoznajo vzorce, ki bi jih tradicionalni modeli morda zgrešili. Modeli strojnega učenja lahko upoštevajo več dejavnikov hkrati glede na pretekle ravni onesnaženosti, vremenske razmere, kot sta vlažnost in hitrost vetra, in celo količino prometa, za bolj natančne napovedi o kakovosti zraka.

Vpliv vremena na nivo PM10 delcev

Vreme ima veliko vlogo pri koncentracijah PM10. Na primer, visoka vlažnost lahko povzroči, da se delci zlepijo skupaj, kar poveča raven onesnaženosti. Po drugi strani pa lahko močni vetrovi razpršijo onesnaževala in tako izboljšajo kakovost zraka. Temperaturne inverzije, kjer topel zrak ujame hladnejši zrak (in delce v njem) blizu tal, kar lahko privede do nevarno visokih koncentracij, zlasti v mestih, ki se nahajajo v kotlinah, npr. Ljubljana, Velenje itd.

Z vključitvijo vremenskih podatkov v napovedne modele lahko dobimo veliko jasnejšo sliko o obnašanju delcev v zraku. Študije so pokazale, da so modeli strojnega učenja, ki vključujejo meteorološke podatke, bistveno boljši od tistih, ki se opirajo zgolj na pretekle vzorce onesnaženosti.

Strojno učenje in napovedovanje onesnaženosti zraka

V zadnjih letih je umetna inteligenca [6] močno napredovala pri napovedovanju onesnaženosti zraka. Modeli globokega učenja, kot so omrežja dolgotrajnega kratkoročnega spomina (LSTM), lahko analizirajo podatke časovnih vrst, se učijo iz

preteklih vzorcev za napovedovanje prihodnjih ravni onesnaženosti. Medtem metode ansambla, kot je XGBoost, združujejo več modelov za izboljšanje natančnosti.

Drug razburljiv razvoj je integracija strojnega učenja s senzorji IoT (Internet of Things). Pametni monitorji kakovosti zraka, nameščeni na ključnih lokacijah, lahko zagotovijo podatke v realnem času, ki se vnesejo neposredno v modele AI, zaradi česar so napovedi še natančnejše. Ta kombinacija umetne inteligence in spremljanja v živo bi lahko spremenila način sledenja in odzivanja na onesnaženje zraka.

Strojno učenje

Strojno učenje predstavlja področje računalništva, ki se ukvarja z iskanjem znanja v kopici podatkov.

Poznamo tri vrste učenja: nadzorovano, nenadzorovano in okrepčevalno.

Pri nadzorovanem učenju so v učnih podatkih na voljo tudi pravilni rezultati. Dober primer nadzorovanega učenja je zaznavanje neželene e-pošte, kjer je označeno katera pošta je neželena.

Nenadzorovano učenje je nasprotje nadzorovanega, saj v učnih podatkih ni pravilnih rezultatov. Takšno učenje se lahko uporabi za predpripravo učnih podatkov.

Pri okrepčevalnem učenju se model uči na poskusih in tako poveča svojo učinkovitost. Ta pristop se tipično uporablja pri strateških igrah (npr. šah).

Načini ugotavljanja napak

Obstaja več vrst izračuna napake:

- **RMSE (povprečna kvadratna napaka) [5]** je kvadratni koren povprečne kvadratne napake .
- **MAE (Mean Absolute Error)** je povprečje absolutnih napak.
- **MSE (srednja kvadratna napaka)** je povprečje kvadratov napak.

RMSE je večji od MAE, kadar so posamezne napake izstopajoče. Sicer so vrednosti RMSE, MAE in MSE podobne.

9. Metodologija

Razvojno okolje in naše delo

V grobem se je naše delo delilo na dva dela: priprava podatkov in strojno učenje.

Za našo raziskavo smo uporabili kombinacijo različnih programov, kar nas je pripeljalo do spremembe pri izbiri programske opreme za strojno učenje. Začeli smo z WEKA-o, vendar smo zaradi težav pri vizualizaciji in analizi prešli na Orange, ki nam je omogočil boljši pregled nad podatki ter učinkovitejšo obdelavo.

Priprava podatkov

Pred učenjem smo učne podatke pripravili z urejevalnikom tabel MS Excel in urejevalnikom besedil Notepad++. Uporabili smo ju za čiščenje, urejanje in preoblikovanje podatkov. Z njima smo podatke oblikovali v obliko, ki jih bereta WEKA in Orange. Odstranjevali smo nepotrebne parametre, popravljali napake pri kodiranju znakov in dopolnjevali manjkajoče vrednosti. Pripravo podatkov smo zaključili s programoma Orange in WEKA: odstranjevanje nerelevantnih podatkov.

- **WEKA** – Naš prvi program za strojno učenje, ki smo ga uporabili za analizo podatkov in modeliranje z različnimi algoritmi. Izkazalo se je, da ima določene omejitve pri vizualizaciji in obdelavi podatkov, zaradi česar smo kasneje prešli na drugo orodje.
- **Orange** – Zamenjal je Weko, saj je omogočal bolj intuitivno analizo podatkov. Omogočil nam je vizualizacijo podatkov, uporabo različnih algoritmov ter enostavnejše prilagajanje modelov.

Eden od največjih izzivov so bili manjkajoči podatki. Na primer, v stolpcu oblačnosti je bilo od 456 vrstic le 176 veljavnih vnosov. Sprva smo razmišljali o ignoriranju manjkajočih vrednosti, vendar smo ugotovili, da je to vplivalo na kakovost modela. Zato smo uporabili različne metode zapolnjevanja vrzeli:

- Pri številskih stolpcih smo vrednosti nadomestili s povprečjem.
- Pri kategoričnih stolpcih smo uporabili najpogostejšo vrednost.
- V določenih primerih smo podatke dopolnili s podatki iz podobnih vrstic.

Poleg tega smo odstranili nepotrebne parametre, ki so podvajali informacije. Na primer, podatki o ledenem dežju vsebujejo skoraj enake informacije kot podatki o dežju. Zato smo ju združili v enega. Prav tako smo naleteli na nekaj nejasnih stolpcev, ki jih zaradi pomanjkljivih podatkov nismo mogli smiselno interpretirati, zato smo jih odstranili.

Prilagajanje podatkov za analizo

Uporaba več različnih programov je pomenila tudi delo s podatki v različnih formatih. Naša naloga je bila priprava podatkov v CSV obliki in njihova pretvorba v ARFF (Attribute-Relation File Format). To se je izkazalo za eno bolj zamudnih opravil, saj WEKA ni mogla prebrati določenih vrstic, kar smo morali ročno popraviti. Poleg tega smo imeli težave s kodiranjem znakov, glede na to, da nekateri stolpci so vsebovali napačne simbole (npr. »Â«, »Ä« in »L«), kar je povzročalo težave pri analizi. Popravili smo te napake, da so bili podatki pravilno interpretirani.

Celoten proces priprave podatkov je bil časovno zahteven, vendar ključen za doseglo kakovostnih rezultatov. S skrbnim urejanjem podatkov in uporabo ustreznih metod smo na koncu pridobili bolj zanesljive modele in natančnejšo analizo.

Prilagoditev časovnih podatkov

Zanimalo nas je kako manjše število primerov vpliva na zanesljivost naše raziskave.

Med testiranjem modelov strojnega učenja smo ugotovili, da manjše število učnih primerov ne poslabša občutno rezultatov učenja.

Naš cilj ni bil le napovedovanje ravni PM10, temveč tudi izbira najboljšega modela. Za primerjavo modelov smo uporabili rezultate FERl (Fakulteta za elektrotehniko, računalništvo in informatiko v Mariboru). Za boljšo primerjavo rezultatov smo uporabili povsem enake učne podatke. Model FERl je dosegel povprečno absolutno napako 8 $\mu\text{g}/\text{m}^3$. Naša priprava učnih podatkov se je razlikoval v tem, da smo podatek o datumu ločili na podatke leto, mesec, dan. S pomočjo urejevalnika Excel, smo iz datumov izračunali še dan v tednu. Na ta način se je rezultat učenja občutno izboljša, saj se med vikendom dejavnost tovarn običajno zmanjša.

