Procjena indeksa kvalitete zraka za lebdeće čestice PM10 i PM2,5 grada Zagreba primjenom metoda strojnog učenja

Klonkay, Paola

Master's thesis / Diplomski rad

2024

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Chemical Engineering and Technology / Sveučilište u Zagrebu, Fakultet kemijskog inženjerstva i tehnologije**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:149:578223

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
SVEUČILIŠNI DIPLOMSKI STUDIJ

Paola Klonkay

DIPLOMSKI RAD

Zagreb, rujan 2024.

SVEUČILIŠTE U ZAGREBU
FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE
POVJERENSTVO ZA DIPLOMSKE ISPITE

Kandidatkinja **Paola Klonkay**

Predala je izrađen diplomski rad dana: 23. rujna 2024.

Povjerenstvo u sastavu:

Izv. prof. dr. sc. Željka Ujević Andrijić, Sveučilište u Zagrebu Fakultet kemijskog inženjerstva i tehnologije
Izv. prof. dr. sc. Marin Kovačić, Sveučilište u Zagrebu Fakultet kemijskog inženjerstva i tehnologije
Dr. sc. Silvije Davila, znan. sur., Institut za medicinska istraživanja i medicinu rada, Zagreb
Prof. dr. sc. Nenad Bolf, Sveučilište u Zagrebu Fakultet kemijskog inženjerstva i tehnologije (zamjena)

povoljno je ocijenilo diplomski rad i odobrilo obranu diplomskog rada pred povjerenstvom u istom sastavu.

Diplomski ispit održat će se dana: 26. rujna 2024.

SVEUČILIŠTE U ZAGREBU

FAKULTET KEMIJSKOG INŽENJERSTVA I TEHNOLOGIJE

SVEUČILIŠNI DIPLOMSKI STUDIJ

Paola Klonkay

Procjena indeksa kvalitete zraka za lebdeće čestice PM_{10} i $PM_{2,5}$ grada Zagreba primjenom metoda strojnog učenja

Mentor: izv. prof. dr. sc. Željka Ujević Andrijić, Fakultet kemijskog inženjerstva i tehnologije

Komentor: dr.sc. Silvije Davila, Institut za medicinska istraživanja i medicinu rada (Zavod za higijenu okoliša)

Članovi ispitnog povjerenstva: izv. prof. dr. sc. Željka Ujević Andrijić

izv. prof. dr. sc. Marin Kovačić

dr. sc. Silvije Davila

Zagreb, rujan 2024.

Ovaj diplomski rad izrađen je na Zavodu za higijenu okoliša na Institutu za medicinska istraživanja i medicinu rada pod neposrednim vodstvom doktorandice Marije Jelene Lovrić Štefiček i komentorstvom dr. sc. Silvija Davile, te na Zavodu za mjerenja i automatsko vođenje procesa Fakulteta kemijskog inženjerstva i tehnologije Sveučilišta u Zagrebu pod mentorstvom izv. prof. dr. sc. Željke Ujević Andrijić.

Diplomski rad izrađen je na opremi i resursima projekta KK.01.1.1.02.0007 "Istraživačko-edukacijski centar za zdravstvenu i medicinsku ekologiju i zaštitu od zračenja - rekonstrukcija i dogradnja Instituta za medicinska istraživanja i medicinu rada", tj. KK.01.1.1.02.0007 "Research and Education Centre of Environmental Health and Radiation Protection – Reconstruction and Expansion of the Institute for Medical Research and Occupational Health", i financiran iz sredstava Europske unije – NextGenerationEU (Programski ugovor od 8. prosinca 2023. KLASA: 643-02/23-01/00016, URBROJ: 533-03-23-0006)-EnvironPollutHealth. Za izradu ovog diplomskog rada korišteni su javno dostupni podaci s mjernih postaja državne mreže za trajno praćenje kvalitete zraka u nadležnosti Ministarstva zaštite okoliša i zelene tranzicije kojom upravlja Državni hidrometeorološki zavod, a mjerenja financira Fond za zaštitu okoliša i energetsku učinkovitost.

Zahvala:

Želim izraziti zahvalnost Državnom hidrometeorološkom zavodu (DHMZ) za meteorološke podatke. Njihov rad i podrška omogućili su prikupljanje i analizu podataka o kvaliteti zraka u Republici Hrvatskoj, što je ključno za ovo istraživanje.

Iskrena zahvala ide mojoj mentorici, izv. prof. dr. sc. Ujević Andrijić, na Fakultetu kemijskog inženjerstva i tehnologije, za stručno vodstvo i podršku tijekom izrade ovog rada, kao i dr. sc. Silviju Davili na pomoći i prenesenom znanju. Vaši savjeti i smjernice bili su od iznimne važnosti. Također, srdačno zahvaljujem doktorandici Mariji Jeleni Lovrić Štefiček na pomoći i strpljenju. Vaša podrška i stručni savjeti uvelike su doprinijeli kvaliteti ovog rada. Posebna zahvala ide asistentu Nikoli Rimcu koji me uputio u osnove strojnog učenja na predavanjima iz kolegija Metode umjetne inteligencije u kemijskom inženjerstvu (MUI) na Fakultetu kemijskog inženjerstva i tehnologije.

Hvala svima koji su svojim doprinosom i podrškom omogućili izradu ovog rada. Na kraju, hvala i mojoj obitelji i prijateljima na njihovoj konstantnoj podršci tijekom mog cjelokupnog akademskog obrazovanja.

SAŽETAK

Onečišćenje zraka predstavlja veliki problem današnjice, kako za okoliš tako i za ljudsko zdravlje. Zato je cilj ovog diplomskog rada razviti modele za procjenu dnevnog indeksa kvalitete zraka (AQI) koji se odnosi na onečišćenje vanjskog zraka lebdećim česticama PM_{10} i $PM_{2,5}$ (lebdeće čestice ekvivalentnog aerodinamičnog promjera manjeg od 10 i 2,5 μm) koristeći metode nadziranog strojnog učenja temeljene na klasifikaciji. Za razvoj modela primijenjene su metoda nasumičnih šuma i metoda potpornih vektora, koristeći programski jezik Python i pripadajuće knjižice. Korišteni su podaci o masenim koncentracijama frakcija lebdećih čestica PM_{10} i $PM_{2,5}$ i meteorološki podaci s tri mjerne postaje državne mreže za trajno praćenje kvalitete zraka. Uz lokalne dnevne meteorološke podatke korišteni su i dnevni podaci prikupljeni na meteorološkoj postaji Maksimir, Državnog hidrometeorološkog zavoda. Za izradu modela korišteni su prikupljeni i izračunati srednji dnevni meteorološki i generirani temporalni podaci. U Pythonu su izračunati i uspoređeni kriteriji vrednovanja razvijenih klasifikacijskih modela. Za korištene podatke za procjenu dnevnog indeksa kvalitete zraka modeli nasumičnih šuma pokazuju bolje rezultate i performanse u odnosu na model potpornih vektora. Oba razvijena modela pokazuju zadovoljavajuću točnost (oko 90%) te se u budućnosti mogu primijeniti za predviđanje indeksa kvalitete zraka na novim skupovima podataka. Točno predviđanje indeksa kvalitete zraka može pridonijeti pravovremenom donošenju učinkovitih mjera za poboljšanje kvalitete zraka.

Ključne riječi: Onečišćenje zraka, strojno učenje, lebdeće čestice PM_{10}, lebdeće čestice $PM_{2,5}$, indeks kvalitete zraka, klasifikacija, metoda nasumičnih šuma, metoda potpornih vektora, Python, meteorološki podaci, temporalni podaci, kriterij vrednovanja klasifikacijskih modela

ABSTRACT

Air pollution is a critical issue of today, both for the environment and for human health. Therefore, the aim of this final thesis is to develop models for estimating the daily air quality index (AQI) related to outdoor air pollution by particulate matter PM_{10} and $PM_{2,5}$ (particulate matter with an equivalent aerodynamic diameter of less than 10 and 2.5 μm) using methods of monitored machine learning based on classification. Random Forest method and Support Vector Machine method were applied for the development of the model, using the Python programming language and associated libraries. The theses used data on mass concentrations of fractions of particulate matter PM_{10} and $PM_{2,5}$ and meteorological data from three measuring stations of the state network for permanent monitoring of air quality. In addition to local daily meteorological data, daily data collected at the Maksimir meteorological station of the State Hydrometeorological Institute were also used. Collected and calculated average daily meteorological and generated temporal data were used to create the model. Validation metrics, evaluation criteria of the developed classification models, were calculated and compared in Python. For the data used to estimate the daily Air Quality Index, the Random Forest models show better results and performance compared to the Support Vector Machine models. Both developed models show satisfactory accuracy (around 90%) and can be applied in the future to predict Air Quality Index on new data sets. Accurate prediction of the Air Quality Index can contribute to timely adoption of effective measures to improve air quality.

Keywords: Air pollution, machine learning, particulate matter PM_{10}, particulate matter $PM_{2,5}$, air quality index, classification, random forest, support vector machine, Python, meteorological data, temporal data, classification model evaluation criteria

SADRŽAJ

1. UVOD

Onečišćenje zraka je veliki problem današnjice, osobito u urbanim sredinama koje su karakteristične po velikom broju stanovnika, gustom prometu, visokoj potrebi za energijom te velikom broju industrijskih postrojenja. Uvođenjem i provođenjem raznih regulativa i zakona kroz zadnjih nekoliko desetljeća uspješno su smanjene razine pojedinih onečišćujućih tvari u zraku. Međutim, lebdeće čestice jedna su od vrsta onečišćujućih tvari koje potječu iz brojnih izvora, što otežava njihovo reguliranje i praćenje. Najopasnije su upravo vrlo sitne lebdeće čestice pa se stoga u okviru rutinskog monitoringa najčešće prate frakcije lebdećih čestica aerodinamičnog promjera manjeg od 10 ili 2,5 μm (PM_{10} i $PM_{2,5}$) zbog sposobnosti prodiranja u kožu, bronhije i krvotok. Osim visokog rizika kojeg predstavljaju za ljudsko zdravlje, također imaju štetan utjecaj na okoliš, direktno i indirektno utječu na klimu i smanjuju vidljivost. Posljedično, bitno je pratiti koncentraciju PM_{10} i $PM_{2,5}$ kako bi se točno procijenila kvaliteta zraka, ublažile opasnosti po zdravlje i razumjeli klimatski učinci aerosola. Čimbenici poput gustoće prometa, vremenskih i meteoroloških uvjeta, te blizine i vrste raznih industrija utječu na koncentraciju čestica u zraku. Kako bi se javnosti olakšala interpretacija i razumijevanje izmjerenih vrijednosti uobičajeno se koristi indeks kvalitete zraka (engl. *Air Quality Index*, AQI) kao pokazatelj onečišćenja.[1,2]

Zbog kompleksnosti problema onečišćenja zraka lebdećim česticama i općenito velikog ekonomskog ulaganja u praćenje i regulaciju onečišćujućih tvari kao i u troškove zdravstvenih posljedica uslijed onečišćenja zraka, bitno je pravovremeno i preventivno djelovanje. Stoga sve veću važnost dobivaju modeli koji omogućavaju predviđanje koncentracija onečišćujućih tvari u zraku. Kao sredstvo za predviđanja onečišćenja zraka može se koristiti strojno učenje. Strojno učenje omogućuje razvoj naprednih pristupa koji mogu autonomno izdvajati uzorke i predviđati trendove. Modeli korišteni u ovom radu izabrani su na temelju ranijih istraživanja koja dokazuju njihov potencijal primjene za predviđanje kvalitete zraka.[1–3]

U ovom radu cilj je usporediti i optimizirati modele za procjenu indeksa kvalitete zraka koji se odnosi na onečišćenje zraka lebdećim česticama PM_{10} i $PM_{2,5}$ u zraku na području grada Zagreba pomoću metoda strojnog učenja. U tu svrhu primijenjeno je nekoliko modela nadziranog strojnog učenja temeljenih na klasifikaciji, a to su model potpornih vektora (engl. *Support Vector Machine*, SVM) i model nasumičnih šuma (engl. *Random Forest*, RF). Za potrebe strojnog učenja i izrade modela korišten je programski jezik Python i

njegove pripadajuće knjižice. Nadalje, za treniranje i razvoj modela korišteni su meteorološki mjerni podaci i mjerni podaci o koncentraciji frakcija lebdećih čestica, prikupljeni na tri mjerne postaje državne mreže za trajno praćenje kvalitete zraka u gradu Zagrebu za razdoblje od 2015. do početka 2024. godine. Modeli su vrednovani klasifikacijskim kriterijima vrednovanja te je uspoređena njihova učinkovitost.

2. TEORIJSKI DIO

2.1. Onečišćenje zraka

Onečišćenje zraka nastaje prirodnim ili antropogenim djelovanjem prilikom čega se onečišćujuće tvari ispuštaju u atmosferu i mijenjaju njezine prirodne karakteristike. Onečišćenje zraka štetno djeluje na okoliš, potiče učinak staklenika i klimatske promjene, djeluje na razaranje ozonskog sloja, smanjenje stratosferskog ozona i stvaranje troposferskog te ima negativne posljedice na vegetaciju i ekosustav.[4,5] Također, smanjuje i kvalitetu života, razara materijalna i kulturna dobra te dovodi do smanjene vidljivosti. Nadalje, štetno djeluje na ljudsko zdravlje, uzrokuje čitav niz bolesti od kardiovaskularnih do raka pluća pa čak i negativno djeluje na psihičko zdravlje. Prema podacima Svjetske zdravstvene organizacije (engl. *World Health Organization*, WHO) čak 99 % svjetske populacije izloženo je zraku koji sadrži onečišćujuće tvari u koncentracijama iznad smjernica WHO, pri čemu su najizloženije zemlje u razvoju (u području Afrike, Azije i Južne Amerike), a od 1990. godine broj globalnih smrti uzrokovanih onečišćenjem zraka nije se promijenio.

Najopćenitija podjela onečišćenja zraka dijeli onečišćenje na ono u vanjskom zraku (engl. *outdoor ili ambient air pollution*) i ono u zatvorenim prostorima (engl. *indoor air pollution*). U ovom radu obrađuju se podaci o onečišćenju vanjskog zraka. S obzirom na izvor, onečišćenje zraka može biti posljedica emisija iz nepokretnih (stacionarnih) izvora poput industrijskih procesa unutar pogona, građevinske industrije, poljoprivrede, itd. ili emisija iz pokretnih (mobilnih) izvora, prijevoznih sredstava. S obzirom na nastanak, onečišćujuće tvari se dijele na:

- primarne (nastaju direktnom emisijom iz izvora): primarne lebdeće čestice, spojevi sumpora, dušikovi oksidi (dušikov oksid, NO, dušikov dioksid, NO_2, didušikov oksid, N_2O), ostali dušikovi spojevi (amonijak, NH_3 i cijanovodik, HCN), ugljikovi spojevi (ugljikov monoksid, CO i ugljikov dioksid, CO_2), hlapljivi organski spojevi (engl. *Volatile organic compounds*, VOC), metali i njihovi halogenidi
- sekundarne (nastaju reakcijama primarnih onečišćujućih tvari i okoliša, npr. djelovanjem sunčeve energije, oksidacijom ili međusobnim reakcijama): ozon, poliakrilonitril (PAN), sekundarne lebdeće čestice, NO_2, VOC, CO_2

Posljedice onečišćenja očituju se na lokalnoj, regionalnoj i globalnoj razini. Razarajuće posljedice iziskuju regulaciju emisija te kontinuirano praćenje njihovih koncentracija u zraku, posljedično potrebna su visoka ekonomska ulaganja za sprječavanje

ovog gorućeg problema današnjice. Pregledom i revizijom znanstvenih dokaza WHO donosi smjernice za kvalitetu zraka (engl. *air quality guidelines*, AQG) kao prijedlog i podlogu za granične vrijednosti u zakonodavstvu. U tablici 1 prikazane su godišnje i dnevne granice za neke onečišćujuće tvari ($PM_{2,5}$, PM_{10}, prizemni ozon (O_3), NO_2, sumporov dioksid (SO_2) i CO).[4,6]

Tablica 1. Predložene granične vrijednosti i privremeni ciljevi za razne onečišćujuće tvari prema WHO.[4]

Onečišćujuća tvar	Period	AQG razina
$PM_{2,5}$, $\mu g/m^3$	Godišnje	5
	24 h [a]	15
PM_{10}, $\mu g/m^3$	Godišnje	15
	24 h [a]	45
O_3, $\mu g/m^3$	vrhunac sezone [a]	60
	8 h [a, b]	100
NO_2, $\mu g/m^3$	Godišnje	10
	24 h [a]	25
SO_2, $\mu g/m^3$	24 h [a]	40
CO, mg/m^3	24 h	4

a) 99. percentil (tj. 3-4 dana prekoračenja godišnje),

b) Prosjek dnevne maksimalne 8-satne srednje koncentracije O_3 u šest uzastopnih mjeseci s najvišom šestomjesečnom prosječnom koncentracijom O_3.

2.1.1. Izvori i širenje onečišćenja u gradovima

Antropogeni izvori onečišćenja razlikuju se od prirodnih izvora po velikom masenom toku emisija s obzirom na malu površinu emisije. Globalizacija i urbanizacija su dva čimbenika koja itekako doprinose onečišćenju zraka.[6] Onečišćenje zraka potaknuto urbanizacijom je multidimenzionalni problem, a istraživački pristupi mogu se podijeliti u tri skupine s obzirom na način promatranja ovoga izazova. Prva skupina istraživanja polazi od hipoteze da postoji linearna ovisnost između rasta stanovništva i onečišćenja okoliša, uzrokovanog selidbom stanovništva iz ruralnih područja u urbane sredine što uzrokuje trošenje velike količine fosilnih goriva. Druga skupina polazi od tvrdnje da se urbanizacijom uspostavlja kompaktnija mreža transporta, učinkovitije se koristi prostor te dugoročno razvijenije i ekonomski uspješnije sredine nastoje više ulagati u očuvanje okoliša. Treća skupina odnosi se na teoriju nelinearnog, obrnutog odnosa između gospodarskog rasta i onečišćenja okoliša (U-krivulja).[7]

U urbanim sredinama antropogeni izvori uključuju promet (automobili, kamioni, zrakoplovi, brodski motori, itd.), industriju i proizvodnju energije.[8] U razvijenim zemljama industrijske emisije značajno su se smanjile, tako da je na nekim lokacijama promet glavni izvor onečišćenja zraka. Kada se količina prometa poveća i što se vozila sporije kreću, kvaliteta zraka se pogoršava.[9] Značajan izvor onečišćenja zraka mogu biti i kućna ložišta tijekom sezone grijanja.[10,11]

Onečišćenje zraka ne određuje samo vrsta i intenzitet emisija. Meteorologija i klima, kao i lokalna topografija imaju velik utjecaj na raspršivanje i pretvorbu onečišćujućih tvari u atmosferi. Koncentracija onečišćujućih tvari u niskim slojevima atmosfere ovisi, između ostalog, o atmosferskom tlaku, vjetru i temperaturi. Ciklone (niski atmosferski tlakovi) povezane su s jakom turbulencijom zraka i stoga s dobrim uvjetima za raspršivanje, dok anticiklone (visoki tlakovi) odgovaraju zračnoj stabilnosti, slabom strujanju zraka i uzrokuju epizode onečišćenja u urbanim sredinama. Raspršivanje tvari povećava se s brzinom i turbulencijom vjetra, a prati se i njegovo usmjerenje. Vertikalni temperaturni gradijent (opisuje koliko i u kojem smjeru oko određenog mjesta se mijenja temperatura po udaljenosti, mjerna jedinica K/m) pomaže uzdizanju onečišćujućih tvari u zrak te je znak nestabilnosti atmosfere. Međutim, u slučaju temperaturne inverzije (porast temperature s visinom)[12], onečišćujuće tvari su zadržane u niskim slojevima atmosfere, što stvara epizode onečišćenja. Na kemijsku transformaciju onečišćujućih tvari u zraku najviše utječu temperatura, vlaga i sunčeva svjetlost. Sunčeve zrake i visoka temperatura ljeti potiču fotokemijske reakcije i stvaranje ozona te nastanak fotokemijskog smoga.[8]

2.1.2. Lebdeće čestice

Jedan od velikih problema u urbanim i ruralnim sredinama predstavlja onečišćenje zraka lebdećim česticama (engl. *particulate matter*, PM). Definiraju se kao heterogene čvrste ili tekuće čestice raspršene u zraku, zato se nazivaju i atmosferskim aerosolom. S obzirom na veličinu najčešće se određuju:

a) Ukupne lebdeće čestice

b) PM_{10}, $PM_{2,5}$, PM_1 –čestice aerodinamičkog promjera manjeg od 10 μm, 2,5 μm i 1 μm

c) Ultrafine čestice (čestice promjera manjeg od 100 nm)

Izvori PM dijele se na prirodne (šumski požari, vulkanske erupcije, pustinjske oluje…) i antropogene (industrijske aktivnosti, izgaranje fosilnih goriva, spaljivanje

biomase…). U okolišu nastaju izravno iz primarnih izvora (primarni aerosoli) te neizravno kemijskim reakcijama i faznim promjenama u atmosferi (sekundarni aerosoli).[6]

Ovisno o veličini, zadržavaju se u atmosferi od par sati do nekoliko tjedana, nakon čega se procesima mokrog ili suhog taloženja talože na površinu Zemlje. PM utječu na klimu te se procjenjuje da antropogene promjene u aerosolima doprinose promjeni od 40 % u kratkovalnom zračenju i povećanju od 60 % u broju kondenzacijskih jezgri oblaka (engl. *Cloud condensation nuclei,* CCN), malih čestica na kojima se kondenzira vodena para.[13] Nadalje, utječu na kruženje hranjivih tvari, mijenjaju pH tla i dovode do smanjene fotosintetske aktivnosti kod biljaka.[14,15]

Čimbenici poput trajanja izloženosti, njihov kemijski sastav i fizikalna svojstva te masa i veličina uvelike utječu na posljedice koje imaju na ljudsko zdravlje i ekosustav. Fine čestice ($PM_{2,5}$) imaju sposobnost dubljeg prodiranja u organizam, apsorbiraju se u krvotok, pri čemu su opasnije za zdravlje ljudi. Epidemiološka istraživanja povezala su izloženost lebdećim česticama s posljedicama poput raka, promjenama u ekspresiji gena, kardiovaskularnih bolesti, prisutnosti toksičnog materijala u krvi (olovo, kadmij, cink), alergijskih reakcija, bakterijskih i gljivičnih infekcija, fibroze (npr. azbest, kvarc), iritacije sluznice (kiseline i lužine) te pojačanim respiratornim simptomima, pogoršanjem astme i preranom smrću. Rizici su najveći za osjetljive skupine kao što su starije osobe, kronični bolesnici i djeca.[13,16]

S obzirom da se ovisno o porijeklu, razlikuju po masi i kemijskom sastavu, vrlo je teško pratiti utjecaje koje imaju na okoliš i ljudsko zdravlje.

2.1.3. Indeks kvalitete zraka

Indeks kvalitete zraka (engl. *Air Quality Index,* AQI) koristi se kao pokazatelj onečišćenosti zraka kako bi se olakšala interpretacija mjernih rezultata. AQI se računa na temelju podataka o koncentracijama onečišćujućih tvari poput lebdećih čestica PM_{10} i $PM_{2,5}$, prizemnog ozona, dušikova dioksida i sumporova dioksida. Mjerna vrijednost koncentracije onečišćujuće tvari svrstava se u određeni raspon koncentracija koja definira AQI kategoriju. Na prometnim lokacijama koncentracije sumporovog dioksida su obično visoke samo u vrlo lokaliziranim područjima, dok su koncentracije prizemnog ozona često vrlo niske. Stoga je za prometne lokacije optimalno izračunavati indeks koristeći podatke o koncentracijama dušikovog dioksida i lebdećih čestica ($PM_{2,5}$, PM_{10} ili obje). Za industrijske i pozadinske mjerne postaje izračunavanje indeksa zahtijeva podatke o najmanje tri onečišćujuće tvari: dušikov dioksid, prizemni ozon i lebdeće čestice ($PM_{2,5,}$ PM_{10} ili obje). Vrijeme usrednjavanja

za izračun indeksa razlikuje se ovisno o tipu onečišćujuće tvari. Za NO_2, O_3 i SO_2, koriste se satne koncentracije. S druge strane, za lebdeće čestice PM_{10} i $PM_{2,5}$ koriste se pomični 24-satni prosjeci kako bi se osigurala točnost i reprezentativnost podataka. U tablici 2. prikazana je kategorizacija kvalitete zraka na temelju izmjerenih koncentracija onečišćujućih tvari i raspona u koji spadaju.[17]

Tablica 2. Oznake AQI kategorije onečišćujućih tvari prema koncentraciji istih.[17]

Onečišćujuća tvar	**Dobro**	**Prihvatljivo**	**Umjereno**	**Loše**	**Vrlo loše**	**Izuzetno loše**
$PM_{2,5}$	0-10	10-20	20-25	25-50	50-75	75-800
PM_{10}	0-20	20-40	40-50	50-100	100-150	150-1200
NO_2	0-40	40-90	90-120	120-230	230-340	340-10000
O_3	0-50	50-100	100-130	130-240	240-380	380-800
SO_2	0-100	100-200	200-350	350-500	500-750	750-1250

Oznake AQI dopunjene su sa smjernicama za javnost kako bi se olakšalo razumijevanje i prevenirale posljedice na zdravlje. Te smjernice prikazane su u tablici 3. za opću populaciju i osjetljive skupine.

Tablica 3. Smjernice ponašanja za javnost s obzirom na razinu indeksa.[17]

Razina indeksa	**Opća populacija**	**Osjetljive skupine građana**
Dobro	Kvaliteta zraka je dobra. Uživajte u svojim svakodnevnim aktivnostima na otvorenom.	Kvaliteta zraka je dobra. Uživajte u svojim svakodnevnim aktivnostima na otvorenom.
Prihvatljivo	Uživajte u svojim svakodnevnim aktivnostima na otvorenom.	Uživajte u svojim svakodnevnim aktivnostima na otvorenom.
Umjereno	Uživajte u svojim svakodnevnim aktivnostima na otvorenom.	Razmislite o smanjenju intenzivnih aktivnosti na otvorenom, ukoliko osjetite simptome.
Loše	Razmislite o smanjenju intenzivnih aktivnosti na otvorenom ukoliko osjetite simptome poput nadražaja očiju, kašlja ili grlobolje.	Razmislite o smanjenju tjelesnih aktivnosti, osobito na otvorenom, posebno ukoliko osjetite simptome.
Vrlo loše	Razmislite o smanjenju intenzivnih aktivnosti na otvorenom ako osjetite simptome poput nadražaja očiju, kašlja ili grlobolje.	Smanjite fizičke aktivnosti, osobito na otvorenom, posebno ukoliko osjetite simptome.
Izuzetno loše	Smanjite fizičke aktivnosti na otvorenom.	Izbjegavajte fizičke aktivnosti na otvorenom.

2.2. Utjecaj meteoroloških uvjeta na kvalitetu zraka

Teško je opisati utjecaj pojedinih čimbenika na onečišćenje zraka jer kombinacije različitih uvjeta i njihovih vrijednosti različito utječu na kvalitetu zraka. U sljedećim poglavljima opisana je opća povezanost meteoroloških uvjeta i onečišćenja zraka u gradovima prema postojećim istraživanjima.

2.2.1. Vjetar

Vjetar je pojava strujanja zraka, najčešće vodoravnog, koja nastaje uslijed razlike tlakova između dvaju područja. Zrak struji iz područja višeg tlaka prema području nižeg tlaka. Što je ta razlika veća, to je vjetar jači (brži). Osim jačine vjetra, određuje se i smjer vjetra (ruža vjetrova). Na njegovo kretanje utječu Zemljina rotacija, Coriolisova sila, centrifugalna sila, sile trenja s podlogom te reljef i temperaturne razlike između mora i kopna.[18]

Brzina vjetra mjeri se anemometrima ili se procjenjuje prema Beaufortovoj ljestvici. Beaufortova ljestvica, koja se koristi od 1874. godine, međunarodno je prihvaćena iskustvena ljestvica za procjenjivanje jakosti vjetra prema učincima na kopnu ili prema stanju morske površine, u 13 stupnjeva (od 0 do 12 bofora).[19] Značenja i rasponi brzine vjetra za svaki stupanj opisani su u tablici 4 koja je preuzeta s mrežne stranice.

Tablica 4. Beaufortova ljestvica jakosti vjetra.[19]

Stupnjevi	**Opis vjetra**	**Glavni učinci vjetra na kopnu**	**Brzina (km/h)**	**Brzina (m/s)**
0	tišina	dim se diže okomito uvis	0–1	0–0,2
1	lahor	smjer vjetra zapaža se po dimu	1–5	0,3–1,5
2	povjetarac	vjetar se osjeća na licu, vjetrulja se pokreće	6–11	1,6–3,3
3	slab vjetar	lišće i grančice stalno se njišu	12–19	3,4–5,4
4	umjeren vjetar	vjetar podiže prašinu i pokreće manje grane	20–28	5,5–7,9
5	umjereno jak vjetar	tanja lisnata stabla počinju se njihati	29–38	8,0–10,7
6	jak vjetar	pokreću se velike grane, čuje se zujanje telefonskih žica	39–49	10,8–13,8
7	žestok vjetar	njišu se cijela stabla, hodanje otežano	50–61	13,9–17,1
8	olujni vjetar	vjetar lomi grane na drveću	62–74	17,2–20,7
9	jak olujni vjetar	nastaju laka oštećenja na zgradama	75–88	20,8–24,4
10	orkanski vjetar	velike štete na zgradama, čupa drveće iz zemlje	89–102	24,5–28,4
11	jak orkanski vjetar	velika razaranja	103–117	28,5–32,6
12	orkan	katastrofalna razaranja	>118	32,7–36,9

Strujanje zraka uzrokuje raspršenje onečišćujućih tvari od njihovog izvora, raspodjela onečišćenja može se pratiti na globalnoj i lokalnoj razini. Veće brzine obično uzrokuju veću disperziju, što rezultira smanjenjem koncentracije onečišćujućih tvari.[20] Kako se tlo zagrijava tijekom dana, zrak općenito postaje turbulentniji, uzrokujući raspršivanje onečišćujućih tvari u zraku. Kada je noću zrak hladniji, nastaju stabilniji uvjeti, zbog čega se onečišćujuće tvari manje raspršuju (detaljno opisano u poglavlju 2.4.2.).[21,22] U radu Cuhadaroglua prosječna koncentracija lebdećih čestica blago se smanjuje s povećanjem brzine vjetra do vrijednosti od 8 m/s.[23]

2.2.2. Temperatura zraka

Temperatura zraka u meteorologiji je temperatura u prizemnom sloju atmosfere koja se mijenja tijekom dana s obzirom na doba dana, vremenske uvjete (vjetar, oborine) i prati sezonski trend ovisno o godišnjem dobu (koje se mijenja s promjenom položaja Zemlje prema Suncu, ovisi o geografskom položaju lokacije i klimatskim promjenama). Temperatura zraka mjeri se na 2 m iznad tla u termometrijskoj kućici.[24]

Temperatura u urbanim područjima je u prosjeku 5 °C viša nego u ruralnim područjima.[25] Fenomen akumulacije topline u urbanim područjima naziva se urbanski toplinski otok (engl. *urban heat island*, UHI). UHI nastaje zbog ljudskih aktivnosti (vozila, industrijske aktivnosti) i karakterističnog načina gradnje urbanih sredina (asfalt, beton, manjak zelenih površina, gustoća izgradnje), a povećanje temperature površine tla uzrokovano UHI-om utječe na energetske tokove u gradovima.[26]

Prijašnja istraživanja prikazala su jasnu negativnu korelaciju između koncentracija lebdećih čestica i temperature, što je više povezano s ljudskom aktivnosti, naročito grijanjem u hladnijem dijelu godine.[10,11,23,27] U istraživanjima[23,28] ispitivan je utjecaj temperature na koncentraciju $PM_{2,5}$, a rezultati su pokazali da su koncentracije $PM_{2,5}$ obrnuto proporcionalne visini mješovitog sloja (engl. *mixing layer height*, MLH). MLH označava visinu atmosferskog sloja u kojem dolazi do miješanja zraka zbog turbulencije, nastalog zbog razlika temperature zraka i tla. Zimi tijekom dana, zagrijavanje tla od strane Sunca povećava MLH, što omogućava bolje razrjeđivanje onečišćujućih tvari. Noću, kada se tlo hladi i tijekom stabilnih atmosferskih uvjeta, MLH se smanjuje, što dovodi do akumulacije i taloženja PM blizu površine.[28]

U istraživanju Vaishali i Das (2023.) uočena je jaka negativna korelacija koncentracije PM i temperature tijekom uvjeta visoke vlažnosti (iznad 50%) i slaba korelacija s temperaturom okoline tijekom uvjeta niske vlažnosti (ispod 50%).[29]

2.2.3. Relativna vlažnost zraka

Relativna vlažnost zraka (engl. *relative humidity*, RH) je fizikalna veličina koja izražava količinu vodene pare prisutne u zraku ili plinovima. Izračunava se kao omjer parcijalnog tlaka vodene pare prisutne u zraku i parcijalnog tlaka zasićene vodene pare pri istoj temperaturi i tlaku zraka. Također se može prikazati kao omjer između trenutne apsolutne vlažnosti i maksimalne moguće apsolutne vlažnosti pod istim uvjetima. Relativna vlažnost se često izražava kao postotak (%). Kada je zrak potpuno suh, relativna vlažnost je 0 %, dok je kod potpuno zasićenog zraka 100 %. Obično se mjeri pomoću psihrometra ili specijaliziranih senzora.[30]

Uvjeti koji su izvan optimalnog raspona od 40-60 % mogu imati značajan utjecaj na ljudsko zdravlje, uključujući olakšavanje prijenosa zaraze i pogoršanje respiratornih bolesti. Kada je RH preniska, može uzrokovati suhoću i iritaciju dišnog trakta i kože, čineći pojedince osjetljivijima na infekcije. S druge strane, kada je RH previsoka, može stvoriti vlažno okruženje koje potiče rast štetnih mikroorganizama poput plijesni, bakterija i virusa.[27] Visoka RH uzrokuje smanjenje cirkulacije zraka što također dovodi do povećanja koncentracije onečišćujućih tvari.[31]

Utjecaj meteoroloških uvjeta na koncentraciju PM u zraku nije jednostavno objasniti za svaki čimbenik jer određene kombinacije čimbenika i njihovih vrijednosti drugačije utječu na kvalitetu zraka. U poglavlju 2.4.2 navedeno je da je u radu[29] uočena je jaka negativna korelacija koncentracije PM i temperature tijekom uvjeta visoke vlažnosti (iznad 50%) i slaba korelacija s temperaturom okoline tijekom uvjeta niske vlažnosti (ispod 50%). Istraživanje[20] pokazalo je da RH te koncentracija PM i ostalih onečišćujućih tvari imaju obrnuti odnos. Tijekom ljetnih mjeseci, kada je RH niža, koncentracija onečišćujućih tvari je visoka, a tijekom zimskih mjeseci kapljice vode u zraku pomažu u taloženju onečišćujućih tvari većeg promjera (PM_{10}), dok koncentracija manjih čestica ovisi o MLH, koji zimi uzrokuje skupljanje čestica u nižim slojevima atmosfere.[28,32]

2.2.4. Tlak zraka

Tlak zraka je skalarna fizikalna veličina, a opisuje se djelovanjem sile na površinu. Mjerna jedinica je Pa (Pascal), a koriste se i bar (1 bar = 105 Pa), normalna atmosfera (1 atm = 101 325 Pa), milimetar stupca žive ili tor (1 mmHg = 1 tor = 133,322 Pa) te milimetar stupca vode (1 mmH2O = 9,806 649 Pa). Atmosferski tlak mjeri se u meteorologiji. On je uzrokovan težinom zraka i određen težinom stupca zraka nad površinom. Standardni

atmosferski tlak (p_0) je tlak zraka mjeren na razini mora, tj. na nadmorskoj visini 0 m, pri temperaturi zraka 0 °C i iznosi 101 325 Pa.[33]

Barometar, manometar i vakuummetar su instrumenti za mjerenje tlaka. Mjeri se kao apsolutni ili kao relativni s obzirom koji se tlak uzima kao nulta vrijednost (vakuum ili atmosferski).[33]

Ne postoje direktni učinci tlaka zraka na onečišćenje. Tlak zraka, kao što je opisano u poglavlju 2.1.1. uzrokuje uvjete ciklone i anticiklone pa na taj način i uvjetuje kvalitetu zraka. Dakle, niski tlak dovodi do nestabilnih vremenskih uvjeta, a visoki tlak dovodi do stabilnih vremenskih uvjeta. U uvjetima visokog tlaka (anticiklona) vjetrovi su slabi i pušu u smjeru kazaljke na satu (na sjevernoj hemisferi). Zrak se spušta što smanjuje stvaranje oblaka i dovodi do slabog vjetra i ustaljenih vremenskih prilika.[34] U uvjetima niskog tlaka (ciklona), zrak se diže i puše u smjeru suprotnom od kazaljke na satu (na sjevernoj hemisferi). Dok se diže i hladi, vodena para se kondenzira stvarajući oblake i moguće su padaline.[35]

2.2.5. Oborine

Oborina ili padalina je voda u tekućem ili čvrstom agregatnom stanju koja pada na tlo iz oblaka ili nastaje na tlu kondenzacijom vodene pare iz zraka u dodiru s tlom. Dijele se na konvekcijske (pljuskovi iz kumulonimbusa), orogene (utjecaj orografije, nastaju prisilnim dizanjem vlažna zraka uz obronke planina pod utjecajem vjetra) i frontalne (utjecaj ciklone). Oborine su vremenski i prostorno promjenjive, a određene su i klimom područja. Razlikuju se ekvatorski (s maksimumom oborina nakon proljetne i jesenske ravnodnevice), tropski (maksimum oborina ljeti), monsunski (maksimum oborina ljeti, zime suhe), suptropski (maksimum oborina zimi, ljeta suha), kontinentalni (ljetne kiše), oceanski (zimske kiše) tip oborina te sredozemni tip oborina (zime kišovite, ljeta suha).[36]