Standardizacija kategoričnih vrednosti

Zaradi nekaterih podatkov kategoriziranih kot »da« ali »ne«, kot so veter, dež in sneg, smo te pretvorili v 1 in 0 za lažjo obdelavo z Orange-om. Prav tako smo prilagodili številčne vrednosti na obseg od 0 do 1, saj nekatere modele moti, če so različni atributi različnega velikostnega razreda.

Proces treniranja

Podatke smo razdelili na učno in testno množico v razmerju 66:33, kar pomeni, da smo 66 % podatkov uporabili za učenje modela, preostalih 33 % pa za testiranje naučenosti. Ta pristop omogoča, da preverimo, kako dobro se model uči in spopada z novimi podatki, ki niso bili vključeni v fazo učenja. Na ta način lahko ocenimo dejansko zmogljivost modela, saj je testna množica ločena od učne, kar preprečuje prekomerno prilagajanje na podatke, ki jih je model že »videl« med učenjem.

Poleg tega smo uporabili 5-kratno križno validacijo, ki vključuje večkratno razdelitev podatkov na različne dele za treniranje in preverjanje modela. Ta tehnika pomaga doseči bolj zanesljive rezultate, saj ocena ne temelji le na eni naključno izbrani testni množici. S temi pristopi smo sistematično raziskali možne nastavitve, da bi izbrali tiste, ki dajejo najboljše rezultate.

WEKA

Na začetku naše raziskave smo se odločili uporabiti program WEKA, ki smo ga bili vaje že od prej. Zdelo se nam je, da bo delo z njim enostavno in učinkovito, saj ponuja različna orodja za analizo podatkov. A že pri prvem koraku smo naleteli na problem. WEKA za strojno učenje uporablja samo datoteke tipa ARFF, kar pomeni, da je treba pri pripravi datoteke natančno določiti attribute (stolpce z našimi podatki) in tipe vrednosti, ki jih lahko vsebujejo (npr. numeric, string, nominal itd.). Tukaj je nastala prva težava, saj so nekateri stolpci vsebovali številske vrednosti v stolpcih določenih kot string, čeprav to ni bila dejanska napaka, je WEKA to prepoznala kot zapis napačnega podatka v stolpec.

Naša nova metoda je sprva obetala, vendar smo kmalu naleteli na težave. Večina podatkov je bila razporejena v kategoriji 1 in deloma v kategoriji 2, medtem ko sta kategoriji 3 in 4 ostali prazni. Ta neravnovesna porazdelitev je onemogočala pravilno

prepoznavanje kategorij 3 in 4. Zaradi tega smo opustili ta pristop in se vrnili k napovedovanju točnih vrednosti koncentracije delcev.

Kljub temu da so nekatere nastavitve WEKA-e dosegle nekoliko boljše rezultate kot Orange, smo opazili, da so modeli pogosto preveč prilagojeni specifičnim vzorcem v učnih podatkih. To je povzročilo, da so bili rezultati manj generalizirani in niso delovali dobro pri napovedovanju novih, nepoznanih podatkov. Prekomerno učenje je bil še posebej izrazit pri razporeditvi podatkov v različne kategorije, kjer so modeli pretirano upoštevali majhne variacije v podatkih, kar je privedlo do natančnih, vendar napačnih napovedi. To nas je prepričalo, da moramo preiti na bolj robusten in fleksibilen pristop, kot je Orange, ki je ponujal boljše možnosti za obvladovanje te težave in omogočal večjo natančnost pri napovedovanju.

Po predlogu mentorja smo prek uradne spletne strani WEKA-e uvozili dodatno funkcijo imenovano Time series forecasting, ki izboljša učenje modela z uporabo podatkov s časovnimi nizi.

Po pripravi naših učnih podatkov smo začeli s testiranjem različnih ponujenih modelov. Pri tem smo se osredotočili na nevronske mreže in naključno drevo, saj smo iz izkušenj iz preteklosti pričakovali, da bosta najboljša. Pri naključnem drevesu smo prilagajali število atributov, ki jih je model uporabljal za gradnjo posameznih dreves. Čeprav smo uspeli ustvariti modele in jih preizkusiti na naših podatkih, smo kmalu ugotovili, da rezultati niso dovolj dobri. Natančnost napovedi PM10 se je gibala med $20 \mu\text{g}/\text{m}^3$ in $25 \mu\text{g}/\text{m}^3$, pri čemer so nevronske mreže pokazale nekoliko boljše rezultate kot naključno drevo. A to še vedno ni bilo dovolj.

Ker smo ugotovili, da natančno napovedovanje količine delcev z obstoječim pristopom ni uspešno, smo se odločili za spremembo strategije.

Ker so omejitve WEKA-e bile zelo očitne, smo iskali še druge načine kako izboljšati rezultate. Ena od idej je bila zbiranje dodatnih podatkov, kot so podatki o prometu ali požarih, v bližini merilnih postaj. Namesto tega smo se odločili raziskati druge programe za strojno učenje, ki ponujajo več funkcionalnosti in lažjo integracijo z večjimi, kompleksnimi podatkovnimi nabori. Tako smo se odločili da poskusimo še z Orange-om, ki smo ga tudi že uporabljali v preteklosti za priprave na tekmovanje. Hitro smo ugotovili, da ponuja številno dodatnih možnosti, ki jih ni v WEKA-i.

Orange

Orange je slovenski program fakultete FRI na univerzi v Ljubljani. Uporaba Orange je zaradi čarovnikov (widgets [2]), bolj prijazna uporabniku.

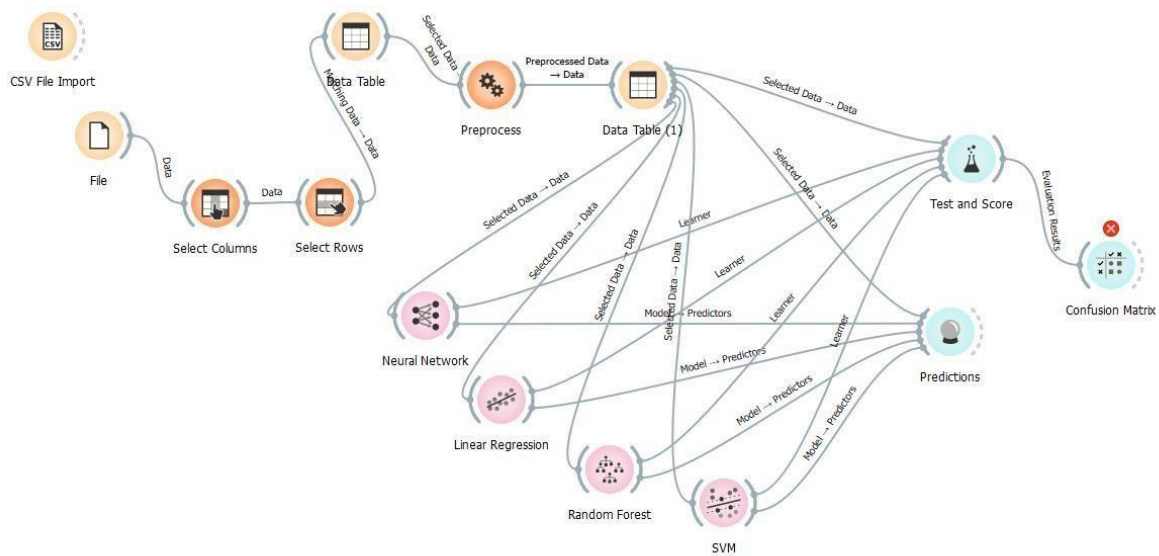
Ena od prednosti je bila uporaba datotek tipa CSV, ki je bistveno bolj preprost kot ARFF in omogoča uporabo Excela. Orange ponuja bolj nazorno prikazovanje rezultatov. Poleg zgolj prikaza nam je omogočil tudi, da podatke spreminjamo neposredno znotraj Oranga, ne da bi s tem vplivali na izvirno datoteko. To je bilo izredno uporabno, saj smo lahko postopoma dodajali in odstranjevali stolpce ter vrstice, kadar smo menili, da je to potrebno za izboljšanje našega modela.

Eden ključnih widgetov, ki smo jih uporabljali, je bil Preprocess. S pomočjo tega orodja smo lahko urejali podatke na različne načine. Na primer prazna polja smo lahko zapolnili z različnimi metodami, kot so naključne vrednosti, mediana ali povprečje obstoječih vrednosti. Preprocess vključuje tudi dodatne funkcionalnosti, kot je normalizacija podatkov, s katero smo vrednosti podatkov preslikali na interval od 0 do 1. To je bil ključen korak za zmanjšanje vpliva različnih velikostnih razredov.