Količina oborina mjeri se kišomjerom (njime se utvrđuje koliko bi milimetara bio visok sloj vode od oborina kada ne bi bilo isparavanja, otjecanja i prokapljivanja kroz tlo). Količina oborina od 1 mm odnosi se na površinu od 1 m^2, što znači da je na svaki kvadratni metar tla pala jedna litra vode.[36]

Oborine u tekućem stanju imaju dvostruki učinak na koncentraciju onečišćujućih tvari u zraku. Kiša ispire onečišćujuće tvari i može uzrokovati otapanje i taloženje istih na tlo.[9,37] Indirektni učinak oborina povećava onečišćenje zbog većeg prometa i smanjenja brzine vozila. Ukupni učinak kiše ovisi o veličini ovih suprotstavljenih učinaka.[9]

2.3. Praćenje kvalitete zraka u gradu Zagrebu

Zagreb je najveći (641,24 km^2) i glavni grad Republike Hrvatske sa 767 131 stanovnika, prema popisu stanovništva provedenom 2021. godine. Nalazi se na 122 m nadmorske visine. Položaj grada u odnosu na opservatorij Grič je 15°59' istočne geografske dužine i 45°49' sjeverne geografske širine.[38] Kroz grad prolazi rijeka Sava (29 km), a grad se nalazi u podnožju gore Medvednice na jugozapadnom rubu Panonske nizine.[39]

Ukupan broj registriranih motornih vozila prema podacima iz 2022. godine iznosi 442 049, što je 3,4 % više u odnosu na 2021. godinu. Od toga broj osobnih vozila iznosi 307 518. 2022. godine prosječan broj osobnih motornih vozila u Europi na 1000 stanovnika je iznosio 560, a u Zagrebu 400.[40] U blizini Zagreba nalazi se i Zračna luka Franjo Tuđman, u 2022. ukupan promet putnika iznosio je 3,1 milijuna, odnosno 22,5 % više nego 2021.[38,41]

Vodeće industrijske grane u Zagrebu su građevinska i prerađivačka industrija, koja se odnosi na proizvodnju prehrambenih proizvoda, pića, tekstila, papira, gume i plastike, metala, električne opreme, motornih vozila, namještaja, farmaceutskih proizvoda…[38]

Područje Zagreba karakterizira kontinentalna klima. Državni hidrometeorološki zavod (DHMZ, državna upravna i znanstvena organizacija) svakodnevno mjeri i prati hidrološke i meteorološke podatke (temperaturu i vlažnost zraka, tlak, količinu oborina, sijanje sunca, jačinu vjetra, smjer vjetra, učestalost kiselih kiša) sinoptičkim, klimatološkim i kišomjernim postajama te na automatskim postajama. Osim toga, širok spektar poslova DHMZ-a uključuje prikupljanje, obradu i objavljivanje meteoroloških i hidroloških podataka te istraživanje atmosfere i vodnih resursa te primjenama meteorologije i hidrologije u područjima klimatologije, pomorske meteorologije, agrometeorologije, zrakoplovne meteorologije, prostornog planiranja i projektiranja te ostalih područja.[42]

Prema podacima DHMZ-a najviša srednja mjesečna temperatura zraka u Zagrebu iznosila je 24,0 °C, u srpnju 2022., a povijesno najviša temperatura iznosila je 40,4 °C u srpnju 1950. godine. Najniža srednja mjesečna temperatura iznosila je -8,6 °C u siječnju 2022., a najniža prosječna mjesečna temperatura zabilježena je 1956. godine i iznosila je -27,3 °C.[41] Ukupna količina oborina 2023. iznosila je 1220,2 mm/m^2.[42]

Kranjčić et al.[43] navode kako je Zagreb relativno naseljen šumama i zelenim urbanim površinama (3 % površine grada). No, glavni problem je u tome što postojeće biljke oko prometnica nemaju sposobnost apsorbiranja onečišćujućih tvari. Ovo istraživanje ne samo da predstavlja novi pristup kombinirajući tehnologiju satelitskih podataka i strojnog učenja za

definiranje urbanih zelenih površina, već također nudi konkretne preporuke za poboljšanje kvalitete života u Zagrebu kroz povećanje broja stabala i smanjenje onečišćenja zraka.

Prva mjerenja kvalitete zraka u Zagrebu proveli su 60-tih godina prošlog stoljeća djelatnici Instituta za medicinska istraživanja i medicinu rada, a sadržavala su mjerenja sumporova dioksida i crnog dima, koji su u to vrijeme u Europi predstavljali značajan ekološki i zdravstveni problem. Bila su to ujedno i prva mjerenja kvalitete zraka u Hrvatskoj te su dovela do formiranja mjerne mreže grada Zagreba sredinom šezdesetih godina prošlog stoljeća. Po uzoru na Zagreb i neki drugi gradovi u Hrvatskoj oformili su svoje mjerne mreže za praćenje kvalitete zraka, a s godinama rastao je broj mjernih postaja kao i broj tvari koji se pratio.[44–46]

Državna mreža za trajno praćenje kvalitete zraka u Republici Hrvatskoj uspostavljena je početkom ovog stoljeća te je sastavni dio praćenja stanja i okoliša, u nadležnosti Ministarstva zaštite okoliša i zelene tranzicije. Portal „Kvaliteta zraka u Republici Hrvatskoj“ [17] sadrži izmjerene koncentracije onečišćujućih tvari u zraku iz državne mreže za trajno praćenje kvalitete zraka te iz lokalnih mreža (u nadležnosti županija, Grada Zagreba, gradova i općina). Prema članku 31. Zakona o zaštiti zraka (NN 127/19, NN 57/22) Državni hidrometeorološki zavod pod stručnim nadzorom Ministarstva zaštite okoliša i zelene tranzicije, upravlja radom državne mreže, osigurava izgradnju novih postaja u državnoj mreži i praćenje kvalitete zraka i odgovoran je za provođenje programa mjerenja kvalitete zraka na tim mjernim postajama. Praćenje kvalitete zraka u postajama iz državne mreže za plinovite onečišćujuće tvari (SO_2, NO i NO_2, benzen, CO, prizemni ozon i prekursore ozona) i lebdeće čestice PM_{10} i $PM_{2,5}$ (automatske metode) obavlja DHMZ.[42] Praćenje kvalitete zraka u postajama iz državne mreže u dijelu koji se odnosi na uzorkovanje i fizikalno kemijske analize lebdećih čestica PM_{10} i $PM_{2,5}$ te ekvivalenciju nereferentnih metoda za određivanje masenih koncentracija lebdećih čestica PM_{10} i $PM_{2,5}$ obavlja Institut za medicinska istraživanja i medicinu rada (IMI).[47] Podaci kvalitete zraka iz državne mreže javni su i objavljuju se na mrežnim stranicama Ministarstva.

Ocjena kvalitete zraka provodi se temeljem Zakona o zaštiti zraka (NN 127/19, NN 57/22) pri čemu se zrak dijeli u dvije kategorije:

- I kategorija - čist ili neznatno onečišćeni zrak: nisu prekoračene granične vrijednosti, ciljne vrijednosti i ciljne vrijednosti za prizemni ozon;
- II kategorija - onečišćen zrak: prekoračene su granične vrijednosti, ciljne vrijednosti i ciljne vrijednosti za prizemni ozon.

Granične i ciljne vrijednosti za pojedine onečišćujuće tvari propisane su Uredbom o razinama onečišćujućih tvari u zraku (NN 77/20).

Zrak u Zagrebu je 2019. godine okarakteriziran kao zrak II. kategorije za NO_2, prizemni ozon i PM_{10}, odnosno onečišćen zrak s obzirom na zaštitu zdravlja ljudi na više mjernih postajama te za H_2S s obzirom na kvalitetu življenja na mjernoj postaji Jakuševac.[48] Problem onečišćenja zraka s PM karakterističan je za zimske mjesece, a problem onečišćenja prizemnim ozonom za ljetne mjesece. Onečišćenje zraka lebdećim česticama u Zagrebu je u prvom redu povezano s izgaranjem nekvalitetnih goriva, drva i ugljena u kućnim ložištima, a tijekom sezone grijanja općenito su povećane emisije plinova i čestica, uz to tijekom zime prisutni su nepovoljni meteorološki uvjeti.[10,49,50] Dugoročni trendovi pokazuju da se razine lebdećih čestica smanjuju, što je rezultat provođenja mjera za zaštitu zraka u okviru lokalnih i globalnih akcijskih planova i propisa.[51]

2.4. Strojno učenje

Umjetna inteligencija je znanstvena disciplina koja se bavi inženjeringom inteligentnih računala, odnosno računalnih sustava reprodukcijom ljudske inteligencije kroz učenje, zaključivanje i samoispravljanje/prilagodbu.[52] Strojno učenje (engl. *Machine Learning*, ML) je grana umjetne inteligencije, a podrazumijeva razvoj matematičkih algoritama i analiziranje složenih podataka imitacijom ljudske inteligencije.[53] Oslanja se na različita područja kao što su neuroznanost, vjerojatnost i statistika, informatika, teorija informacija, psihologija, teorija kontrole i filozofija. Integrirajući ideje iz ovih disciplina, ML-u je cilj humanizirati računala, na način da uče iz svog okruženja i prošlih iskustava, sa ili bez izravnih uputa.[54]

ML počelo se razvijati prvom polovicom 20. stoljeća. U radu *"A logical calculus of the ideas immanent in nervous activity"* Waltera Pittsa i Warrena McCulloha, prvi puta se koristi matematički model neuronskih mreža.[55]

Posljednjih par desetljeća, zbog značajnog rasta i dostupnosti podataka (engl. *Big data*), razvijaju se i primjenjuju tehnike strojnog učenja kako bi se omogućilo razumijevanje problema iz raznih područja znanosti (računalni vid, robotika, ekologija, biologija, medicina, ekonomija, inženjerstvo).[54]

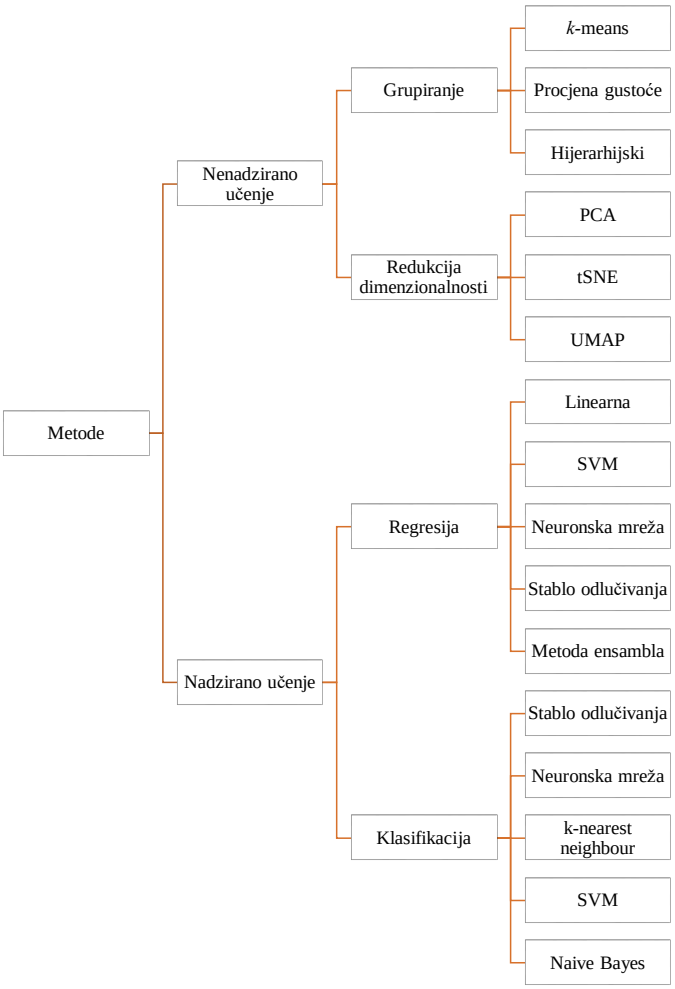
Dakle, ML omogućuje računalima da se prilagode i poboljšaju svoje performanse tijekom vremena učenjem iz podataka, čime igraju ključnu ulogu u pokretanju inovacija i napretka u brojnim domenama. Prema definiciji Alpaydina (2009.)[56] strojno učenje jest programiranje računala na način da optimiziraju neki kriterij uspješnosti temeljem

podatkovnih primjera ili prethodnog iskustva. Učenjem se smatra optimizacija parametara modela temeljem podataka. Modeli ML dijele se na 3 kategorije prema načinu učenja.[57]

1. nadzirano učenje,
2. nenadzirano učenje,
3. polu-nadzirano učenje
4. ojačano učenje.

Nadzirano učenje temelji se na klasifikaciji (izlaz je nebrojčana, diskretna vrijednost) ili regresiji (izlaz je kontinuirana, brojčana vrijednost), modeli se uvježbavaju pomoću dostupnih poznatih ulaza i izlaza i pronalaze vezu između njih procesom optimizacije te se kasnije primjenjuju za predviđanje izlaza samo na temelju novih ulaznih podataka.

Nenadzirano učenje temelji se na grupaciji ili smanjenju dimenzionalnosti, a podrazumijeva analizu i grupiranje neoznačenih skupova podataka, odnosno pronalazi skrivene uzorke i izvodi zaključke samo na temelju ulaznih podataka. Razlikuju se 4 vrste grupiranja, odnosno klasteriranja podataka: ekskluzivno klasteriranje (engl. *Exclusive Clustering*), preklapajuće klasteriranje (engl. *Overlapping Clustering*), hijerarhijsko klasteriranje (engl. *Hierarchical Clustering*) i vjerojatnosno klasteriranje (engl. *Probabilistic Clustering*).[58] Na slici 1. shematski je prikazana podjela metoda te algoritmi koji im pripadaju.[59]



Slika 1. *Shematski prikaz podjele vrsta modela i algoritama strojnog učenja*

Polu-nadzirano učenje koristi se u situacijama kada postoji mala količina označenih podataka s velikom količinom neoznačenih podataka. Polu-nadzirano učenje ima veliku praktičnu vrijednost jer može smanjiti troškove označavanja podataka potrebnih za treniranje modela posebno u situacijama kada je neizvedivo označiti sve značajke. U mnogim situacijama, ručno označavanje svih podataka može biti vrlo skupo i vremenski zahtjevno jer zahtijeva ljudski rad i ekspertizu. Na primjer, u medicinskoj dijagnostici, potrebno je stručno znanje za točno označavanje podataka, što može biti neizvedivo za velike skupove podataka. Polu-nadzirano učenje omogućava korištenje manjeg broja označenih podataka u kombinaciji s velikom količinom neoznačenih podataka, čime se smanjuje potreba za sveobuhvatnim ručnim označavanjem, ali se i dalje postižu točni i korisni rezultati. Model se može trenirati pomoću samo nekoliko ulazno/izlaznih parova.

Ojačano učenje za razliku od prva tri uči putem interakcije sa svojim okruženjem i sakupljanjem iskustva iz tog procesa, istražuje kako inteligentni agenti trebaju djelovati u svom okruženju kako bi postigli najbolje moguće rezultate.[58]

Postupak modeliranja modela strojnog učenja može se podijeliti u 6 koraka, a to su prikupljanje podataka, predobrada podataka, odabir strukture modela, razvoj modela, validacija modela i na kraju primjena modela.

1. Prikupljanje podataka je korak u kojem se identificira i prikuplja potrebna količina podataka za rješavanje formuliranog problema. Rezultat ovog koraka su neobrađeni podaci.
2. Predobrada podataka je najzahtjevniji i najvažniji korak, zahtjeva dobro razumijevanje problema i podataka. Podrazumijeva deskriptivnu statistiku, identifikaciju i uklanjanje ekstremnih vrijednosti (engl. *outliers)* tj. vrijednosti koje izrazito odskaču od drugih, odabir značajki (obično se počinje s velikim brojem značajki *P* i na kraju odabire *n* najinformativnijih za model), odlučivanje o skaliranju i normalizaciji vrijednosti te istraživanje korelacija između značajki.
3. S obzirom na korelacije, količinu podataka i vrstu problema odabiru se strukture modela, odnosno algoritama strojnog učenja te se podaci dijele na skup za treniranje i za validaciju.
4. Razvoj modela podrazumijeva treniranje ili učenje modela. Korištenjem seta za treniranje različiti algoritmi ML-a se treniraju, podešavanjem njihovih parametara i hiperparametara. Potrebno je voditi računa da ne dođe do pretreniranja (engl. *overfitting*), tj. prekomjernog podešavanja parametara za određeni skup podataka, čime model postaje neprimjenjiv za druge podatke. Također, treba izbjegavati podtreniranje (engl. *underfitting*), što znači korištenje premalo podataka i neinformativnih značajki ili razvoj prejednostavnih modela.
5. Validacija ili vrednovanje modela odvija se na skupu za validaciju. U ovom koraku zapravo se uspoređuju podaci dobiveni modelom i stvarni podaci. Kvaliteta dobivenih modela procjenjuje se na temelju standardne metrike izvedbe (npr. srednje kvadratne pogreške, korijena iz srednje kvadratne pogreške, koeficijenta višestruke determinacije, srednje apsolutne pogreške, kriterija slaganja modela…) i odabire se model s najboljim rezultatima.

6. Model se koristi za predviđanje koristeći nove neobrađene podatke. Često je potrebno kontinuirano praćenje performansi modela u stvarnom vremenu i ukoliko je potrebno ponovno se provodi treniranje modela.[58]

Najveća prednost strojnog učenja je sposobnost detektiranja složenih obrazaca i nelinearnih odnosa unutar velikih skupova podataka. Stoga korištenje ovih tehnika može poboljšati točnost i preciznost predviđanja onečišćenja zraka u usporedbi s tradicionalnim pristupima modeliranju, što dovodi do pouzdanijih informacija za donošenje odluka.[60] U ovom radu razvijaju se dva modela nadziranog učenja i uspoređuje njihova uspješnost u predviđanju indeksa kvalitete zraka za PM_{10} na području grada Zagreba. To su klasifikacijski model "nasumičnih šuma" (engl. *Random Forest Classification,* RFC) i klasifikacijski model "potporni vektorski stroj" (engl. *Support Vector Machine,* SVM).