Poleg funkcij za urejanje podatkov nam je Orange ponujal širok spekter modelov in prilagodljivih funkcij za vsak model posebej. Delo z modeli je potekalo na podoben način kot v programu WEKA, vendar smo tukaj imeli več svobode pri sprotnem prilagajanju atributov. To pomeni, da smo lahko neposredno med analizo izbirali, katere lastnosti bomo vključili v model in katere bomo uporabili kot tarčo napovedovanja. Za testiranje natančnosti modelov smo uporabljali dva widgeta. Prvi widget je bil »Test and Score«, ki omogoča preverjanje natančnosti modela s pomočjo sistema cross-validation. To orodje je zelo učinkovito, saj razdeli podatke na več delov, pri čemer en del uporabi za testiranje, ostale pa za učenje modela. Drugi widget, ki smo ga uporabljali, je bil Predictions, ki je posebej prilagojen za napovedovanje rezultatov na podlagi obstoječih modelov.

Rezultati so pokazali, da je napaka našega modela padla pod $5 \mu\text{g}/\text{m}^3$, kar smatramo kot uspeh, ko se primerjamo z rezultatom FERI, ki znaša $8 \mu\text{g}/\text{m}^3$. Ta pristop je potrdil učinkovitost naše metodologije in orodij. Na tej točki smo raziskavo zaključili, saj smo dosegli svoje cilje.

Pripomočki programa Orange



Slika 2: Predstavitev povezave v Orange-u (lasten vir)

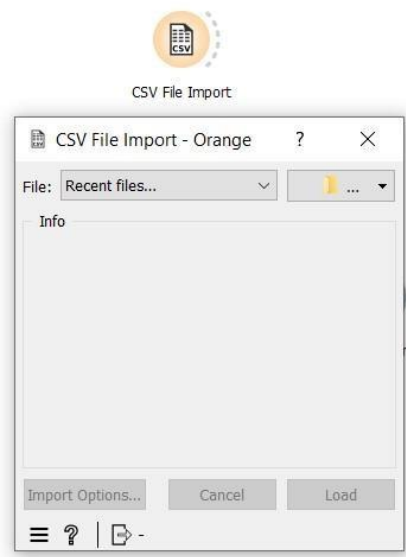
V procesu raziskave smo preizkusili širok spekter različnih gradnikov (widgetov) in jih povezali, da bi dosegli učinkovito predstavitev modela, testiranja in napovedi. Na začetku smo se osredotočili na različne vrste orodij za obdelavo podatkov, tako da smo lahko optimizirali našo raziskavo in testirali različne pristope. Celoten postopek smo razdelili v različne skupine, vsaka pa je imela svoje specifične naloge, kar nam je omogočilo boljši nadzor nad obdelavo podatkov.

Podatkovna skupina

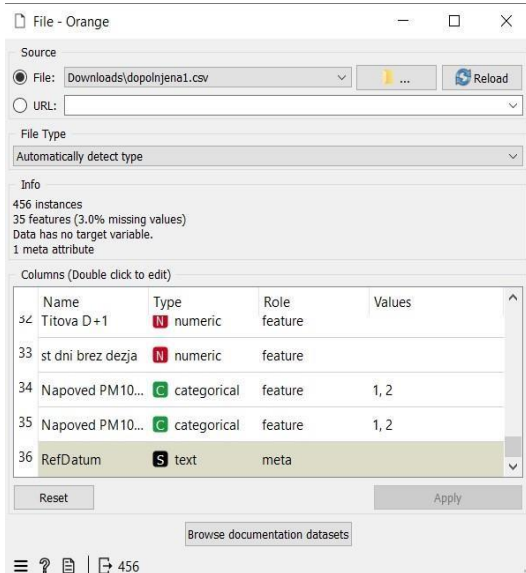


Slika 3: Widget-i, ki obdelujejo delovanje datoteke (lasten vir)

Podatkovna skupina je bila ključna pri začetnem uvozu in prikazu podatkov. Začeli smo s funkcijo za uvoz datotek, kjer smo uporabili CSV datoteke, saj so te omogočale enostaven uvoz podatkov v naš program. Prav tako smo se odločili za večjo fleksibilnost in uporabljali "File" widget, ki nam je omogočil uvoz različnih vrst datotek, kar je pripomoglo k širokemu spektru podatkovnih virov. V tej fazi smo si prizadevali, da smo imeli popolno kontrolo nad tem, katere podatke uvažamo, saj smo želeli, da so bile vse potrebne informacije vključene v analizo



Slika 4: Widget za vnos CSV datoteke (lasten vir)



Slika 5: Prikaz wigeťa, ki nasplošno uvažã datoteke (lasten vir)

Vizualizacija podatkovne tabele nam prikazuje podatke, kot so učni primerki, funkcije, meta atributi in številčni rezultati.

Data Table - Orange

Info
445 instances
30 features (2.3 % missing data)
Numeric outcome
1 meta attribute

Variables
 Show variable labels (if present)
 Visualize numeric values
 Color by instance classes

Selection
 Select full rows

Restore Original Order
 Send Automatically

	Titova D+1	RefDatum	st dni brez deolja	dez	Mesec	povpdnevna T [°C]	maxT [°C]	minT [°C]	minT na 5cm [°C]	kategorija_Vrbanska
1	13.7	2.10.2022	0	da	10	14.6	21.1	8.0	5.2	1
2	16.8	3.10.2022	1	ne	10	12.3	20.5	8.7	5.4	1
3	21.3	4.10.2022	2	ne	10	11.0	19.2	3.5	0.5	1
4	25.2	5.10.2022	3	ne	10	12.9	20.5	4.5	1.5	1
5	27.5	6.10.2022	4	ne	10	14.0	22.3	7.0	4.0	1
6	25.8	7.10.2022	5	ne	10	14.8	21.4	8.5	5.2	1
7	18.2	8.10.2022	6	ne	10	13.8	21.0	8.3	5.2	1
8	27.9	9.10.2022	7	ne	10	12.3	17.1	9.3	7.7	1
9	37.6	10.10.2022	8	ne	10	12.0	18.0	7.4	3.9	1
10	33.1	11.10.2022	9	ne	10	14.5	20.2	6.9	3.9	1
11	36.2	12.10.2022	10	ne	10	11.8	19.4	7.5	4.2	1
12	46.8	13.10.2022	11	ne	10	11.5	19.3	5.6	2.9	1
13	47.2	14.10.2022	12	ne	10	10.9	18.4	5.4	2.6	1
14	39.5	15.10.2022	13	ne	10	12.6	17.0	7.4	4.9	1
15	39.8	16.10.2022	14	ne	10	13.2	21.6	7.9	5.4	1
16	38.5	17.10.2022	15	ne	10	13.4	22.6	7.0	4.7	1
17	31.5	18.10.2022	16	ne	10	13.7	23.0	6.5	4.1	1
18	22.3	19.10.2022	17	ne	10	13.1	19.0	8.8	6.0	1
19	35.6	20.10.2022	18	ne	10	11.9	16.9	9.8	8.3	1
20	21.3	21.10.2022	19	ne	10	11.9	18.3	5.8	3.8	1
21	20.6	22.10.2022	0	da	10	13.2	18.4	9.7	10.0	1
22	40.3	23.10.2022	1	ne	10	13.8	21.1	6.1	4.7	1
23	14.2	24.10.2022	0	da	10	16.4	21.2	10.6	8.0	1
24	23.1	25.10.2022	0	da	10	13.1	20.0	9.9	11.6	1
25	21.7	26.10.2022	1	ne	10	12.5	19.5	6.2	3.6	1
26	29.0	27.10.2022	2	ne	10	13.1	21.9	8.0	5.3	1
27	22.1	28.10.2022	3	ne	10	12.9	22.0	6.0	2.7	1
28	28.7	29.10.2022	4	ne	10	14.3	20.0	9.3	6.7	1
29	23.6	30.10.2022	5	ne	10	13.0	21.0	8.8	7.6	1

Slika 6: Tabela podatkov (lasten vir)

Pretvorna skupina



Slika 7: Widgeti pretvorne skupine (lasten vir)

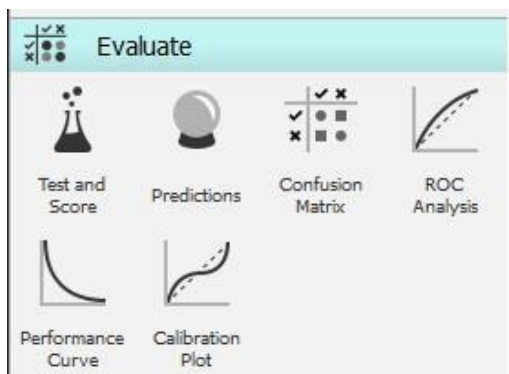
Pretvorna skupina je bila ključna za manipulacijo z našimi podatki in njihovo prilagoditev za nadaljnjo obdelavo. Tukaj smo uporabili več funkcij, ki so omogočale manipulacijo z vrsticami in stolpci ter omogočale boljšo strukturo podatkov. Pomemben del v tem koraku je bila uporaba "Select column" in "Select rows" widgetov, saj smo

lahko natančno določili, katere stolpce in vrstice bomo uporabili kot cilje ali meta podatke. Te manipulacije so bile ključne, saj smo lahko natančno definirali podatke, ki so bili pomembni za naše modeliranje.

Z uporabo "Preprocess" widgeta smo izvedli dodatne obdelave, kot je imputacija manjkajočih vrednosti, naključno izbiranje, diskretizacija zveznih spremenljivk, kar nam je omogočilo, da smo podatke optimalno prilagodili za naslednje korake.

Evaluacijska skupina

Ko smo obdelali podatke, smo jih prenesli v evaluacijsko skupino, kjer smo začeli s testiranjem in vrednotenjem modelov. Ta faza je bila zelo pomembna, saj smo lahko preizkusili različne testne metode in jih povezali z našimi modeli, kot so nevronske mreže in naključni gozd. Za izračun natančnosti modelov smo uporabili "Predictions" in "Confusion matrix", kar nam je omogočilo, da smo lahko natančno ocenili, kako dobro so naši modeli delovali na testnih podatkih.



Slika 8: Evaluacijska skupina (lasten vir)

Skupina modelov



Slika 9: Slika strojnih modelov (lasten vir)

Ko smo se odločili za določene teste in evaluacijo, smo prešli na fazo modeliranja, kjer smo preizkusili več različnih modelov strojnega učenja. Naši modeli so bili zasnovani tako, da so omogočali prilagodljivost pri obdelavi zahtevnih podatkov in iskanju vzorcev v njih. Preizkusili smo različne pristope, kot so nevronske mreže in naključni gozd, saj sta se izkazala za zelo natančna pri napovedovanju. V tej fazi smo se osredotočili na analizo rezultatov, ki so bili včasih zelo različni pri uporabi iste količine podatkov. Ta visoka variabilnost med rezultati nas je spodbudila k temu, da smo temeljiteje raziskali, kateri model bi bil najbolj primeren za našo nalogo, saj so razlike v napovedih močno vplivale na končne odločitve o tem, kateri model bomo uporabljali v nadaljevanju raziskave.

Uporabljeni modeli strojnega učenja

V tej raziskavi smo uporabili več različnih vrst modelov strojnega učenja, da bi ugotovili, kateri najbolj natančno napoveduje rezultate in s tistim modelom nadaljevali našo raziskavo. Modele smo izbrali zaradi njihove raznolikosti, prilagodljivosti in sposobnosti obdelave zahtevnih podatkov in povezav med njimi. Tukaj so opisani uporabljeni modeli, njihove prednosti, slabosti in kako delujejo v različnih situacijah.

Glavni dejavnik za globljo raziskavo naštetih modelov je bil to, da so bile razlike med rezultati pri isti količini podatkov visoke. To je pogosto vodilo v nejasnosti pri odločitvi izbire modela na katerem bo temeljilo naše delo.

Nevronska mreža

Nevronska mreža je zasnovana za učenje iz izkušenj, prilagajanje na podlagi danih podatkov in je osrednji koncept strojnega učenja, saj predstavlja nevronske veze v človeških možganih. Ta vozlišča, ki jih lahko vidimo kot nevrone, so povezana med seboj preko uteži, ki predstavljajo moč povezave med vsakim nevronom, kar ima za posledico generiranje izhoda, ki služi kot vhod za naslednje plasti nevronov v omrežju.

Način učenja prek prilagoditvenih uteži med nevroni omogoča omrežju, da optimizira svoje odločitve glede nabora nalog. Predstavlja krivuljo učenja skozi obdelavo vnaprej določenega cilja in določa težo posameznika, zato je poleg tega, da je zelo prilagodljiv, tudi zelo močan, glede na to, da nenehno eksperimentirajo s kompleksnimi podatki.

V Orange-u ta pripomoček uporablja večplastni perceptronski algoritem sklearn, ki se lahko nauči nelinearnih modelov in tudi linearnih, medtem pa identificira aktivacije brez operacije in uporablja različne logistike, kot je sigmoidna funkcija, da pride do odgovora.

Linearna regresija

Linearna regresija je algoritem z izbirno L1, L2 in L1L2 regulacijo, z algoritmom učenja linearne regresije, ki konstruira učenca ali napovedovalec, ki se uči iz svoje linearne funkcije z vhodnim naborom podatkov.

L1 ali Lasso doda vsoto absolutnih vrednosti koeficientov funkciji izgube.

$$Izguba = MSE + \lambda \sum |w_i|$$

Equation 1: Lasso enačba w_i -

koeficienti modela λ -

Nastavitveni parameter

Regresija L2 ali Ridge je še ena metoda laso, ki skrči vse koeficiente, vendar jih ne postavi na nič, hkrati pa zmanjša multikolinearnost in pomaga funkcijam prispevati k napovedim.

$$Izguba = MSE + \lambda \sum w_i^2$$

Equation 2: Ridge enačba

L1 L2 ali elastična mreža je kombinacija kazni L2 in L2, ki so uporabne za številne korelirane funkcije, in zagotavlja prilagodljiv kompromis med L1 in L2.

$$Izguba = MSE + \lambda_1 \sum |w_i| + \lambda_2 \sum w_i^2$$

Equation 3: Elastična enačba

Naključni gozd

Naključni gozd predvideva z uporabo niza odločitvenih dreves z vhodnim naborom podatkov in metodo predprocesiranja. Ta model naredi natanko to, kar se imenuje, zgradi nabor naključnih odločitvenih dreves, pri čemer se vsako drevo razvije iz zagonskega vzorca iz fata usposabljanja. Pri razvoju teh dreves se nariše poljubna podmnožica atributov in izbere najboljši atribut za razdelitev. Končni model temelji na večini glasov posamično razvitih dreves v gozdu.

10. Rezultati

Za oceno uspešnosti naših napovednih modelov smo uporabili klasična merila za strojno učenje. Med izbranimi so bila: korelacijski koeficient, povprečna absolutna napaka (MAE), povprečna kvadratna napaka (MSE) in kvadratna srednja napaka (RMSE). Za lažje prikazovanje in interpretiranje rezultatov bodo tudi vizualno prikazani.

11. Primerjava modelov

V okviru našega raziskovanja smo preizkusili več modelov strojnega učenja, da bi našli najprimernejši pristop za napovedovanje koncentracije prašnih delcev v zraku. Za vsakega od teh modelov smo izvedli obsežno testiranje tako v programu WEKA kot Orange, da bi ugotovili, kateri model bi dal najboljše rezultate. V nadaljevanju so

opisani rezultati testiranj in pristopi, ki smo jih uporabili, da bi optimizirali naše modele in izboljšali napovedi.

Linearna regresija

WEKA

Model je dosegel koeficient determinacije 0,989, kar pomeni, da lahko pojasni 99 % razlik v podatkih. Kljub temu je napisano da povprečna absolutna napaka znaša 13,87 %.

```
=== Summary ===
Correlation coefficient          0.9894
Mean absolute error             1.7722
Root mean squared error         2.3334
Relative absolute error         13.8712 %
Root relative squared error     14.7006 %
Total Number of Instances      144
Ignored Class Unknown Instances 11
```

Slika 10: Rezultati Linearne regresije v WEKI (lasten vir)

Orange

Ko primerjamo modele, se je linearna regresija izkazala za najboljšo izbiro, saj je imela najnižjo napako. Njen RMSE je bil $6.273 \mu\text{g}/\text{m}^3$, MAE pa $4.836 \mu\text{g}/\text{m}^3$, kar pomeni, da so bile njene

Model	MSE	RMSE	MAE	MAPE	R2
Neural Network	40.795	6.387	4.898	0.264	0.508
Linear Regression	39.344	6.273	4.836	0.260	0.525
Random Forest	41.439	6.437	4.862	0.262	0.500
SVM	55.454	7.447	5.644	0.285	0.331

Slika 11: Rezultata Linearne regresije v Orange-u (lasten vir)

napovedi precej natančne. Tudi vrednost MSE ($39.344 \mu\text{g}/\text{m}^3$) potrjuje, da model dobro zajame vzorce v podatkih.