2.4.1. Algoritam nasumičnih šuma

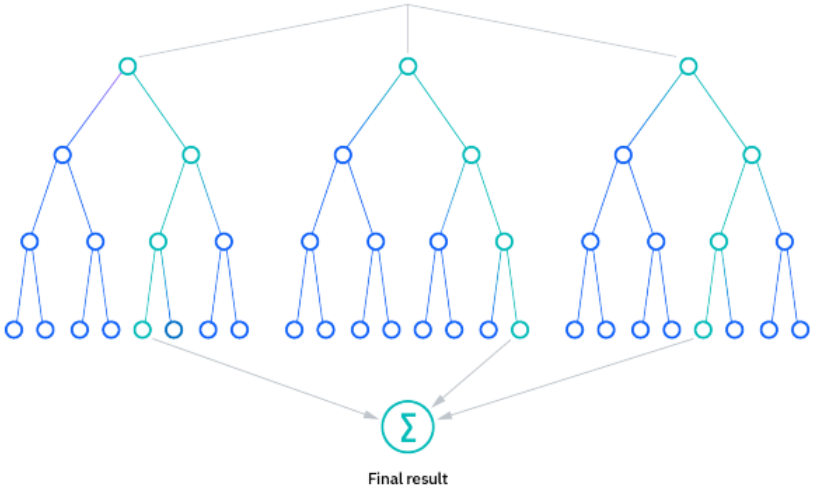
RF algoritam se temelji na modelu stabla odluka (engl. *Decision Tree,* DT), a prema načinu učenja pripada u skupinu nadziranog učenja. U stablu odluka, svaki čvor predstavlja ulaznu varijablu, svaka grana predstavlja odluku ili uvjet koji treba ispuniti, a svaki list na kraju grane odgovarajuću izlaznu varijablu. Čvorovi unutar stabla su poredani prema procijenjenoj važnosti, a odluke se donose temeljem *if* naredbi. Stabla odluka koriste binarni rekurzivni algoritam klasifikacije koji stvara "čiste" čvorove dijeljenjem podataka u dvije homologne grupe. Rekurzivna priroda algoritma znači da se dijeljenje ponavlja dok se ne postigne određena čistoća čvorova. Cijeli niz ovih podjela podataka naziva se stablo. Ovaj algoritam, kada je dopušteno da stablo raste do maksimalne dubine, će ispravno klasificirati ulazne podatke.[61,62]

DT modeli se nazivaju pohlepni jer donose odluke na temelju trenutne najbolje opcije, bez razmatranja budućih posljedica. Pohlepno ponašanje može rezultirati vrlo dubokim stablima koja rijetko generaliziraju na nove podatke, što čini stabla odluka sklonima pretreniranju. Uz pretreniranje modela, jedan od glavnih nedostataka stabala odluka je njihova nepouzdanost s podacima koji sadrže ekstremne vrijednosti. Ako se stablo trenira na skupu podataka bez ekstremnih vrijednosti, može pokazati značajne pogreške kada se primijeni na drugi skup podataka koji ih sadrži.[61,62]

U takvim situacijama, koristi se model slučajnih šuma (engl. *Random Forest*) model. RF stvara više stabala tako da za svako stablo nasumično uzima podskup podataka iz skupa za treniranje. Ta metoda treniranja na nasumičnom podskupu podataka korištenjem *bootstrap*

uzorkovanja, naziva se *bagging*. Ona se koristi za poboljšanje stabilnosti i točnosti modela strojnog učenja, odnosno smanjuje varijancu modela. *Bootstrap* uzorci sadrže neka opažanja (instance, primjere) iz originalnog skupa više puta, dok neka opažanja mogu biti potpuno izostavljena.[63]

Podaci koji nisu uključeni u *bootstrap* uzorak za određeno stablo nazivaju se *out-of-bag* (OOB) podaci. OOB podaci omogućuju da se model testira i procijeni bez potrebe za dodatnim skupom za validaciju, jer svako stablo može biti procijenjeno na temelju opažanja koja nisu bila korištena za njegovo treniranje. Ovaj postupak osigurava da RF model bude robustan i otporan na *overfitting*, te omogućuje procjenu njegove točnosti bez potrebe za dodatnim skupom podataka za validaciju.[63,64] Grafički prikaz načina donošenja odluka prikazan je na slici 2.



Slika 2. *Grafički prikaz načina donošenja odluka RFR modela* [62]

RF se može koristiti u svrhu regresije (rezultat je aritmetička sredina pojedinačnih rezultata, stabala) ili klasifikacije (rezultat je najčešći rezultat pojedinačnih stabala), a prihvaća numeričke i kategorijske varijable te ga je jednostavnije optimizirati u odnosu na druge modele.[65] Tablica 5 sadrži prikaz važnih parametara RF modela i njihovo značenje.

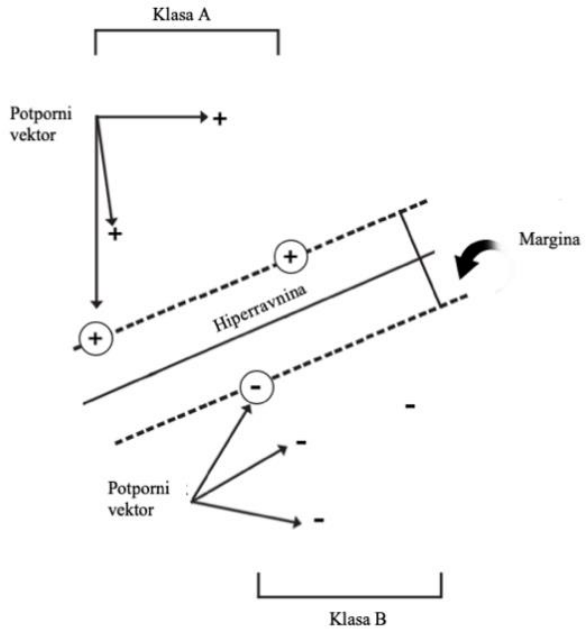
Tablica 5. Parametri RF modela.[63]

Parametar	Opis
criterion	Mjeri kvalitetu podjele.
max_depth	Maksimalna dubina stabla.
min_samples_split	Minimalni broj uzoraka za podjelu čvora.
min_samples_leaf	Minimalni broj uzoraka koji mogu biti u listu.
max_features	Broj značajki za podjelu.
bootstrap	Koristi li se *bootstrap* ili ne.

n_estimators	Broj stabala u šumi.
oob_score	Koriste li se OOB uzorci za provjeru valjanosti i generalizaciju.
random_state	Kontrola nasumičnosti.
class_weight	Težine klasa.

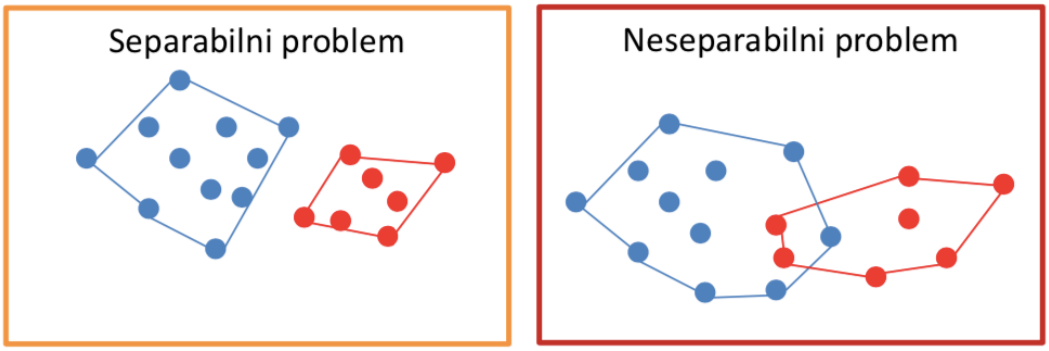
2.4.2. Algoritam potpornih vektora

Model potpornih vektora (SVM) spada u skupinu nadziranog učenja, koristi se za klasifikaciju, regresiju i otkrivanje ekstremnih vrijednosti.[66] SVM koristi različite geometrijske strukture koje služe za odvajanje podataka u različite klase, kao što su hiperplan ili hiperravnina, margina i potporni vektori, prikazani na slici 3.[67]



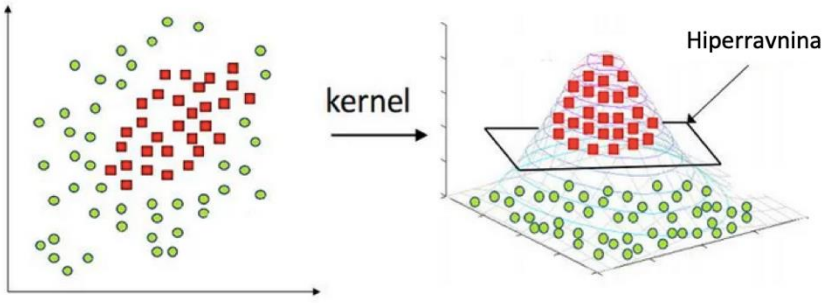
Slika 3. *Prikaz temeljnog načela rada SVM* [67]

Razlikuju se diskriminativni (učenje linije koja razdvaja klase) i generativni (učenje modela za svaku pojedinu klasu) algoritmi. U diskriminativne algoritme spada SVM klasifikacija. SVM klasifikacija provodi optimalne transformacije podataka, koje određuju granice između podatkovnih točaka na temelju unaprijed definiranih klasa, oznaka ili izlaza tražeći maksimalnu marginu. Postoje linearno separabilni i linearno neseparabilni problemi. Linearno separabilni su oni koji imaju prazan presjek konveksnih ljusaka zasebnih klasa.[68] Prikaz različitih vrsta problema je na slici 4.



Slika 4. *Linearno separabilni problem bez preklapanja podataka i konveksne ljuske različitih klasa i linearno neseparabilan problem s preklapanjem* [69]

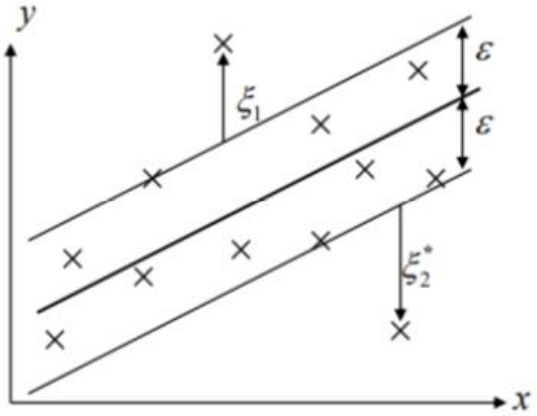
U skupu podataka dijelimo varijable na ulazne, *X*, koje se prikazuju kao vektori značajki i *Y*, koje predstavljaju ciljane varijable. Svaka vrsta značajki predstavlja jednu klasu. Podaci su prikazani u višedimenzionalnom prostoru (prostoru značajki) kao točke. Uz pretpostavku da su klase linearno odvojive postoji beskonačno mnogo rješenja. Cilj SVM-a je pronaći optimalnu hiperravninu koja razdvaja različite vrste podataka. Optimalna hiperravnina je ona s najvećom marginom, a predstavlja udaljenost između točaka različitih klasa najbližih hiperravnini, odnosno između potpornih vektora. Dakle, funkcija odluke definirana je preko primjera iz skupa za učenje, tzv. potpornih vektora (engl. *support vectors*). Ova metoda je prvenstveno osmišljena za binarne probleme, ako se radi s višeklasnim problemima (nelinearni problem) dolazi do modifikacije rada SVM. U tom slučaju koriste se metode jezgre (engl. *Kernel methods*) za transformaciju značajki podataka upotrebom funkcija jezgre. Jezgrene funkcije oslanjaju se na proces preslikavanja složenih skupova podataka u više dimenzije na način koji olakšava odvajanje točaka. Funkcija pojednostavljuje granice podataka za nelinearne probleme dodavanjem viših dimenzija za mapiranje složenih podatkovnih točaka. Jedan primjer mapiranja u višedimenzionalni prostor prikazan je na slici 5. Metode jezgre mogu biti linearne, polinomne, RBF (engl. *Radial Basis Function*) ili sigmoidne.[66,68,70]



Slika 5. *Prikaz rješavanja nelinearnih problema pomoću funkcija jezgre* [71]

Stroj potpornih vektora je algoritam primarno osmišljen za rješavanje klasifikacijskih problema, no koristi se i za regresijske probleme pod nazivom regresija potpornih vektora (engl. *Support Vector Regression*, SVR).

Cilj je pronaći funkciju koja najbolje aproksimira dane primjere, kao i kod linearne regresije. Međutim, za razliku od linearne regresije, SVR traži funkciju $f(X)$ koja u najgorem slučaju za primjer $x(i)$ ima ε odstupanje od "prave" funkcijske vrijednosti $y(i)$ te je što monotonija. Monotonost se odnosi na to da model reagira na predvidljiv i konzistentan način s obzirom na promjenu ulaznih značajki. Konstanta $C > 0$ određuje kompromis između monotonosti funkcije f i iznosa do kojeg se toleriraju odstupanja veća od ε.[72] Na slici 6 prikazan je regresijski postupak.



Slika 6. *Regresija potpornih vektora* [72]

Ovaj pristup omogućava precizniju i robusniju predikciju u prisutnosti šuma u podacima, a također olakšava razumijevanje značajnosti pojedinih ulaznih varijabli u modelu. Regresija potpornih vektora ima sposobnost učinkovitog upravljanja kompleksnim i nelinearnim odnosima u podacima.[72] Tablica 6 sadrži prikaz važnih parametara SVM modela i njihovo značenje.

Tablica 6. Prikaz važnih parametara SVM modela i njihovo značenje.

Parametar	Objašnjenje
C	Parametar određuje jačinu regularizacije, koja pomaže u sprječavanju prenaučenosti modela. Njegova vrijednost je obrnuto proporcionalna snazi regularizacije, što znači da veće vrijednosti C smanjuju regularizaciju.
kernel	Jezgra definira način na koji model transformira ulazne podatke kako bi pronašao optimalnu granicu razdvajanja između klasa. Postoji nekoliko opcija:
gamma	Koliko daleko pojedinačni uzorci podataka utječu na konačnu granicu razdvajanja. Kod nižih vrijednosti, svaki podatak ima širi utjecaj na model, dok veće vrijednosti daju veću važnost pojedinačnim podacima u blizini granice razdvajanja. *Scale* (automatski postavlja vrijednost u skladu s varijabilnošću podataka), *Auto* (postavlja vrijednost na 1 podijeljeno s brojem značajki).

2.4.3. Python

Python je interpretirani programski jezik, što znači da se kod izvršava direktno bez potrebe za prethodnom kompilacijom u strojni jezik. Ova karakteristika omogućuje brže testiranje i razvoj jer se promjene u kodu mogu odmah isprobati. Python je također visoko razinski jezik, što ga čini bližim ljudskom jeziku nego strojnom jeziku, olakšavajući tako učenje i pisanje koda.

Python se ističe objektno orijentiranim pristupom, podržavajući paradigmu programiranja u kojoj se koncepti modeliraju kao "objekti" s vlastitim podacima (atributima) i funkcionalnostima (metodama). Ova struktura omogućuje bolju organizaciju programa.

Također, Python koristi dinamičku semantiku, što znači da se mnoge odluke o vrsti podataka, poput toga je li neka varijabla broj ili tekst, donose tijekom izvođenja programa, a ne unaprijed. Ovo pruža fleksibilnost i olakšava pisanje generičkog koda koji može raditi s različitim vrstama podataka. Python dolazi s opsežnom standardnom knjižicom koja pokriva širok spektar funkcionalnosti, uključujući rad s datotekama, internet protokolima, upravljanje sustavom i još mnogo toga. Također, podržava modularnost koda putem modula i paketa, što omogućuje bolje organiziranje i ponovnu upotrebu koda.[73]

Programi napisani u Pythonu obično su puno kraći od ekvivalentnih C, C++ ili Java programa jer omogućuje izražavanje složenih operacija u jednoj izjavi, grupiranje iskaza uvlačenjem umjesto početnim i završnim zagradama te nisu potrebne deklaracije varijabli ili argumenata. Jednostavnost, fleksibilnost i moćne knjižnice ovog programskog jezika omogućuju primjenu Pythona u raznim područjima, uključujući web razvoj, znanost o podacima, strojno učenje, automatizaciju, razvoj igara…[74]

U ovom radu od Pythonovih knjižica koristile su se:

1. *Pandas* – za čišćenje, pripremu i analizu podataka.
2. *SciPy* - knjižnica za znanstveno i tehničko računalstvo. Uključuje module za optimizaciju, integraciju, interpolaciju, probleme svojstvenih vrijednosti i statistiku.
3. *Seaborn (sns)* - za vizualizaciju podataka, pruža sučelje visoke razine za crtanje raznih vrsta grafova.
4. *Statsmodels (seasonal_decompose)* - Knjižnica za statističko modeliranje i ekonometriju, koja pruža klase i funkcije za procjenu mnogih različitih statističkih modela, kao i za provođenje statističkih testova.

5. *Matplotlib (plt)* - sveobuhvatna biblioteka za stvaranje statičnih, animiranih i interaktivnih vizualizacija u Pythonu.
6. *Windrose (WindroseAxes)* - Posebna biblioteka za izradu dijagrama ruže vjetra, koji se koriste za predstavljanje raspodjele smjera i brzine vjetra.
7. *Scikit-learn* – koristi se za treniranje i razvoj modela, standardizaciju vrijednosti, PCA analizu…

Ove su knjižice bitne za različite faze analize podataka, od predobrade i istraživačke analize podataka do izgradnje, evaluacije i interpretacije modela strojnog učenja.[75]

2.4.4. Predobrada podataka

Najvažniji i najzahtjevniji korak u izradi modela strojnog učenja je predobrada podataka. U predobradi podataka koriste se numerički izračuni i grafovi za prikaz raspodjele vrijednosti podataka. Pomoću deskriptivne statistike računaju se mjere centralne tendencije (aritmetička sredina, medijan, kvartili i dominantna vrijednost, mod), mjere disperzije (raspon i interkvartil) i mjere rasipanja (standardno odstupanje, varijanca i koeficijent varijance). Crtaju se grafovi za vizualni prikaz vrijednosti podataka poput normalne raspodjele, *box-plot* dijagrama i histograma. To se radi u svrhu detekcije i uklanjanja ekstremnih vrijednosti te za bolje razumijevanje ponašanja značajki.[76–78]

2.4.5. Inženjerstvo značajki

Inženjerstvo značajki je postupak kojim se „sirove“ informacije pretvaraju u informacije koje su u odgovarajućem obliku za potrebe modeliranja, odabiru se najvažnije, te se izbacuju i/ili kreiraju nove vrste značajki, kako bi model što bolje predviđao izlazne varijable. Provodi se prilikom predobrade podataka i najzahtjevniji je korak u razvoju modela strojnog učenja.

Vrlo bitno je tehničko i teorijsko razumijevanje značajki, a tako i samog problema, s obzirom da je upravo kvaliteta ulaznih podataka za treniranje modela odgovorna za uspješnost rada modela. Postoji više postupaka inženjerstva značajki koji se provode ovisno o vrsti podataka i problema.