Naključni gozd (Naključni gozd)

WEKA:

Model naključnega gozda je v Weki dosegel determinacijski koeficient 0,954 $\mu\text{g}/\text{m}^3$, vendar je imel precej visoko povprečno absolutno napako 30,3 %. Proces "Bagginga" v našem primeru ni bil učinkovit.

```
=== Summary ===
Correlation coefficient      0.9542
Mean absolute error        3.8717
Root mean squared error    5.8489
Relative absolute error    30.3038 %
Root relative squared error 36.8489 %
Total Number of Instances  144
Ignored Class Unknown Instances 11
```

Slika 12: Random fores v WEKI (lasten vir)

Orange:

Pri Naključni gozdu so bile napake malenkost večje, RMSE (6.437 $\mu\text{g}/\text{m}^3$), MAE (4.862 $\mu\text{g}/\text{m}^3$) in MSE (41.439 $\mu\text{g}/\text{m}^3$). To kaže, da bi lahko z boljšimi nastavitvami hiperparametrov model še izboljšali.

Model	MSE	RMSE	MAE	MAPE	R2
Neural Network	40.795	6.387	4.898	0.264	0.508
Linear Regression	39.344	6.273	4.836	0.260	0.525
Random Forest	41.439	6.437	4.862	0.262	0.500
SVM	55.454	7.447	5.644	0.285	0.331

Slika 13: Naključni gozd v Orange (lasten vir)

Metoda podpornih vektorjev (SVM)

WEKA:

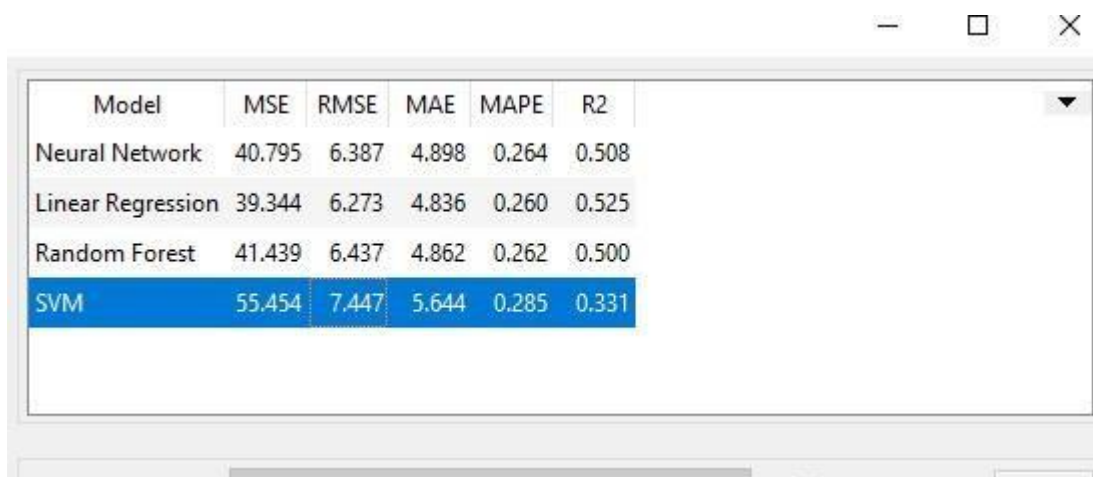
Rezultati SVM so bili zadovoljivi. Natančnost korelacijskega koeficienta je dosegla 98,9 %, kar nas je pozitivno presenetilo. Povprečna absolutna napaka je znašala 14,15 %, kar je dokaj dober rezultat.

```
=== Summary ===  
  
Correlation coefficient          0.9887  
Mean absolute error             1.8077  
Root mean squared error        2.3872  
Relative absolute error        14.1492 %  
Root relative squared error    15.0397 %  
Total Number of Instances      144  
Ignored Class Unknown Instances 11
```

Slika 14: SVM v WEKI (lasten vir)

Orange:

Najslabše se je odrezal SVM, saj je imel največje napake, RMSE ($7.447 \mu\text{g}/\text{m}^3$), MAE ($5.644 \mu\text{g}/\text{m}^3$) in MSE ($55.454 \mu\text{g}/\text{m}^3$). Očitno ta model ni tako dobro zajel značilnosti podatkov, zato bi ga bilo smiselno prilagoditi ali zamenjati.



Model	MSE	RMSE	MAE	MAPE	R2
Neural Network	40.795	6.387	4.898	0.264	0.508
Linear Regression	39.344	6.273	4.836	0.260	0.525
Random Forest	41.439	6.437	4.862	0.262	0.500
SVM	55.454	7.447	5.644	0.285	0.331

Slika 15: SVM v Orangeu (lasten vir)

Nevronska mreža

WEKA:

Nevronske mreže so v Weki dosegale slabše rezultate kot drugi modeli. Korelacijski koeficient je znašal le 97,3 %, medtem ko je relativna absolutna napaka znašala kar 25 %.

```

=== Summary ===

Correlation coefficient          0.973
Mean absolute error            3.1995
Root mean squared error        4.1778
Relative absolute error        25.0423 %
Root relative squared error    26.3206 %
Total Number of Instances      144
Ignored Class Unknown Instances 11
  
```

Slika 16: Nevronska mreža v WEKI (lasten vir)

Orange:

Nevronska mreža se je odrezala zelo podobno, saj je njen RMSE znašal $6.387 \mu\text{g}/\text{m}^3$, MAE $4.898 \mu\text{g}/\text{m}^3$, MSE pa $40.795 \mu\text{g}/\text{m}^3$. To pomeni, da je skoraj enako dobra kot linearna regresija, a vseeno rahlo slabša.

Model	MSE	RMSE	MAE	MAPE	R2
Neural Network	40.795	6.387	4.898	0.264	0.508
Linear Regression	39.344	6.273	4.836	0.260	0.525
Random Forest	41.439	6.437	4.862	0.262	0.500
SVM	55.454	7.447	5.644	0.285	0.331

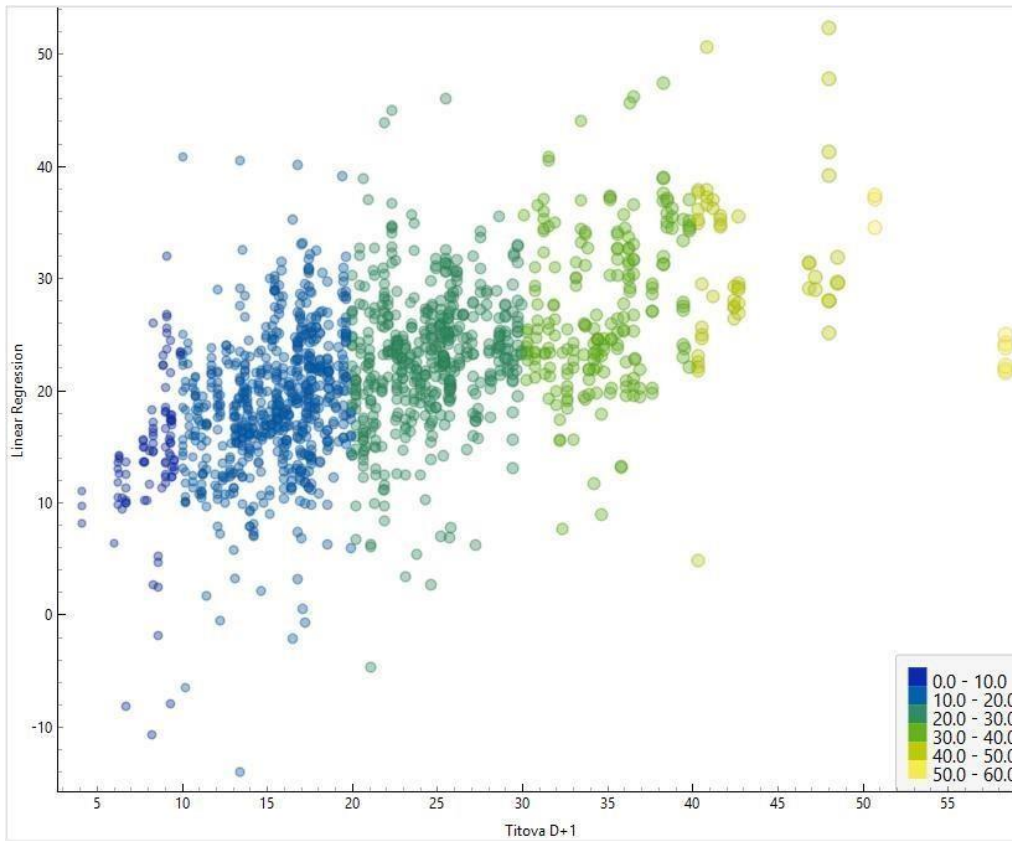
Slika 17: Nevronska mreža v Orangeu (lasten vir)