Ovi postupci uključuju izdvajanje i odabir značajki, u tu svrhu se koriste analiza glavnih komponenti (engl. *Principal Component Analysis*, PCA) i linearna diskriminantna analiza (engl. *Linear discriminant analysis*, LDA), koje kombiniraju i transformiraju izvorne značajke te proizvode nove. Glavna razlika je u tome što PCA proizvodi nove sastavne varijable namijenjene maksimiziranju varijance podataka, dok LDA proizvodi sastavne

varijable prvenstveno namijenjene maksimiziranju klasne razlike u podacima. Osim prethodno opisanih dviju metoda mogu se koristit i dijagrami korelacije ili raspršenja koji pokazuju koliko jedna varijabla utječe na drugu (npr. *Pearsonov* ili *Spearmanov* koeficijent korelacije). Analizom dijagrama odlučuje se o uklanjanju značajki ili dodavanju novih te se može provesti množenje i dijeljenje značajki kako bi se smanjio broj istih.[79]

Nadalje, uključuju jednokratno kodiranje (engl. *One hot encoding*) i razvrstavanje (engl. *Binning*), koji se koriste za pretvaranje numeričkih vrijednosti u kategorijske i obrnuto. Provodi se i skaliranje podataka, kako bi se izbjeglo treniranje modela na temelju brojčanih vrijednosti značajki. Dakle, ova tehnika se provodi u svrhu normalizacije podataka u fiksnom rasponu. Naime, često su varijable (značajke) različite fizikalne veličine (različitih mjernih jedinica) te je moguće da će se povećati značaj brojčano većih veličina nad manjima prilikom treniranja modela. Razlikuje se *Min-Max, z-score* normalizacija i robusno skaliranje. U ovom radu korištena je *MinMaxScaler* funkcija procjene koja se nalazi unutar *Scikit-learn* knjižnice.[80,81]

2.4.6. Treniranje modela

Prije treniranja modela cijeli skup podataka dijeli se na 3 dijela, jedan za treniranje (obuku), jedan za testiranje i jedan za validaciju modela. Trening set podataka je najveći, dok su test i validacijski skup često iste veličine. Skup za obuku koristi se za procjenu ili učenje parametara modela. To se zove "prilagodba modela". Skup za validaciju koristi se za procjenu kriterija odabira za odabir modela, a testni podaci koriste se za procjenu pogreške generalizacije konačno odabranog modela.

Kako bi se ostvarilo što bolje predviđanje modela bitno je trenirati i validirati model na različitom setu podataka, u tu svrhu može se koristiti unakrsna validacija (engl. *Cross validation*, CV). Ova tehnika se temelji na iteraciji, nasumično se odabiru podaci koji će se koristiti za validaciju modela, dok se ostali koriste za trening i testiranje modela. Konačno, rezultati iz svakog koraka provjere valjanosti izračunavaju se u prosjeku kako bi se dobila robusnija procjena izvedbe modela. Ona pomaže osigurati da je model robustan i dobro generaliziran na nove podatke (sprječava pretreniranje modela). Tehnike CV uključuju *k-fold* CV, *leave-one-out* CV, *holdout* validaciju i *stratified* CV.[82,83] U ovom istraživanju korištena je *k-fold* unakrsna validacija.

2.4.7. Optimizacija modela

Kako bi se dobio najoptimalniji model potrebno je odabrati najvažnije značajke u odgovarajućem obliku i optimalne hiperparametre za kreiranje modela.

Općenito, svaki model koristi znatan broj hiperparametara koji određuju njegov način rada. Hiperparametre je moguće podešavati metodom pokušaja i pogreške, no s obzirom na različite kombinacije vrijednosti parametara i različite vrste parametara koje modeli koriste taj proces bi bio mukotrpan i dugotrajan. Zato se koriste algoritmi koji traže optimalne parametre modela kojima se vladanje modela najmanje razlikuje od vladanja stvarnog procesa. Kada se govori o optimizaciji hiperparametara, tada se parametri algoritma nazivaju hiperparametri, dok se koeficijenti koje pronalazi sam algoritam optimizacije nazivaju parametrima. Nakon što se postave odgovarajuće vrijednosti za hiperparametre, izvedba modela može se značajno poboljšati.[84]

Algoritam *GridSearchCV* često se koristi za pronalazak optimalnih parametara modela. *GridSearchCV* automatski prolazi kroz sve kombinacije unaprijed ručno definiranih hiperparametara, te za svaku kombinaciju hiperparametara koristi unakrsnu validaciju kako bi procijenio izvedbe modela.[85,81] Na temelju najbolje izvedbe odabire optimalne parametre za model. *GridSearch* koristi parametre zapisane u tablici 7.

Tablica 7. Parametri algoritma *GridSearchCV*.[81]

Parametar	Zadana vrijednost	Opis
estimator	N/A	procjenitelj koji implementira *scikit-learn* sučelje (model za koji se koristi)
param_grid	N/A	rječnik s nazivima parametara kao ključevima i listama mogućih vrijednosti kao vrijednostima
scoring	None	strategija za evaluaciju performansi modela
n_jobs	None	broj paralelnih poslova, `-1`koristi sve procesore
refit	TRUE	ponovno treniranje procjenitelja s najboljim pronađenim parametrima na cijelom skupu podataka
cv	None	strategija podjele za unakrsnu validaciju, ako je None, koristi se zadana 5-*fold* unakrsna validacija.
verbose	0	razina detaljnosti izlaza, veće vrijednosti daju detaljnije poruke (0-3+)
pre_dispatch	'2×*n_jobs*'	kontrolira broj poslova koji se distribuiraju tijekom paralelne izvedbe
return_train_score	FALSE	ako je *True*, vraća rezultate treniranja

2.4.7. Validacija modela

U svakom projektu strojnog učenja koristi se više različitih modela, treniranih na istom skupu podataka i odabire se onaj s najboljom izvedbom na temelju prediktivnosti ili deskriptivnosti. Kvaliteta modela, odnosno podudarnost vladanja modela s vladanjem stvarnog procesa, određuje se pomoću kriterija za vrednovanja modela. Oni se temelje na usporedbi predviđenih vrijednosti i stvarnih vrijednosti ciljane varijable. Kriteriji vrednovanja modela se razlikuju za klasifikaciju i regresiju. Model bi trebao imati što nižu vrijednost

pogreške generalizacije koja karakterizira njegovu izvedbu predviđanja.[86–88] U tablici 8 su prikazani kriteriji, njihove formule za klasifikaciju, a u tablici 9 za regresiju.

Tablica 8. Kriterij vrednovanja modela i njihove formule za klasifikaciju.[86]

Kriterij	Formula	Metoda
Točnost (engl. *Accuracy*)	Acc = (TP + TN) / (TP + FP + FN + TN)	Binarna i višeklasna klasifikacija
Preciznost (engl. *Precision*, P)	P = TP / (TP + FP)	Binarna i višeklasna klasifikacija
Odziv (engl. *Recall*, R)	R = TP / (TP + FN)	Binarna i višeklasna klasifikacija
Specifičnost (engl. *Specificity*)	S = TN / (TN + FP)	Binarna klasifikacija
Mjera F1 (engl. *F1 Score*)	F1 = 2 * (P * R) / (P + R)	Binarna i višeklasna klasifikacija

U formulama TP (engl. *True Positive*) je broj točno predviđenih pozitivnih primjera, a TN (engl. *True Negative*) je broj točno predviđenih negativnih primjera. FP (engl. *False Positive*) je broj netočno predviđenih pozitivnih primjera (pozitivno predviđenih, ali stvarno negativnih), FN (engl. *False Negative*) je broj netočno predviđenih negativnih primjera (negativno predviđenih, ali stvarno pozitivnih). Matrica zabune (engl. *Confussion matrix*) se koristi za prikaz broja točnih i netočnih predikcija modela u usporedbi sa stvarnim oznakama. Kod višeklasne klasifikacije ($K > 2$), matrica zabune dimenzija je $K \; x \; K$.

Da bi se dobile mjere preciznosti, odziva i F1 za višeklasnu klasifikaciju, mogu se koristiti dvije metode:

1. Makro-uprosječivanje (engl. *Macro-averaging*), gdje se izračunava kriterij vrednovanja za svaku klasu zasebno. Dobivene vrijednosti se zatim prosječno zbroje da se dobije konačna makro-mjera.
2. Mikro-uprosječivanje (engl. *Micro-averaging*), gdje se zbrajaju svi elementi svih binarnih matrica zabune po klasama. Iz združene matrice zabune izračunava se kriterij vrednovanja kao da se radi o običnoj binarnoj klasifikaciji.

Makro-mjere daju jednaku težinu svim klasama, dok mikro-mjere favoriziraju klase s većim brojem primjera. Točnost (engl. *Accuracy*) se izračunava izravno iz matrice zabune dimenzija $K \; x \; K$, dok se za ostale mjere koristi makro ili mikro uprosječivanje kako bi se dobile vrednovane vrijednosti za cijeli skup podataka.

Nadalje za prikaz uspješnosti predviđanja modela koriste se i razni dijagrami i grafovi. Kod klasifikacije to su krivulja preciznost-odziv (engl. *precision-recall curve*, PR krivulja) i površina pod krivuljom ROC (engl. *Receiver Operating Characteristic*).

PR krivulja je grafički prikaz odnosa između preciznosti i odziva. Na *x*-osi prikazuje se odziv, a na *y*-osi preciznost, a krivulja se dobiva tako da se za trenirani model vrijednost praga postepeno smanjuje od 1 do 0. Preferiraju se krivulje koje su bliže točki (P=1, R=1).

Površina pod krivuljom ROC, skraćeno AUC (engl. *area under ROC curve*). ROC krivulja je vrijednost stope stvarnih pozitivnih vrijednosti (engl. *True Positive Rate*, TPR), odnosno odziva, kao funkcije stope lažno predviđenih pozitivnih instanci (engl. *False Positive Rate*, FPR), odnosno ispadanja (engl. *fall-out*). FPR je omjer broja lažno pozitivnih instanci (FP) i ukupnog broja stvarnih negativnih instanci (FP + TN). X-os predstavlja FPR, a Y-os predstavlja TPR. AUC označava površinu ispod ROC krivulje, a ima raspon od 0 do 1 (označava dobro predviđanje).[86,87]

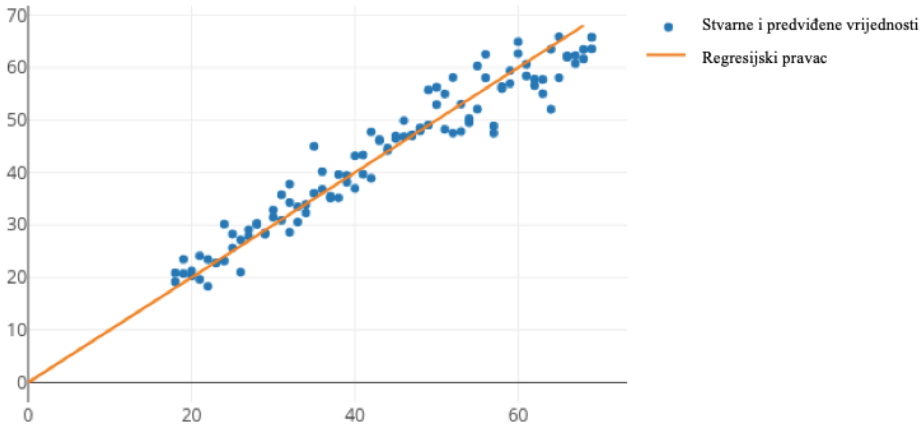
Tablica 9. Kriteriji vrednovanja modela i njihove formule za regresiju.[89,90]

Naziv funkcije	Značenje	Formula
Koeficijent determinacije (engl. *R-squared*, R^2)	Mjeri koliko dobro regresijski model objašnjava varijabilnost zavisne varijable.	$R^2 = \Sigma((\hat{y}_i - \bar{y})^2) / \Sigma((y_i - \bar{y})^2),\ 0 \leq R^2 \leq 1$
Srednja apsolutna pogreška (engl. *Mean Absolute Error*, MAE)	Prosjek apsolutnih razlika između predviđenih i stvarnih vrijednosti.	$MAE = (1/n) * \Sigma\lvert y_i - \hat{y}_i\rvert$
Srednja pristranost pogreške (engl. *Mean Bias Error*, MBE)	Srednja razlika između predviđenih i stvarnih vrijednosti.	$MBE = (1/n) * \Sigma(y_i - \hat{y}_i)$
Relativna apsolutna pogreška (engl. *Relative Absolute Error*, RAE)	Omjer ukupne apsolutne pogreške i apsolutne razlike između srednje i stvarne vrijednosti.	$RAE = \Sigma\lvert y_i - \hat{y}_i\rvert / \Sigma\lvert y_i - \bar{y}\rvert$
Srednja apsolutna postotna pogreška (engl. *Mean Absolute Percentage Error*, MAPE)	Apsolutna razlika između stvarnih i predviđenih vrijednosti podijeljena sa stvarnom vrijednošću.	$MAPE = (100/n) * \Sigma\lvert y_i - \hat{y}_i\rvert / \lvert y_i\rvert$
Srednja kvadratna pogreška (engl. *Mean Squared Error*, MSE)	Srednja kvadratna pogreška, računa srednju kvadratnu razliku između predviđenih i stvarnih vrijednosti.	$MSE = (1/n) * \Sigma(y_i - \hat{y}_i)^2$
Korijen srednje kvadratne pogreške (engl. *Root Mean Squared Error*, RMSE)	Kvadratni korijen srednje kvadratne pogreške.	$RMSE = \sqrt{(MSE)}$
Relativna kvadratna pogreška (engl. *Relative Squared Error*, RSE)	Omjer srednje kvadratne pogreške i kvadratne razlike između stvarne i srednje vrijednosti.	$RSE = \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$
Normalizirani korijen srednje kvadratne pogreške (engl. *Normalized Root Mean Squared Error*, NRMSE)	RMSE se dijeli s varijantom kao što su maksimalna vrijednost, srednja vrijednost itd.	$NRMSE = RMSE / std(y)$
Relativni korijen srednje kvadratne pogreške (engl. *Relative Root Mean Squared Error*, RRMSE)	RMSE koja se normalizira rasponom ciljne varijable.	$RRMSE = \sqrt{(\Sigma(y_i - \hat{y}_i)^2 / \Sigma\, \hat{y}_i^2)}$
Korijen srednje kvadratne	Primjena logaritama na stvarne i	$RMSLE = \sqrt{((1/n) * \Sigma(\log(y_i + 1) -}$

logaritamske pogreške (engl. *Root Mean Squared Logarithmic Error*, RMSLE)	predviđene vrijednosti i računa njihove razlike.	$\log(\hat{y}_i + 1))^2$)
Huberova pogreška (engl. *Huber Loss*, HL)	Kombinacija linearnih i kvadratnih metoda ocjenjivanja s hiperparametrom delta (δ).	HL = ako je $\lvert y_i - \hat{y}_i \rvert \leq \delta$: 0,5 * $(y_i - \hat{y}_i)^2$, inače: δ * ($\lvert y_i - \hat{y}_i \rvert$ - 0,5 * δ^2)
Log Cosh gubitak (engl. *Log Cosh Loss*, LCL)	Računa logaritam hiperboličkog kosinusa pogreške.	LCL = $\Sigma\log(\cosh(\hat{y}_i - y_i))$
Kvantilna pogreška (engl. *Quantile Loss*, QL)	Primjenjuje se za predviđanje kvantila, proširenje MAE osim za 50-ti percentil (MAE).	QL = ako je $y_i >= \hat{y}_i$: γ * $\lvert y_i - \hat{y}_i \rvert$, inače: (1-$\gamma$) * $\lvert y_i - \hat{y}_i \rvert$

*Gdje je n broj uzoraka u skupu podataka, y_i je predviđena vrijednost za i-ti uzorak, a $\bar{y}$ je ciljna vrijednost za i-ti uzorak

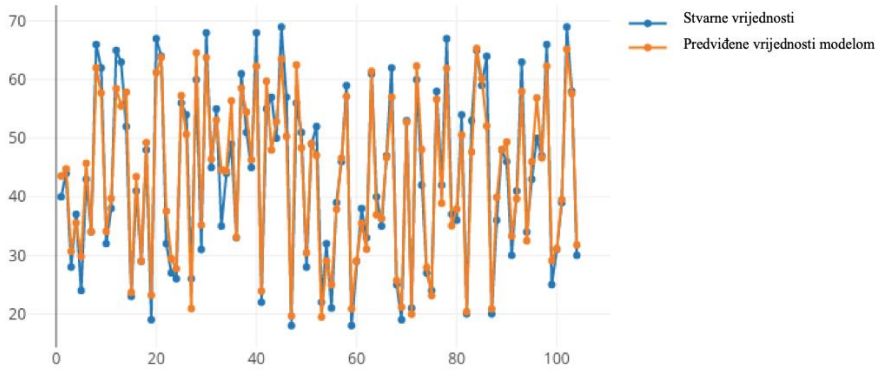
Kod regresijskih problema računa se *loss* funkcija, tj. funkcija pogreške, koja računa razliku između predviđenih i stvarnih vrijednosti izlaza modela. Kod regresije prikazuju se numeričke vrijednosti pomoću grafova poput dijagrama raspršenja i pravca linearne regresije, prikazanih na slici 7. Iz dijagrama se izračunava *R*, koeficijent korelacije (-1 do 1), odnosno R^2.[90–93]



Slika 7. *Dijagram raspršenja i pravac linearne regresije* [91]

R^2 je omjer zbroja kvadrata odstupanja protumačenog modelom i zbroja kvadrata odstupanja eksperimentalnih podataka, što je iznos R^2 bliža 1, to model ima bolje predviđanje. Također se računaju reziduali i crtaju grafovi reziduala (histogram ili dijagram raspršenja). Rezidual je pogreška i računa se za svaki par predviđena-stvarna vrijednost.[94]

Za klasifikaciju i regresiju se koriste dijagrami raspršenja koji prikazuju stvarne i predviđene vrijednosti modelom i njihovu promjenu tijekom vremena. Primjer dijagrama raspršenja za regresijski problem je prikazan na slici 8.