Vizualizacija rezultatov

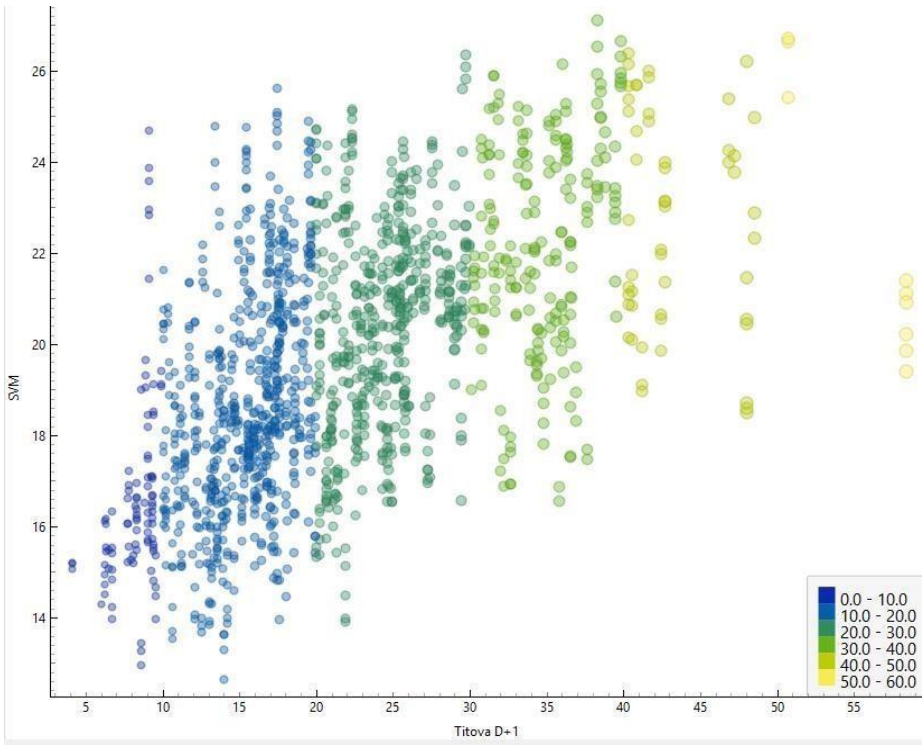
Za boljšo ponazoritev rezultatov smo uporabili grafe in diagrame, ki prikazujejo primerjave med dejanskimi in napovedanimi vrednostmi.

WEKA žal ne ponuja orodij prek katerih bi lahko naše rezultate predstavili, zato bodo v nadaljevanju vizualno prikazani rezultati le v Orange-u.

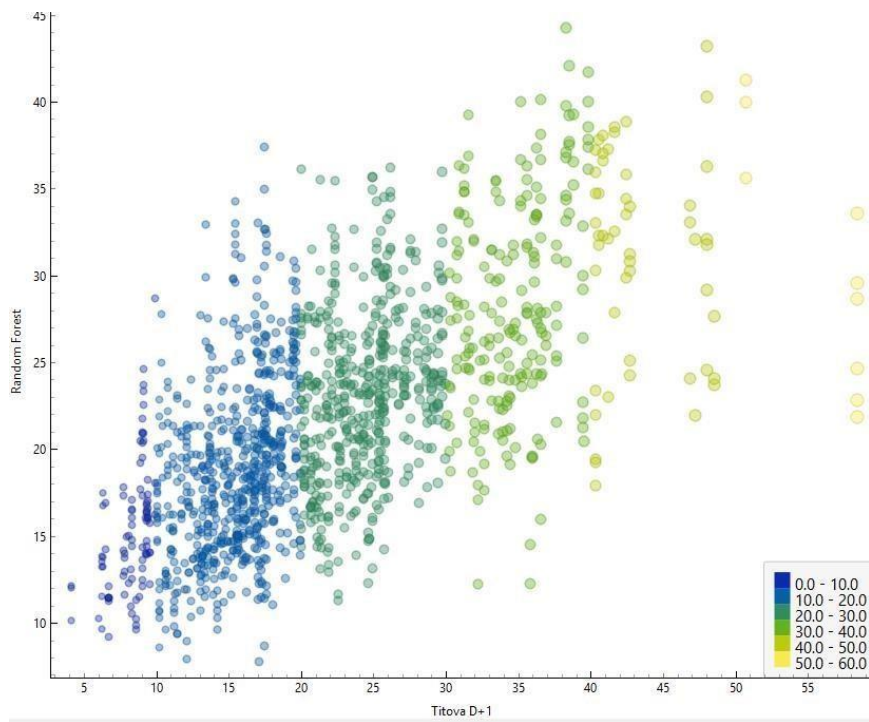
Podatki so predstavljeni tako, za uporabo gradnika (widget) za prikaz matrike raztresenosti (Confusion Matrix). Na slikah 18,19, 20,21 je prikazana razlika med posameznimi napovednimi modeli. Na x oseh so prikazane iskalne vrednosti, na y oseh pa napovedane.



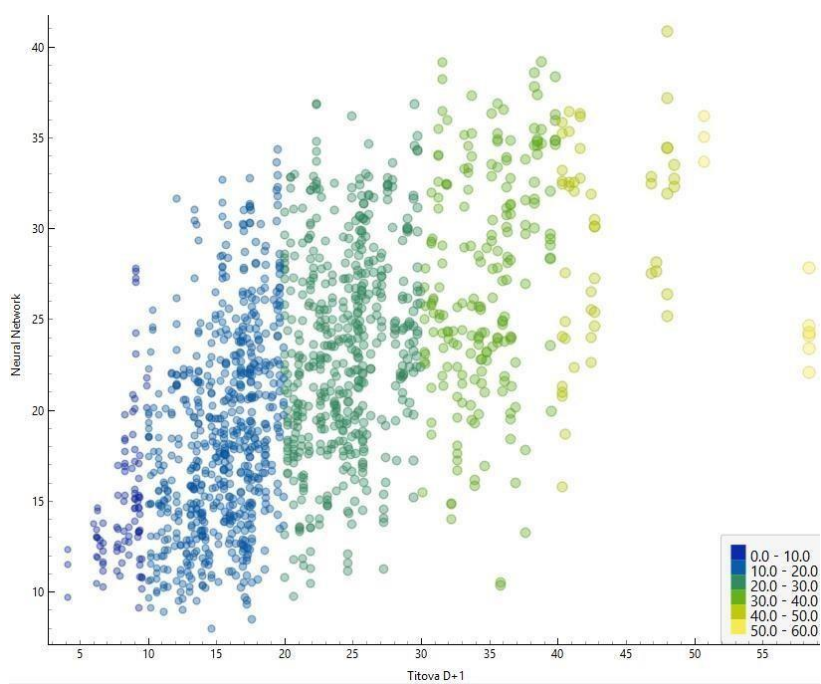
Slika 18: Vizualizacija delovanja linearne regresije (lasten vir)



Slika 19: Vizualizacija delovanja SVM-a (lasten vir)



Slika 20: Vizualizacija delovanja Naključni gozd (lasten vir)



12. Diskusija

Med raziskovanjem smo ugotovili da so nekateri parametri podani v neprimerni obliki. Ko smo datum razdelili na leto, mesec in dan ter dodali dan v tednu, je natančnost napovedovanja narastla. S to razdelitvijo datuma je model lahko upošteval značilnosti dni v tednu, mesecev in let.