Slika 8. *Dijagram raspršenja sa stvarnim i predviđenim vrijednostima za regresiju* [91]

2.5. Pregled literature

U radu Grange et al. RF se pokazao kao dobar model za predviđanje koncentracije PM_{10} na 31 lokaciji (R^2 do 71%).[61] U radu Lovrić et al. (2022.) korišteni su RF i LightGBM modeli za predviđanje koncentracije PM_{10} u gradu Zagrebu (R^2 veći od 0,77)[51], dok su u radu Šimić et al. korišteni regresijski modeli Lasso i RF, koji u usporedbi s drugim modelima pokazuju najveće prediktivne mogućnosti za koncentraciju PM, s Lassoom koji je pobijedio tri od četiri puta.[95] AQI za New Delhi, Bangalore, Kolkata i Hyderabad predviđan je pomoću tri različite ML metode, SVR, RFR i CatBoost regresijom. RFR se pokazao kao najuspješniji.[96] SVR model se koristi i u istraživanju S. Bhattacharya, model predviđa razine različitih onečišćujućih tvari, kao i indeks kvalitete zraka (AQI), s točnošću od 93,4 %.[97]

U radu M.A. Haq korišten je model klasifikacije SMOTEDNN (engl. *Synthetic Minority Oversampling Technique with Deep Neural Network)*, Tehnika sintetskog manjinskog preduzorkovanja, SVM s jezgrenom funkcijom RBF, RF klasifikacija, XGBoost klasifikacijski model i k-algoritam najbližih susjeda (engl. *k-nearest neighbors*, KNN) za predviđanje AQI klasa. SMOTE je tehnika sintetskog manjinskog preduzorkovanja, koristi se za preduzorkovanje vrijednosti manjinske klase na temelju dupliciranja vrijednosti manjinske klase (KNN metodom). Svi modeli (osim KNN) imaju točnost veću od 99 %.[98]

U radu Ravindiran et al. AQI se predviđao pomoću RF, *CatBoost*, *LightGBM*, Adaboost i XGBoost metoda, svi modeli pokazuju vrlo dobra predviđanja s vrijednosti R2 blizu 1.[3] U istraživanju Imam et al. upotrjebljeni su nadzirani modeli klasifikacije, naivni Bayes, logistička regresija, DT klasifikator, SVC i RFC za predviđanje AQI klasa. Pokazalo se kako su najuspješniji modeli SVM i RFC.[99]

U ovom radu koristit će se klasifikacijski modeli SVM i RF za predviđanje AQI kategorija.

3. MATERIJALI I METODE

Za razvoj modela i predobradu podataka korišten je programski jezik Python. U narednim poglavljima biti će prikazani prikupljeni obrađeni dnevni podaci za svaku ispitivanu lokaciju za cijeli vremenski period mjerenja. Na kraju se za model stvaraju novi dokumenti ovisno o lokaciji prema postojećim prikupljenim podacima tako da je razdoblje korištenih podataka ujednačeno.

3.1. Prikupljanje i predobrada podataka

Podaci o koncentracijama frakcije lebdećih čestica $PM_{2,5}$ i PM_{10} u zraku i meteorološki podaci prikupljeni su na 3 mjerne postaje državne mreže za trajno praćenje kvalitete zraka na području Grada Zagreba; Zagreb-1 (ZG1), Zagreb-2 (ZG2) i Zagreb-3 (ZG3). ZG1 i ZG2 su karakterizirane kao prometne mjerne postaje dok je ZG3 karakterizirana kao pozadinska mjerna postaja. Na tim postajama prikupljani su satni podaci. Više detalja o mjernim postajama dostupno je na portalu Kvaliteta zraka u Republici Hrvvatskoj.[100] Uz to, korišteni su meteorološki podaci s meteorološke postaje Zagreb-Maksimir (Maksimir) DHMZ-a. Svi podaci osim oborina na lokaciji Maksimir mjereni su u 7, 14 i 21 h, a oborine jednom dnevno. Na slici 9 prikana je satelitska karta grada Zagreba i označene su lokacije mjernih postaja.

Slika 9. *Lokacije mjernih postaja u Zagrebu*

U predobradi podataka provedeno je i inženjerstvo značajki.

Na gore navedene tri mjerne postaje masene koncentracije PM_{10} i meteorološki podaci prikupljeni su za razdoblje od 2015. do 2020. Masene koncentracije $PM_{2,5}$ i meteorološki podaci prikupljeni su za razdoblje od 2022. do kraja 2023. (s kontinuiranim mjerenjima frakcije lebdećih čestica $PM_{2,5}$ na ovoj lokaciji započelo se 2021. godine). Mjerenja na lokacijama ZG1, ZG2 i ZG3 provodila su se kontinuirano pri čemu su u sustav pohranjivani satni prosjeci koncentracija iz čega su se zatim izračunati dnevni (24-satni) prosjeci za svaki dan. Na lokaciji Maksimir također su izračunati dnevni prosjeci iz dostupnih podataka od 2015. do kraja 2023. godine.

3.2. Meteorološki podaci

Predobrada meteoroloških podataka uključivala je promjenu vrste podataka ukoliko je bilo potrebno (npr. datum i vrijeme u *Datetime* formatu), uklanjanje *outlier-a* pomoću *rolling-mean* funkcije i zamjena nedostajućih vrijednosti (engl. *Not a Number*, NaN) pomoću vremenske interpolacije i sezonske dekompozicije. Za varijable temperatura, relativna vlažnost i tlak zraka, dnevne vrijednosti izračunate su kao aritmetičke sredine na temelju satnih podataka. Dnevne vrijednosti smjera i brzine vjetra određene su kao medijan iz satnih vrijednosti za taj određeni dan. U sljedećim poglavljima prikazani su dijagrami i grafovi podataka nakon predobrade. Tablice deskriptivne analize uključuju ukupni broj vrijednosti u skupu podataka, prosječnu vrijednost, standardnu devijaciju (mjera disperzije skupa podataka), minimalnu i maksimalnu vrijednost te 25., 75. percentil i medijan (50. percentil) za određenu varijablu.

Vrste meteoroloških podataka, broj podataka, razdoblje i mjerne postaje na kojima su prikupljeni prikazani su u tablici 10 i 11.

Tablica 10. Dnevni meteorološki podaci korišteni za model predviđanja AQI na temelju PM_{10}.

Postaje	Meteorološki podaci	Period	Broj podataka pojedine varijable
Zagreb-1	Smjer i brzina vjetra, temperatura, relativna vlažnost zraka	21.4.2016 – 31.12.2020.	1716
Zagreb-2	Smjer i brzina vjetra, temperatura, relativna vlažnost zraka	12.05.2016. – 20.06.2020.	1501
Zagreb-3	Smjer i brzina vjetra	21.04.2016 – 31.12.2020.	1716
Zagreb-Maksimir	Smjer i brzina vjetra, temperatura, relativna vlažnost zraka, tlak zraka, količina oborina	01.01.2016. – 31.12.2020.	2192

Tablica 11. Dnevni meteorološki podaci korišteni za model predviđanja AQI na temelju $PM_{2,5}$.

Postaja	Meteorološki podaci	Period	Broj podataka pojedine varijable
Zagreb-3	Smjer i brzina vjetra, temperatura, relativna vlažnost zraka	09.03.2022 – 31.12.2023.	663
Zagreb-Maksimir	Smjer i brzina vjetra, temperatura, relativna vlažnost zraka, tlak zraka, količina oborina	09.03.2022 – 31.12.2023.	663

Postaje (ZG1, ZG2 i ZG3) su opremljene uređajima za automatsku analizu praćenih parametara i opremom za aktivno sakupljanje. Nalaze se u urbanoj zoni i u blizini prometnih ulica. Postaja Maksimir nalazi se u blizini parka Maksimir. Tijekom koraka predobrade podataka vremenski raspon korištenih podataka se smanjuje (za ZG1, ZG2 i ZG3) jer nedostaju podaci na svim lokacijama za sve meteorološke varijable kroz dulje vremenske periode. Tako skraćeni skupovi se koriste za izradu modela i pridodaju im se podaci s lokacije Maksimir za to određeno razdoblje, a u narednim poglavljima prikazivat će se statistički podaci i grafovi meteoroloških varijabli za razdoblje od 2016. do 2021. godine.

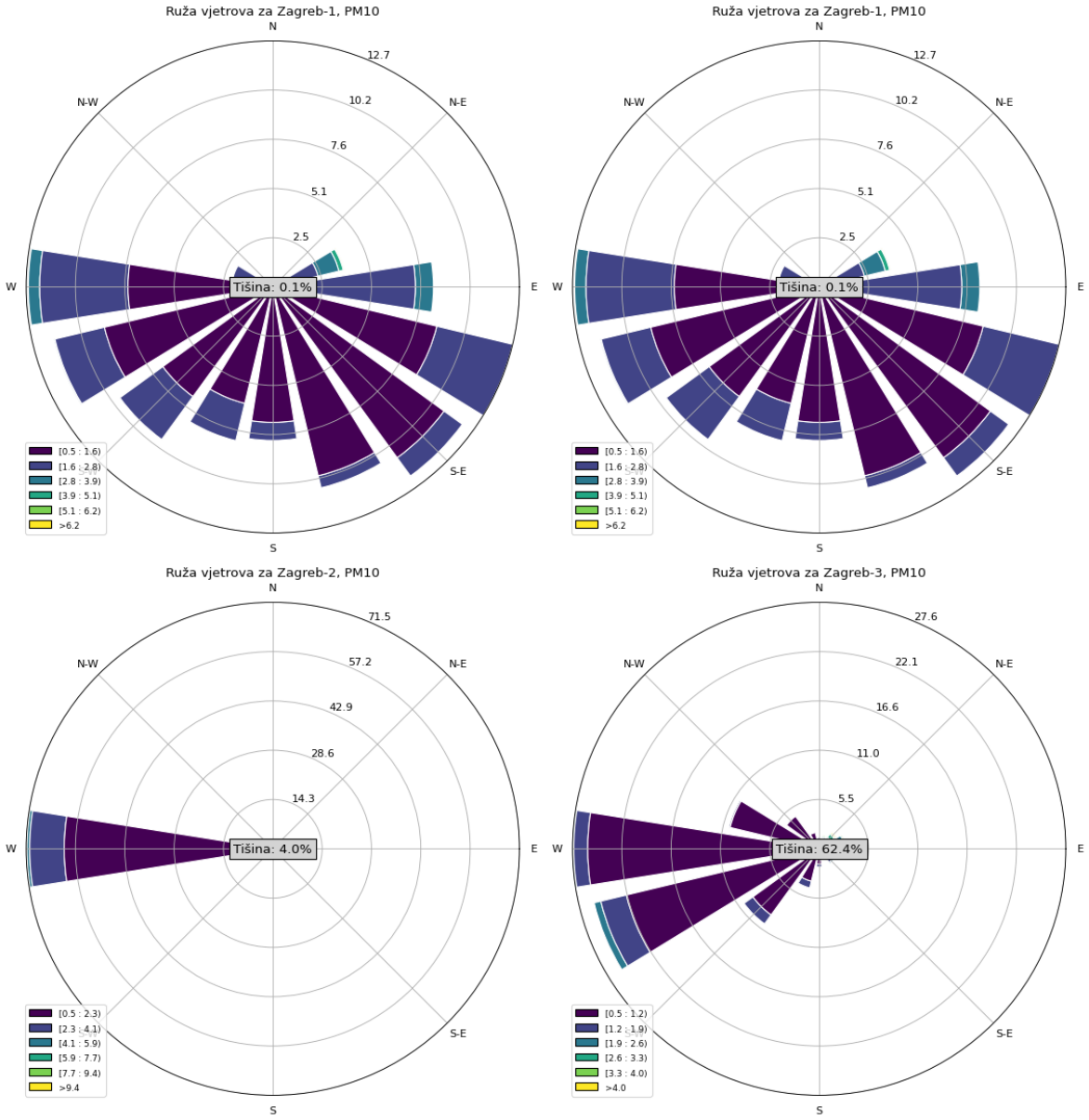
3.2.1 Vjetar

Na slici 10 prikazani su dijagrami ruže vjetrova za tri mjerne postaje u gradu i za postaju Maksimir za razdoblje od određenih datuma 2016.-2021. (prikazano u tablici 10), korištene za predviđanje AQI (PM_{10}). Na slici 11 prikazani su dijagrami ruže vjetrova za mjernu postaju Zagreb-3 i Maksimir za razdoblje od 2022.-2024. (prikazano u tablici 11), korištene za predviđanje AQI ($PM_{2,5}$). Oni prikazuju raspodjele brzine i smjera vjetra tijekom određenog razdoblja. Boje na slikama označuju jačinu vjetra u ms^{-1} i prikazane su u legendi. Smjer vjetra je prikazan pomoću stupaca, a veličina stupca odgovara učestalosti puhanja vjetra u određenom smjeru (postotak je označen na kružnicama). Smjer vjetra podijeljen je u 16 skupina, redom od 0° do 360° s pomakom od 22,5° (tablica 12). Za kod je korištena i brzina vjetra prema Beaufortovoj ljestvici (tablica 4).

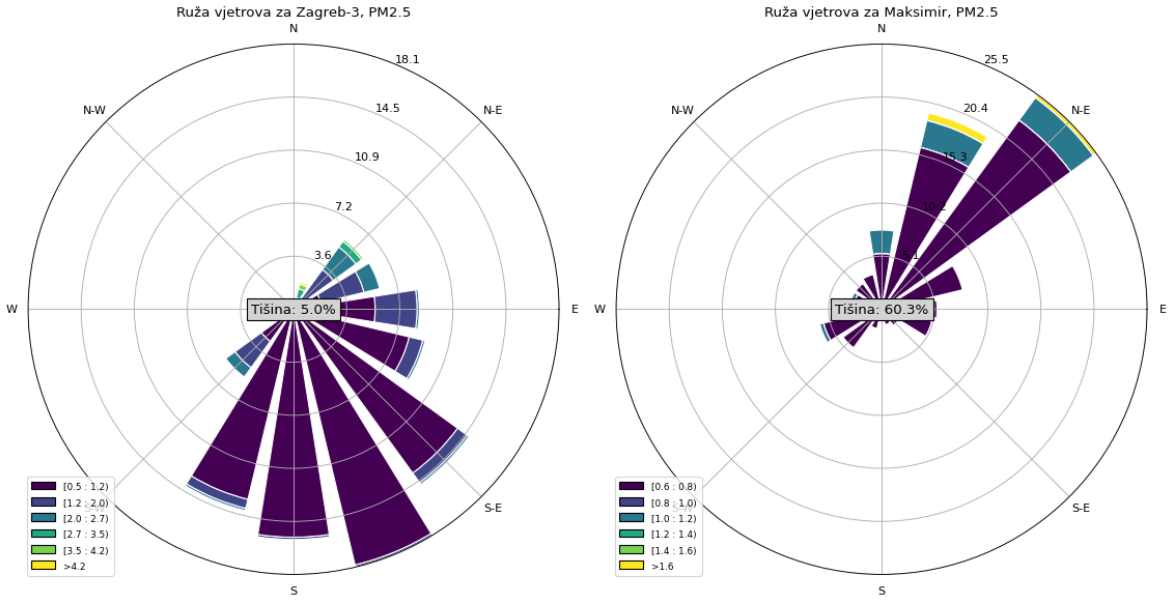
Tablica 12. Prikaz raspona stupnjeva i odgovarajućeg smjera vjetra.

Smjer vjetra	Stupnjevi	Smjer vjetra	Stupnjevi
N	0,00	**S**	180,00
NNE	22,50	**SSW**	202,50
NE	45,00	**SW**	225,00
ENE	67,50	**WSW**	247,50

E	90,00	**W**	270,00
ESE	112,50	**WNW**	292,50
SE	135,00	**NW**	315,00
SSE	157,50	**NNW**	337,50



Slika 10. *Dnevne učestalosti smjera vjetra [%] i dnevne vrijednosti brzine vjetra [m/s] po smjerovima označene u legendi, na sve 4 mjerne postaje tijekom pripadajućeg razdoblja mjerenja (tablica 10)*



Slika 11. *Dnevne učestalosti smjera vjetra [%] i dnevne vrijednosti brzine vjetra [m/s] po smjerovima označene u legendi, na mjernoj postaji Zagreb-3 i Zagreb-Maksimir tijekom pripadajućeg razdoblja mjerenja (tablica 11)*

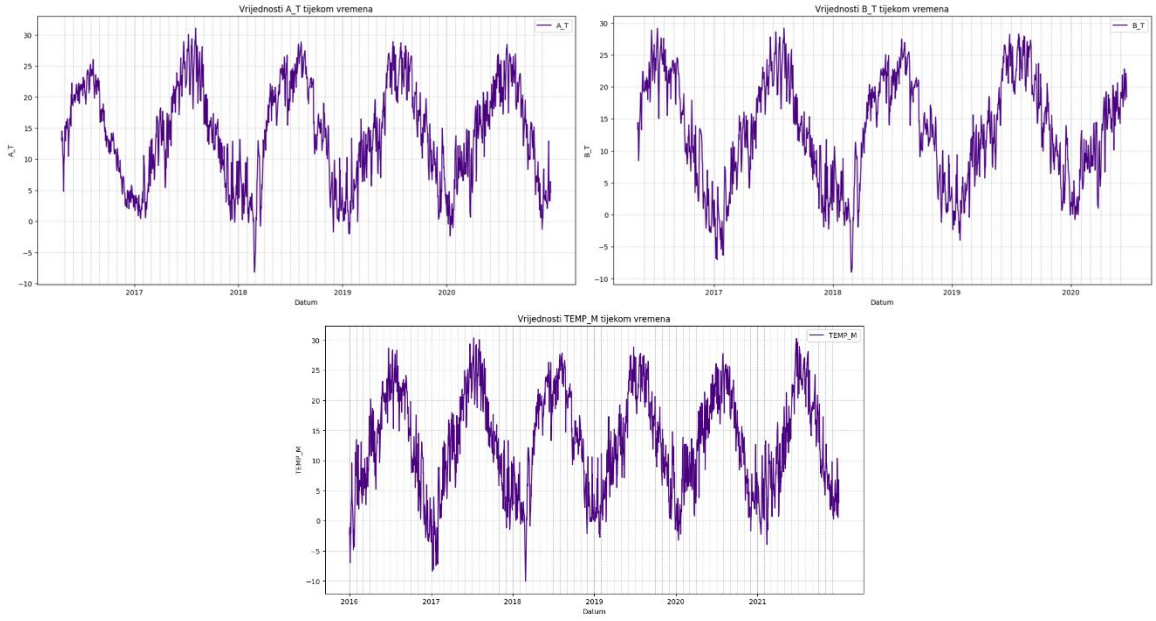
U tablici 13 prikazana je deskriptivna analiza za brzinu vjetra (ms^{-1}).

Tablica 13. Deskriptivna analiza za brzinu vjetra

Brzina vjetra (ms^{-1})						
Razdoblje	2016.-2021.	2016.-2021.			2022.-2024.	
Statistika	Maksimir	ZG1	ZG2	ZG3 (PM_{10})	Maksimir ($PM_{2,5}$)	ZG3 ($PM_{2,5}$)
broj	2192	1716	1501	1716	663	663
prosječna vrijednost	0,29	1,48	1,55	0,43	0,30	0,96
standardno odstupanje	0,29	0,52	0,84	0,49	0,29	0,55
minimalna vrijednost	0,00	0,40	0,13	0,00	0,00	0,30
25. percentil	0,00	1,15	1,00	0,00	0,08	0,60
medijan	0,00	1,40	1,40	0,30	0,08	0,80
75. percentil	1,00	1,70	1,80	0,70	0,58	1,10
maksimalna vrijednost	2,00	6,25	9,45	4,05	1,58	4,20

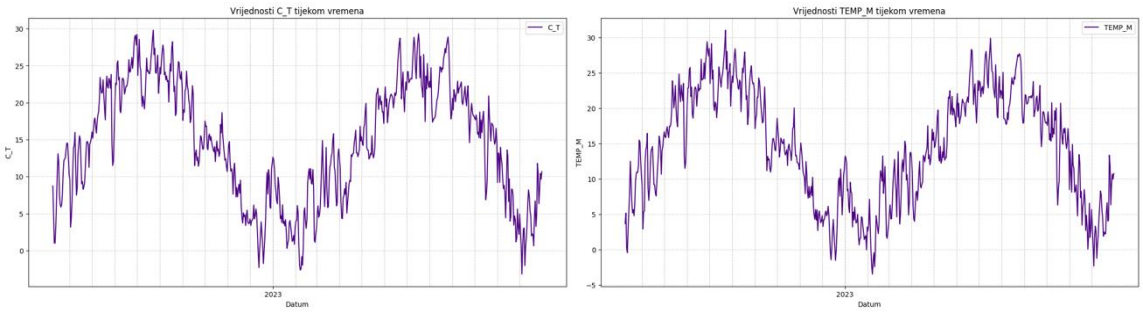
3.2.2. Temperatura zraka

Na slici 12 prikazano je kretanje temperature tijekom razdoblja mjerenja za lokacije Zagreb-1 (A_T), Zagreb-2 (B_T) i Maksimir (TEMP_M) korištene za predviđanje AQI (PM_{10}), a na slici 13 za lokaciju Zagreb-3 (C_T) i Maksimir (TEMP_M) korištene za predviđanje AQI ($PM_{2,5}$).



Slika 12. *Kretanje srednjih dnevnih temperatura [°C] na lokacijama a) ZG1, b) ZG2 i c) Maksimir u periodu 2016. – 2021. (točno razdoblje u tablici 10)*

Iz slike 12. vide se varijacije temperature karakteristične za klimu u gradu Zagrebu i ovom dijelu Hrvatske. Siječanj je općenito najhladniji mjesec, a srpanj najtopliji. 2018. godine najhladniji mjesec je veljača. Ljeti su najviše temperature zabilježene na lokaciji ZG1 u 2017. godini, a najniže temperature izmjerene su zimi na mjernoj postaji Maksimir 2017. i 2018. godine. Vrijednosti temperatura su približno jednake na lokacijama ZG1, ZG2, a kreću se u rasponu od -10 do 31 °C u razdoblju od 2016. do 2021. Na lokaciji Maksimir u razdoblju od 2016. do 2021. raspon temperatura se kreće između -10 i 31 °C. U tablici 14 prikazana je deskriptivna analiza za sve lokacije.



Slika 13. *Kretanje srednjih dnevnih temperatura [°C] na lokacijama a) ZG3, b) Maksimir u periodu 2022. – 2024. ($PM_{2,5}$)*

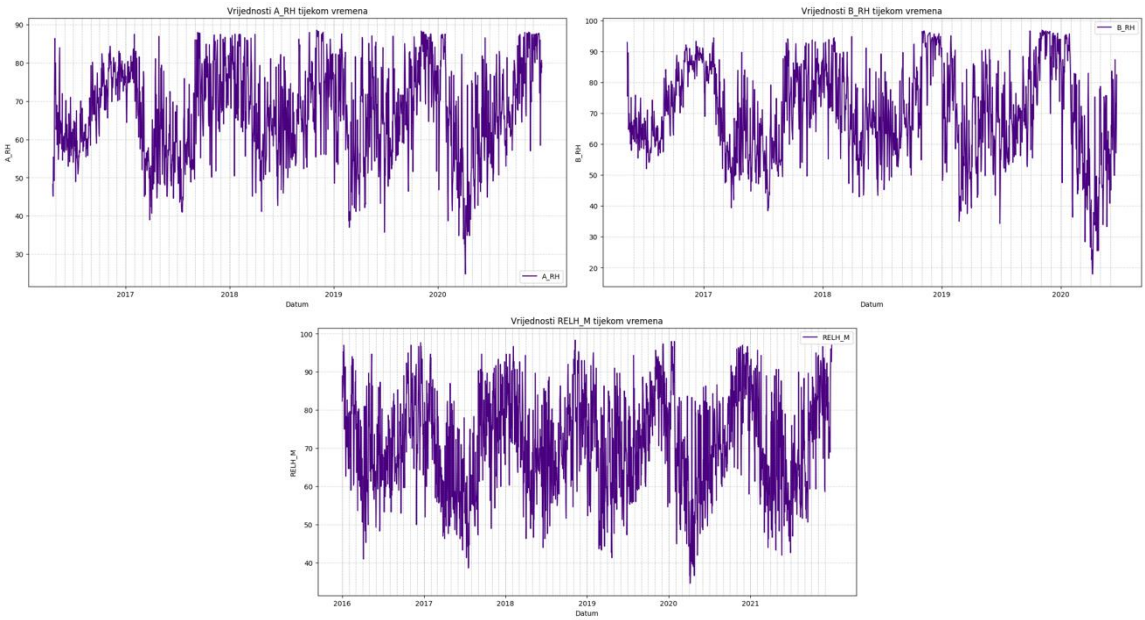
Kao na slici 12, na slici 13 vide se varijacije temperature karakteristične za klimu u gradu Zagrebu i ovom dijelu Hrvatske. Na lokaciji ZG3 srednje dnevne temperature se kreću između -4 i 30 °C, a na lokaciji Maksimir temperature se kreću između -4 i 30 °C za razdoblje od 2022. do 2024. Iz slike 13 vidi se da je veljača najhladniji mjesec, a srpanj najtopliji mjesec. U tablici 14 prikazana je deskriptivna analiza za sve lokacije.

Tablica 14. Deskriptivna analiza dnevnih vrijednosti temperatura na 4 lokacije, nakon predobrade.

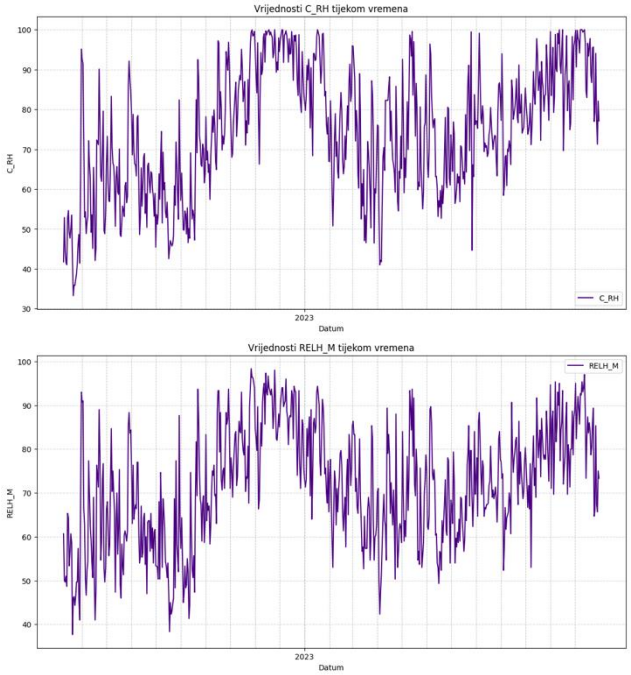
Temperatura (°C)					
Razdoblje	2016.-2021.	2016.-2021.		2022.-2024.	
Statistika	Maksimir	ZG1	ZG2	Maksimir ($PM_{2,5}$)	ZG3 ($PM_{2,5}$)
broj	2192	1716	1501	663	663
prosječna vrijednost	12,90	14,02	12,98	14,56	14,06
standardno odstupanje	8,40	7,97	8,35	7,84	7,85
minimalna vrijednost	-10,07	-8,16	-8.97	-3,47	-3,18
25. percentil	6,03	7,26	6,57	8,17	8,33
medijan	12,95	14,43	13,39	14,73	14,63
75. percentil	20,00	20,72	19,93	21,37	21,32
maksimalna vrijednost	30,37	31,07	29,18	31,07	29,80

3.2.3. Relativna vlažnost zraka

Relativna vlažnost, kao i temperatura se mjeri na iste 3 lokacije. Relativna vlažnost označena s A_RH mjerena je na postaji ZG1, B_RH na ZG2, a RELH_M na postaji Maksimir te s C_RH za ZG3 ($PM_{2,5}$). Grafovi promjene RH u vremenu prikazani su na slikama 14 i 15.



Slika 14. *Kretanje dnevnih prosjeka relativne vlažnosti zraka na lokacijama a) ZG1, b) ZG2 i c) Maksimir u periodu 2016. – 2021.*



Slika 15. *Kretanje dnevnih prosjeka relativne vlažnosti zraka na lokacijama a) ZG3, b)Maksimir u periodu 2022. – 2024. ($PM_{2,5}$)*

Promjene u vlažnosti su prikazane s puno varijabilnosti, što ukazuje na fluktuacije u dnevnoj relativnoj vlažnosti te pokazuju sezonske varijacije, kao i temperatura zraka, s višim vrijednostima tijekom zime i nižim tijekom ljeta. Iz slike 14 moguće je na lokaciji Maksimir očitati raspon RH od 34 % do 99 %, na lokaciji ZG1 vrijednosti se kreću između 24 % i 89

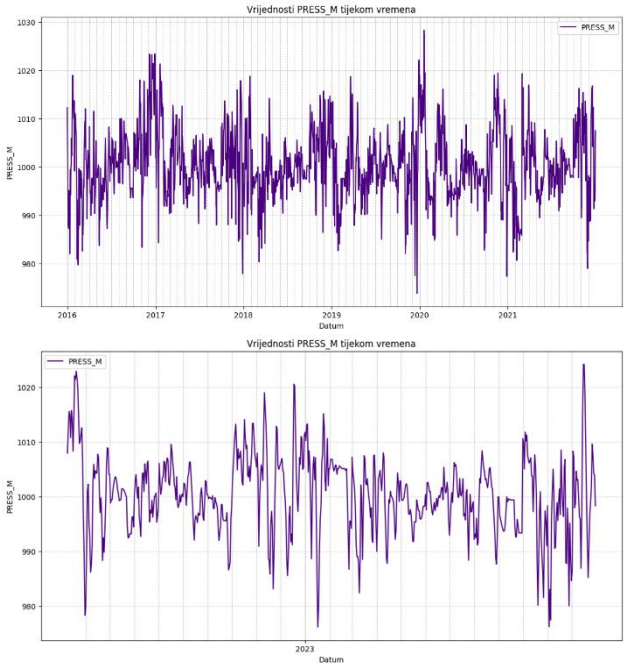
%, dok su na lokaciji ZG2 vrijednosti između 17 % i 97 %. Iz slike 15 vrijednosti RH na lokaciji ZG3 su u rasponu od 33 % do 100 %, a na lokaciji Maksimir vrijednosti se kreću od 37 % do 99 %. Na obje slike je jasno vidljivo da je samo nekoliko izuzetaka vrijednosti RH ispod 40 %, najčešće u travnju. U tablici 15 prikazana je deskriptivna analiza za sve lokacije.

Tablica 15. Deskriptivna analiza dnevnih vrijednosti RH na 4 lokacije, nakon predobrade.

Relativna vlažnost (%)					
Razdoblje	2016.-2021.	2016.-2021.		2021.-2024.	
Statistika	Maksimir	ZG1	ZG2	Maksimir ($PM_{2,5}$)	ZG3 ($PM_{2,5}$)
broj	2192	1716	1501	663	663
prosječna vrijednost	71,22	67,11	70,34	71,70	74,63
standardno odstupanje	12,74	12,12	15,15	13,47	15,94
minimalna vrijednost	34,67	24,79	17,94	37,67	33,28
25. percentil	61,33	57,87	59,53	62,17	62,78
medijan	71,00	67,18	70,00	71,67	75,36
75. percentil	81,33	76,80	83,10	82,67	87,48
maksimalna vrijednost	98,33	88,53	96,84	98,33	100,00

3.2.4. Tlak zraka

Tlak zraka mjeren je isključivo na postaji Maksimir. Promjena tlaka u vremenu prikazana je na slici 16.



Slika 16. *Kretanje dnevnih prosjeka tlaka zraka na lokaciji Maksimir u periodu a) 2016.-2021. (PM_{10}) i b) 2022.-2024. ($PM_{2,5}$)*

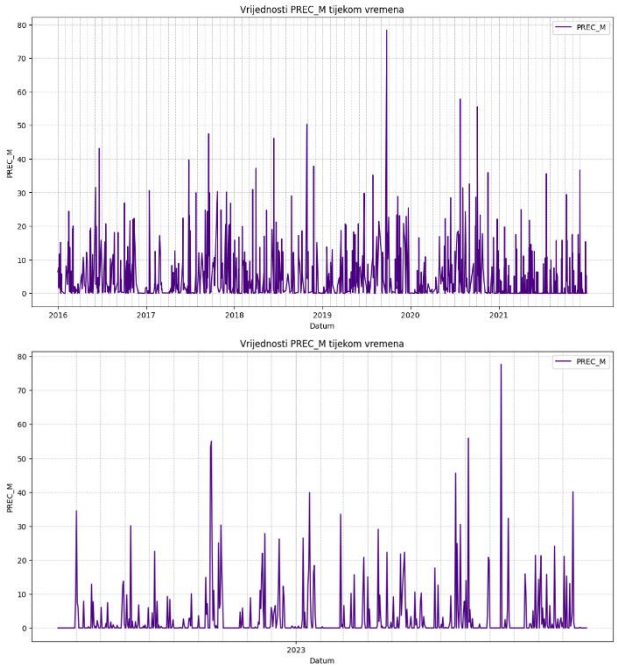
Tlak zraka je obično viši zimi, a niži tijekom ljetnih mjeseci što se može očitati na slici 16. Uočene vrlo niske i visoke vrijednosti tlaka mogu karakteristične za olujno vrijeme ili anticiklone, a takve visoke fluktuacije uočene su zimi. Iako postoje kratkoročne fluktuacije, ukupne vrijednosti tlaka se kreću unutar relativno stabilnog raspona od približno 960 hPa do 1030 hPa. Velike promjene tlaka uzrokuju povećanje onečišćenja PM_{10} (usporedba s slikom 18 i 19). Za razdoblje od 2016. do kraja 2020. srednja vrijednost tlaka iznosi 1000,38, a za razdoblje od ožujka 2022. do kraja 2023. srednja vrijednost tlaka iznosi 1000,64 hPa. U tablici 16 prikazana je deskriptivna analiza.

Tablica 16. Deskriptivna analiza dnevnih vrijednosti tlaka zraka na lokaciji Maksimir, nakon predobrade.

Tlak zraka (hPa)		
Razdoblje	2016.-2021.	2022.-2024.
Statistika	Maksimir	Maksimir
broj	2192	663
prosječna vrijednost	1000,38	1000,64
standardno odstupanje	7,29	6,48
minimalna vrijednost	973,83	976,23
25. percentil	995,95	996,57
medijan	999,74	1000,50
75. percentil	1004,60	1005,00
maksimalna vrijednost	1028,31	1022,93

3.2.5 Oborine

Količina oborina, kao i tlak zraka, mjerena je samo na lokaciji Maksimir. Označava se s PREC_M. Na slici 17 prikazana je promjena količine oborina u vremenu.



Slika 17. *Kretanje dnevnih prosjeka količina oborina na lokaciji Maksimir u periodu a) 2016.-2021. (PM_{10}) i b) 2022.-2024. ($PM_{2,5}$)*

Veća količina oborina karakteristična je za ljetne mjesece na području grada Zagreba, zbog ljetnih oluja. Vidi se kako je zimi prisutna najmanja količina oborina, dok su proljeće i jesen razdoblja s više padalina. Rujan je općenito mjesec s najvećom količinom oborina. Veću količinu oborina prati smanjenje onečišćenja u zraku (usporedba s slikom 18 i 19) i obrnuto. U tablici 17 prikazana je deskriptivna analiza.

Tablica 17. Deskriptivna analiza dnevnih vrijednosti količina oborina na lokaciji Maksimir, nakon predobrade.

Količina oborina (mm)		
Razdoblje	2016.-2021.	2022.-2024.
Statistika	Maksimir	Maksimir
broj	2192	663
prosječna vrijednost	4,48	2,99
standardno odstupanje	7,09	7,80
minimalna vrijednost	0,00	0,00
25. percentil	0,01	0,00
medijan	1,43	0,00
75. percentil	6,08	1,25
maksimalna vrijednost	78,30	77,6

3.3 Masene koncentracije PM_{10} i $PM_{2,5}$ i indeks kvalitete zraka

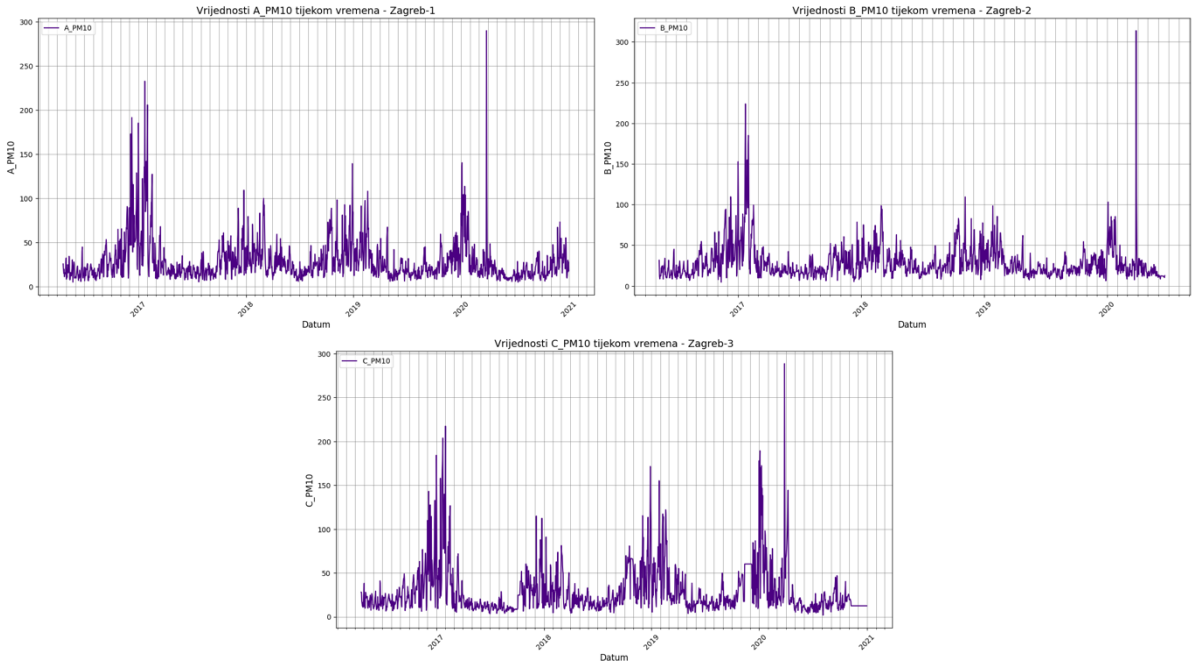
Koncentracije PM_{10} mjerene su kontinuirano na lokacijama ZG1, ZG2 i ZG3 u razdoblju od 2015. do 2020. godine, a prosjeci, izraženi u $\mu g/m^3$, pohranjivani su svaki sat. Mjerenja koncentracija $PM_{2,5}$ započela su na ovim mjernim postajama kasnije; stoga su u predobradu uzeti podaci na lokacijama ZG1, ZG2 u 2023 godini, a na ZG3 u razdoblju od 2020. do 2023. godine. Izmjerene koncentracije $PM_{2,5}$ također su automatski pohranjivane kao satni prosjeci izraženi u μgm^{-3}. Za mjerenje koncentracije PM korišteni je automatski analizator, ESM ANDERSEN FH-61 IR. Analizator daje satne koncentracije lebdećih čestica u realnom vremenu. Podaci s analizatora nakon toga prolaze proces validacije te zatim ekvivalencije, odnosno korekcije izmjerenih rezultata na pojedinom mjernom mjestu s obzirom na referentnu, gravimetrijsku metodu. Ova korekcija je nužna s obzirom na objektivne čimbenike poput sastava lebdećih čestica karakterističnog za to mjerno mjesto kao i s obzirom na čimbenike kao što su mjerni princip i tip uređaja, način i učestalost održavanja mjernog uređaja, utjecaj dizajna ulaza zraka u sakupljač (povišene radne temperature u uređaju uzrokuju gubitak hlapivih komponenti već sakupljenih čestica) i sl. Rezultati testova ekvivalencije omogućuju sezonske korekcije kao i godišnje korekcije mjernih rezultata te korekcije statističkih parametara za kategorizaciju kvalitete zraka (srednju godišnju vrijednost i broj prekoračenja dnevne granične vrijednosti od 50 μgm^{-3} tijekom jednogodišnjeg razdoblja). Ekvivalencije provode nacionalni referentni laboratoriji, u ovom slučaju Institut za medicinska istraživanja i medicinu rada, po uputama danim od strane Joint Resarch Centra i opisanim u dokumentu: "*EUR 23216 EN – 2008 Demonstration of Equivalence of Ambient Air Analytical Method*". Podaci koji su prošli proces ekvivalencije smatraju se referentnim podacima i mogu se koristiti u davanju ocjene kvalitete zraka. Podaci prikupljeni na mjernim postajama za $PM_{2,5}$, s obzirom da se radilo o novo uspostavljenim mjerenjima, nisu prošli proces validacije i korekcije, stoga se na temelju njih ne provodi ocjena kvalitete zraka sukladno Zakonu o zaštiti zraka. Posljedično, rezultati modela za predviđanje AQI kategorija za $PM_{2,5}$ trebaju se uzeti s određenom rezervom.