Vplivnost atributov smo določili z orodjem v WEKA-i. Nato smo manj vplivne attribute izločili.

Med uporabo orodji Orange in WEKA, smo ugotovili, da Orange omogoča večjo fleksibilnost in vključuje široko ponudbo funkcij za zapolnitev manjkajočih vrednosti, odstranjevanje nepomembnih podatkov, urejanje informacij in vizualizacijo rezultatov, s čimer bistveno presega zmogljivosti orodja WEKA. Poleg tega je v Orange-u hitrejše tudi strojno učenje in možna uporaba datotek z različnih formatov, kar močno olajša delo. Prav tako nam je Orange omogočil bolj pregledno primerjavo rezultatov, kar je močno pospešilo naš proces analize.

Naši modeli uspešno napovedujejo stopnjo onesnaženosti zraka z delci PM10. Najmanjša napaka najboljšega modela je dosegla $5,3 \mu\text{g}/\text{m}^3$. Ugotovili smo, da je optimizacija vhodnih podatkov, ter izbira podatkov, ki so uporabljeni bistveno pomembnejša kot smo sprva pričakovali.

Potrjevanje hipotez

Hipoteze:
1) Ugotavljali bomo ali podatki vsebujejo nerelavantne attribute. To so atributi za katere predvidevamo da nima ključnega vpliva na natančnost napovedovanja.
2) Napako napovedovanja onesnaženosti zraka z delci PM10 bomo poskušali zmanjšati na vrednost manj kot $8 \mu\text{g}/\text{m}^3$ (natančnost FERl-ja).
3) Ugotavljali bomo katera vrsta napovednega modela dosega najboljše rezultate.
4) Ugotavljali bomo s katerim orodjem (WEKA ali Orange) lahko dosežemo boljše rezultate.

Potrditve hipotez

1. Najboljši rezultat smo dosegli ko so bili določeni atributi izločeni, s tem je prva hipoteza potrjena.
2. Napako napovedovanja onesnaženosti zraka PM10 smo zmanjšali na vrednost $6,3 \mu\text{g}/\text{m}^3$. S tem je druga hipoteza potrjena.
3. V tretji hipotezi ugotavljamo, da najboljše rezultate napovedovanja nudi model linearna regresija (Linear Regression).
4. Ugotovili smo, da je orodje Orange uporabnejše zaradi boljše predstavitev rezultatov in hitrejšega strojnega učenja.

Razprava o rezultatih

Analiza rezultatov je pokazala, da sta izbira pravega modela in kakovost vhodnih podatkov ključna za natančnost napovedi.

- **Najboljše se je izkazala linearna regresija**, saj je imela najmanjše napake (RMSE: $6.273 \mu\text{g}/\text{m}^3$, MAE: $4.836 \mu\text{g}/\text{m}^3$, MSE: $39.344 \mu\text{g}/\text{m}^3$).
- **Nevronska mreža** je bila zelo blizu, ampak še vedno nekoliko manj točna (RMSE: $6.387 \mu\text{g}/\text{m}^3$, MAE: $4.898 \mu\text{g}/\text{m}^3$, MSE: $40.795 \mu\text{g}/\text{m}^3$).
- **Naključni gozd** je imel nekoliko večje napake (RMSE: $6.437 \mu\text{g}/\text{m}^3$, MAE: $4.862 \mu\text{g}/\text{m}^3$, MSE: $41.439 \mu\text{g}/\text{m}^3$), a je kljub temu dajal solidne rezultate.
- **Najslabše se je odrezal SVM**, kjer so bile napake precej večje (RMSE: $7.447 \mu\text{g}/\text{m}^3$, MAE: $5.644 \mu\text{g}/\text{m}^3$, MSE: $55.454 \mu\text{g}/\text{m}^3$), kar kaže, da ta model ni bil najbolj primeren za ta nabor podatkov.

Pomembno je, da so modeli ustrezno prilagojeni specifičnim podatkom, saj v nasprotnem že manjše spremembe vhodnih podatkov lahko občutno poslabšajo natančnost napovedi.

	error	Linear Regression	error	Random Forest	error	SVM	error	Titova D+1	RefDatum	st dni brez dezia	dez	Mesec	ovpnevna T [°C]	maxT [°C]	minT [°C]	ninT na Scm [°C]	ategorija_Vrba	
1	185.1	4.8	17.9	4.2	14.9	1.2	17.5	3.8	13.7	2.10.2022	0.00	da	0.8182	0.60064	0.62745	0.56146	0.54839	1
2	180.1	1.2	19.2	2.4	15.6	-1.2	19.4	2.6	16.8	3.10.2022	0.0455	ne	0.8182	0.52716	0.61064	0.58472	0.55484	1
3	22.3	1.0	23.8	2.4	19.4	-1.9	21.5	0.2	21.3	4.10.2022	0.0909	ne	0.8182	0.48562	0.57423	0.41196	0.39677	1
4	24.7	-0.5	23.8	-1.4	26.4	1.2	20.9	-4.3	25.2	5.10.2022	0.1364	ne	0.8182	0.54633	0.61064	0.44518	0.42903	1
5	29.1	1.6	29.6	2.1	27.9	0.4	23.5	-4.0	27.5	6.10.2022	0.1818	ne	0.8182	0.58147	0.66106	0.52824	0.50968	1
6	26.1	0.3	28.2	2.4	27.0	1.2	23.2	-2.6	25.8	7.10.2022	0.2273	ne	0.8182	0.60703	0.63585	0.57807	0.54839	1
7	22.2	4.0	25.3	7.1	22.5	4.3	22.1	3.9	18.2	8.10.2022	0.2727	ne	0.8182	0.57508	0.62465	0.57143	0.54839	1
8	30.0	2.1	25.5	-2.4	24.1	-3.8	21.9	-6.0	27.9	9.10.2022	0.3182	ne	0.8182	0.52716	0.51541	0.60465	0.62903	1
9	31.4	-6.2	28.0	-9.6	30.2	-7.4	25.4	-12.2	37.6	10.10.2022	0.3636	ne	0.8182	0.51757	0.54062	0.54153	0.50645	1
10	36.1	3.0	34.8	1.7	32.6	-0.5	27.2	-5.9	33.1	11.10.2022	0.4091	ne	0.8182	0.59744	0.60224	0.52492	0.50645	1
11	35.2	-1.0	32.9	-3.3	33.3	-2.9	27.0	-9.2	36.2	12.10.2022	0.4545	ne	0.8182	0.51118	0.57983	0.54485	0.51613	1
12	38.7	-8.1	35.9	-10.9	41.5	-5.3	28.4	-18.4	46.8	13.10.2022	0.50	ne	0.8182	0.50160	0.57703	0.48173	0.47419	1
13	41.6	-5.6	38.0	-9.2	46.3	-0.9	28.2	-19.0	47.2	14.10.2022	0.5455	ne	0.8182	0.48243	0.55182	0.47508	0.46452	1
14	37.8	-1.7	36.1	-3.4	38.3	-1.2	25.9	-13.6	39.5	15.10.2022	0.5909	ne	0.8182	0.53674	0.51261	0.54153	0.53871	1
15	41.9	2.1	37.4	-2.4	37.6	-2.2	29.0	-10.8	39.8	16.10.2022	0.6364	ne	0.8182	0.55591	0.64146	0.55614	0.55484	1
16	42.2	3.7	36.2	-2.2	42.4	3.9	28.8	-9.7	38.5	17.10.2022	0.6818	ne	0.8182	0.56230	0.66947	0.52824	0.53226	1
17	39.4	7.9	37.3	5.8	39.1	7.6	28.7	-2.8	31.5	18.10.2022	0.7273	ne	0.8182	0.57188	0.68867	0.51163	0.51290	1
18	31.2	8.9	32.0	9.7	24.9	2.6	26.0	3.7	22.3	19.10.2022	0.7727	ne	0.8182	0.55272	0.56863	0.58804	0.57419	1
19	34.7	-0.9	30.8	-4.8	35.3	-0.3	24.4	-11.2	35.6	20.10.2022	0.8182	ne	0.8182	0.51438	0.50980	0.62126	0.64839	1
20	22.6	1.3	30.8	9.5	18.2	-3.1	23.1	1.8	21.3	21.10.2022	0.8636	ne	0.8182	0.51438	0.54902	0.48837	0.50323	1
21	17.8	-2.8	14.9	-5.7	22.5	1.9	17.7	-2.9	20.6	22.10.2022	0.00	da	0.8182	0.55591	0.55182	0.61794	0.70323	1
22	29.2	-11.1	24.8	-15.5	30.9	-9.4	20.9	-19.4	40.3	23.10.2022	0.0455	ne	0.8182	0.57508	0.62745	0.49834	0.53226	1
23	17.9	3.7	25.4	11.2	22.4	8.2	18.6	4.4	14.2	24.10.2022	0.00	da	0.8182	0.65815	0.63025	0.64784	0.63871	1
24	18.4	-4.7	17.1	-6.0	22.6	-0.5	16.9	-6.2	23.1	25.10.2022	0.00	da	0.8182	0.55272	0.59664	0.62458	0.75484	1
25	21.6	-0.1	25.1	3.4	22.3	0.6	21.4	-0.3	21.7	26.10.2022	0.0455	ne	0.8182	0.53355	0.58263	0.50166	0.49677	1
26	28.8	-0.2	27.6	-1.4	28.0	-1.0	22.9	-6.1	29.0	27.10.2022	0.0909	ne	0.8182	0.55272	0.64986	0.56146	0.55161	1
27	28.3	6.2	27.5	5.4	26.6	4.5	24.1	2.0	22.1	28.10.2022	0.1364	ne	0.8182	0.54633	0.65266	0.49502	0.46774	1

Slika 22: Primerjava modelov v Orange-u (lasten vir)

Analiza dela

Naša raziskovalna naloga se je osredotočala na čim bolj uspešno napovedovanje stanja PM10 delcev v zraku, saj ti močno vplivajo na naše zdravje. Naš glavni cilj je bil analizirati podatke do takšne mere, da lahko naš napovedni model čim bolje napove kakovost zraka za naslednji dan. To smo dosegli tako, da smo primerjali rezultate pridobljene z različnimi parametri, pri čemer smo se poglobili v znanje o delcih PM10, kako ti nastajajo, ter kaj vpliva na njihovo količino (npr. vremenske razmere, promet in industrijska dejavnost).

Iskanje najboljšega modela za napoved PM10

Pri napovedovanju PM10 smo preizkusili različne modele strojnega učenja in jih primerjali glede na njihovo natančnost. Najbolje se je izkazala linearna regresija, ki je imela najmanjše napake in se je pokazala kot najprimernejši model za naš nabor podatkov. Nevronska mreža se ji je močno približala, a je bila nekoliko manj natančna. Naključni gozd je dal solidne rezultate, vendar z nekoliko večjimi napakami. Najslabše pa se je odrezal model podpornih vektorjev (SVM), kjer so bile napake občutno večje, kar kaže, da ta pristop ni bil najbolj učinkovit. Rezultati jasno kažejo, kako pomembna je izbira pravega modela in kakovost vhodnih podatkov. Tudi najmanjše spremembe pri čiščenju podatkov ali nastavitvah modela lahko bistveno vplivajo na natančnost napovedi.

Raziskava napovedovanja ni pomembna samo s tehničnega vidika ampak tudi je njeni rezultati lahko pripomorejo k boljšemu spremljanju kakovosti zraka. Če lahko PM10 napovemo natančneje, lahko pomagamo pri sprejemanju ukrepov za izboljšanje kakovosti zraka, zmanjšanju onesnaženja in s tem zaščiti javnega zdravja.

Družbena odgovornost in trajnostni razvoj

Naša raziskava ni zgolj prikaz modela, ki je sposoben napovedovanja, temveč s to raziskavo pokažemo tudi, kako mi sami vplivamo na onesnaženost zraka, ki ga dihamo. Iz raziskave je razvidno da človekov vpliv na okolje ni majhen in da je naše navade potrebno spremeniti, preden bo prepozno.

Z boljšim vpogledom v trenutno stanje kakovosti zraka so nam odprte nove možnosti o odločanju, kdaj je možno prosto gibanje brez skrbi. Na primer, posamezniki z boleznimi dihal bodo sposobni presoditi, ali je varno stopiti na prosto brez maske, športniki pa bodo prilagodili svoje treninge glede na trenutno kakovost zraka. Prav tako se lahko okoljevarstveniki na podlagi teh podatkov odločijo, kdaj in kje je potrebno zostriti ukrepe za zmanjševanje našega onesnaževanja. S tem raziskava postane več kot le zbirka podatkov, temveč postane ključni vir znanja, ki spodbuja človeka k ozaveščanju in odgovornem ravnanju. Ko ljudje vidijo neposredne učinke slabega zraka na zdravje, kot so povečana tveganja za bolezni dihal, srčno-žilne težave in celo vplivi na delovanje možganov, se bodo začele spreminjati njihove navade. Možno je, da se bodo ljudje odločili za uporabo javnega prevoza ali kolesarjenje, s čimer bodo prispevali k zmanjšanju emisij. Poleg tega bi več oseb podprlo ukrepe, kot so uporaba čistejših virov energije.

Z nadaljnjim razvojem našega modela, bi v prihodnosti bilo možno prek mobilnih aplikacij prebivalstvo v realnem času opozarjati o kakovosti zraka in jim nasvetovati, kako naj se zaščitijo, na primer s spremembo časa odhoda ali uporabo mask.

Ena izmed največjih prednosti uporabe umetne inteligence pri spremljanju kakovosti zraka je njena sposobnost stalnega učenja in prilagajanja. Z zbiranjem vedno več podatkov, od aktivnosti prometa do onesnaževanja industrijskih objektov, bo model vedno natančneje napovedoval obdobja, kadar je zrak bolj ali manj onesnažen.

S pravilnim napovedovanjem pripomoremo tudi k prihodnjem razvoju gospodarstva in zdravstva. Natančnejši modeli za napovedovanje onesnaženosti bi lahko spodbujali

nova, bolj "zelena", odkritja v prometu, energetiki in industriji. S tem bi lahko pripomorili k razvoju krožnega gospodarstva, kjer se viri uporabljajo učinkoviteje, odpadki pa se zmanjšujejo na minimum. Poleg tega ima čistejši zrak neposreden vpliv na javno zdravje in manjše onesnaženje zraka pomeni manj bolezni, manj potrebnega zdravljenja in s tem nižje stroške zdravstvene oskrbe.

Trajnost

Ta raziskava ni zgolj združevanje podatkov z različnimi modelov, temveč poglobljena analiza vpliva izbranih metod strojnega učenja na natančnost napovedi. S povezovanjem znanosti, tehnologije in učnih podatkov smo raziskali, kako lahko prilagajanje parametrov modelov prispeva k zmanjšanju napake in s tem k bolj zanesljivim napovedim. Ko združimo znanje in prizadevanja, lahko naredimo korak k čistejšemu in varnejšemu okolju ter prispevamo k trajnostni prihodnosti za nas in prihodnje generacije.

13. Viri

1. Board, C. A. (brez datuma). *Inhalable Particulate Matter and Health (PM2.5 and PM10)*.
 - a. Pridobljeno iz California Air Resources Board:
 2. <https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health>
3. Mining, O. D. (brez datuma). *Widget Catalog*. Pridobljeno iz Widget Catalog: <https://orangedatamining.com/widget-catalog/>
4. Okolje, A. -A. (brez datuma). *Republika Slovenija- Ministarstvo za Okolje, Podnebje in Energijo*. Pridobljeno iz Agencija Republike Slovenije za Okolje:
 - a. <http://arso.si/zrak/kakovost%20zraka/>
5. OpenAI, C. (7. February 2025). *Kaj pomeni RSE, RMSE, MAE, MAPE in R2?* Pridobljeno iz <https://www.openai.com>
6. Ploj, B. (2019). *Bionska umetna inteligenca*: Knjiga o napravah, ki se učijo in se samostojno odločajo, pri čemer lahko presežejo tudi človeštvo. Založnik Visoka šola na Ptuju.

14. Viri slik

Slika 1: Inhalable Particulate Matter and Health (dostopno na <https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health>, 7. februar 2025, nastanek: 2025)