Prvi korak u predobradi podataka koncentracije PM_{10} je zamjena negativnih vrijednosti s *NaN* vrijednostima. Zatim se koristi metoda pomičnog prosjeka (engl. *rolling mean*) s prozorom od 300 uzastopnih vrijednosti za popunjavanje *NaN* vrijednosti. Ekstremne vrijednosti su uklonjene koristeći metodu interkvartilnog raspona (engl. *Interquartile Range*, IQR) jer ostale metode nisu omogućile uklanjanje odstupajućih vrijednosti. Kod IQR donja granica se izračunava kao Q1 minus 1,5 puta IQR, dok se gornja granica izračunava kao Q3

plus 1,5 puta IQR. Podaci koji padaju ispod donje granice ili iznad gornje granice smatraju se odstupanjima. Vrijednosti izvan definiranih granica smatraju se ekstremnim ili nepravilnim vrijednostima koje mogu iskriviti analizu podataka. One su zamijenjene s *NaN* vrijednostima, a *NaN* vrijednosti su nadalje popunjene novim podacima koristeći *bfill* i *ffill* linearnu interpolaciju.

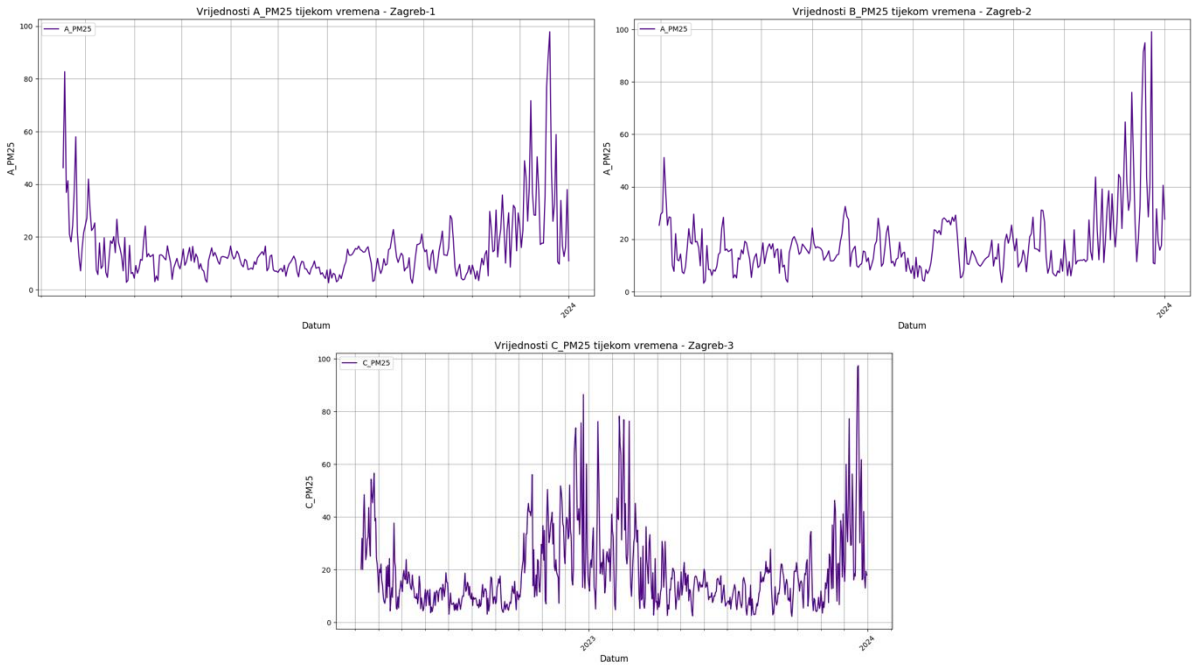
U podacima $PM_{2,5}$ upotrjebljena je samo *rolling-mean* funkcija za popunjavanje *NaN* vrijednosti, negativnih vrijednosti i ekstremnih vrijednosti nije bilo u podacima.

Iz tako predobrađenih podataka izračunate su dnevne vrijednosti putem medijana na temelju satnih podataka za koncentraciju PM_{10}. Rezultati predobrade prikazani su na grafovima na slici 18. Varijabla A_PM10 predstavlja dnevnu koncentraciju PM_{10} na lokaciji ZG1, B_PM10 predstavlja dnevnu koncentraciju PM_{10} na lokaciji ZG2, a C_PM10 predstavlja dnevnu koncentraciju PM_{10} na lokaciji ZG3.



Slika 18. *Kretanje koncentracija PM_{10} za lokacije a) ZG1, b) ZG2 i c) ZG3 za razdoblje 2016.-2021.*

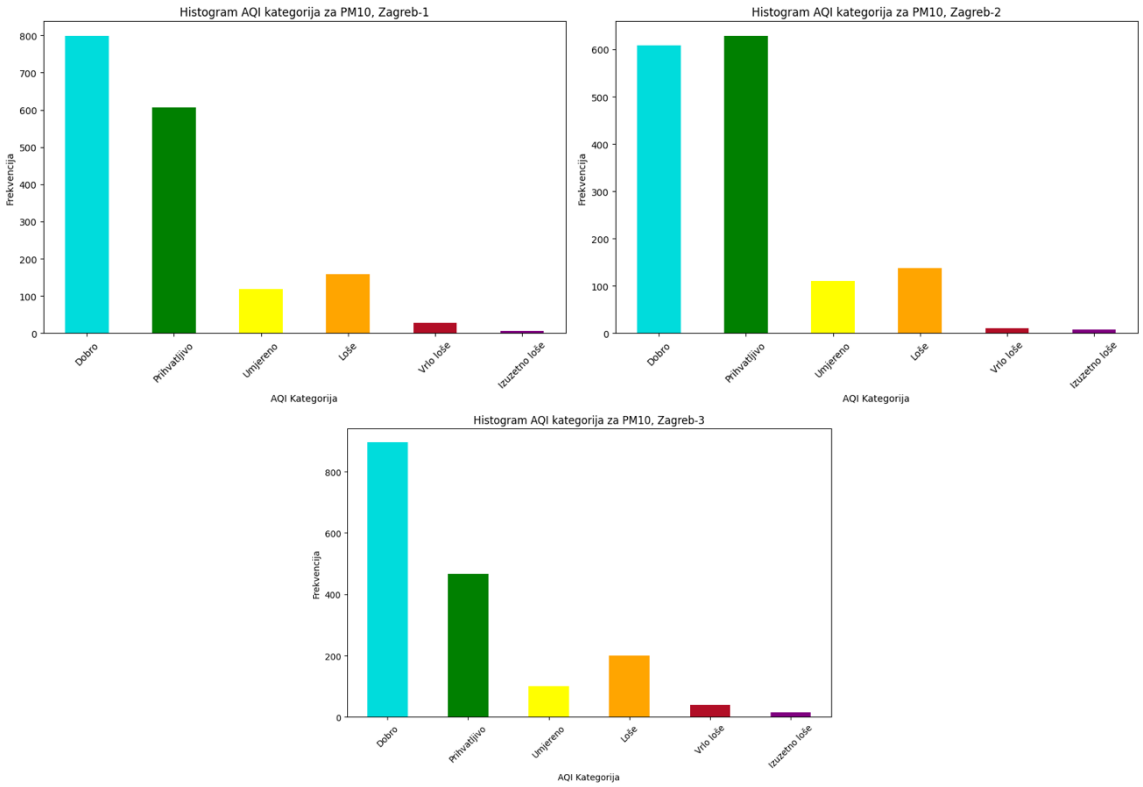
Na isti način izračunate su dnevne vrijednosti $PM_{2,5}$ koje su prikazane na grafovima na slici 19. A_PM25 predstavlja dnevnu koncentraciju $PM_{2,5}$ na lokaciji ZG1, B_PM25 predstavlja dnevnu koncentraciju $PM_{2,5}$ na lokaciji ZG2, a C_PM25 predstavlja dnevnu koncentraciju $PM_{2,5}$ na lokaciji ZG3.



Slika 19. *Kretanje koncentracija $PM_{2,5}$ na lokacijama, a) ZG1 za razdoblje od* veljače 2023 do kraja 2023.*, b) ZG2 za razdoblje od veljače 2023. do kraja 2023. i c) ZG3 za razdoblje od ožujka 2021. do kraja 2023.*

Na slikama 18 i 19. vidljiv je trend koji odgovara teorijskim navodima, zimi su veća onečišćenja lebdećim česticama, dok su ljeti znatno niža. Također, više vrijednosti RH i niže vrijednosti temperature zimi prati povećana koncentracija onečišćenja. Najviše koncentracije PM_{10} primjećuju se u od 24. do 28. ožujka 2020. godine, a uzrok čega je izuzetno jak doprinos pustinjskog pijeska. U prosjeku, najviše koncentracije PM_{10} izmjerene su tijekom zime 2016. i 2017. godine. Uočava se padajući trend koncentracija lebdećih čestica. U ožujku 2020. godine u Hrvatskoj je proglašen *lockdown* kao mjera suzbijanja širenja pandemije COVID-19, što je uključivalo zatvaranje škola, ugostiteljskih objekata, trgovina te ograničavanje kretanja građana, što je moglo utjecati na koncentraciju PM_{10} u tom razdoblju.

AQI kategorija se definira prema tablici 2 s obzirom na koncentraciju PM_{10} i $PM_{2,5}$. Histogrami AQI kategorija za PM_{10} prikazani su na slici 20.



Slika 20. *Raspodjela AQI kategorija za PM_{10} na lokacijama a) ZG1, b) ZG2 i c) ZG3 za razdoblje 2016.-2021.*

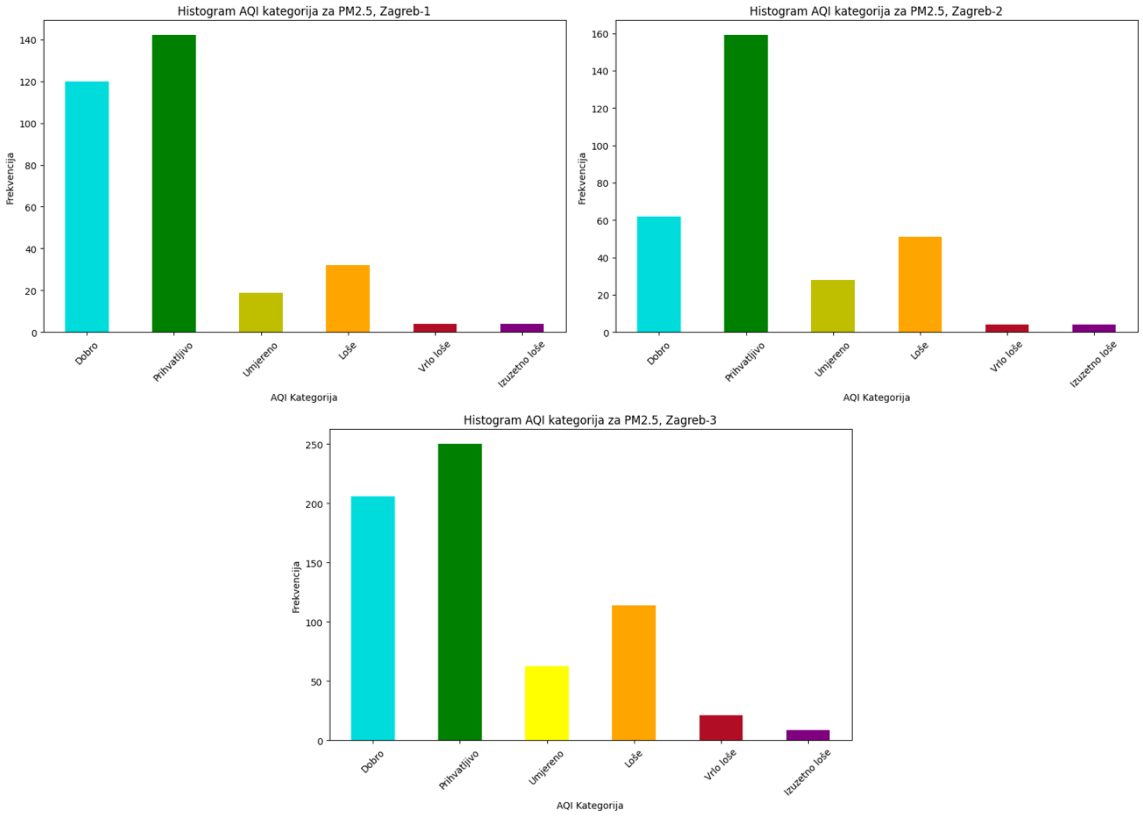
Na lokacijama ZG1 i ZG3 kategorija "Dobro" ima najvišu frekvenciju, slijedi kategorija "Prihvatljivo". Na lokaciji ZG2 kategorija "Prihvatljivo" ima višu frekvenciju od kategorije "Dobro". Na sve 3 lokacije kategorija "Umjereno" je manje zastupljena od kategorije "Loše". Kategorije "Vrlo loše" i "Izuzetno loše" značajno su manje zastupljene u svim skupovima podataka. Raspodjela kvalitete zraka pokazuje da većina podataka (više od 81,87% i 82,41% na lokacijama ZG1 i ZG2 te oko 79,32% na lokaciji ZG3) spada u pozitivnije kategorije ("Dobro" i "Prihvatljivo"), dok su ekstremne negativne kategorije ("Vrlo loše" i "Izuzetno loše") rijetko zastupljene. U tablici 18 prikazan je broj svih AQI kategorija na temelju PM_{10} za lokacije ZG1, ZG2 i ZG3.

Tablica 18. Broj AQI kategorija i odgovarajući postotak kategorije u skupu za lokacije Zagreb-1, Zagreb-2 i Zagreb-3 (za PM_{10}).

Kategorija	Zagreb-1		Zagreb-2		Zagreb-3	
	broj	postotak (%)	broj	postotak (%)	broj	postotak (%)
Dobro	798	46,50	609	40,57	895	52,16
Prihvatljivo	607	35,37	628	41,84	466	27,16

Umjereno	118	6,88	110	7,33	101	5,89
Loše	158	9,21	137	9,13	200	11,66
Vrlo loše	28	1,63	10	0,67	39	2,27
Izuzetno loše	7	0,41	7	0,47	15	0,87
Ukupno	1716		1501		1716	

Histogrami AQI kategorija za $PM_{2,5}$ prikazani su na slici 21.



Slika 21. *Raspodjela AQI kategorija za $PM_{2,5}$ na lokacijama, a) ZG1 za razdoblje od veljače* 2023 do kraja 2023.*, b) ZG2 za razdoblje od veljače 2023. do kraja 2023. i c) ZG3 za razdoblje od ožujka 2021. do kraja 2023.*

Kategorija "Prihvatljivo" ima najvišu frekvenciju na sve tri lokacije, ZG1, ZG2 i ZG3. Slijedi kategorija "Dobro", zatim "Loše" i "Umjereno", a kategorije "Vrlo loše" i "Izuzetno loše" pojavljuju se značajno manjom frekvencijom. Raspodjela kvalitete zraka pokazuje da većina podataka (za lokaciju ZG1 81,62 %, za ZG2 71,75 % i za ZG3 68,78 %) spada u pozitivnije kategorije ("Dobro" i "Prihvatljivo"), dok su ekstremne negativne kategorije ("Vrlo loše" i "Izuzetno loše") rijetko zastupljene (do 4%). U tablici 19 prikazan je broj svih AQI kategorija na temelju $PM_{2,5}$ za lokacije ZG1, ZG2 i ZG3.

Tablica 19. Broj AQI kategorija i odgovarajući postotak kategorije u skupu za lokacije Zagreb-1, Zagreb-2 i Zagreb-3 (za $PM_{2,5}$).

Kategorija	Zagreb-1		Zagreb-2		Zagreb-3	
	broj	postotak (%)	broj	postotak (%)	broj	postotak (%)
Dobro	142	44,24	159	51,62	206	31,07
Prihvatljivo	120	37,38	62	20,13	250	37,71
Umjereno	32	9,97	51	16,56	63	9,50
Loše	19	5,92	28	9,09	114	17,19
Vrlo loše	4	1,25	4	1,30	21	3,17
Izuzetno loše	4	1,25	4	1,30	9	1,36
Ukupno	321		308		663	

Dnevna AQI kategorija je kategorijska ciljana varijabla čije se vrijednosti žele predvidjeti modelima.

3.4. Inženjerstvo značajki

Kod inženjerstva značajki provedeno je nekoliko postupaka kako bi se podaci pripremili za treniranje modela. Sve kategorijske tekstualne varijable (AQI i smjer vjetra po kompasu) bilo je potrebno pretvoriti u numeričke. AQI se ručno označuje od kategorije "Dobro" do "Izuzetno loše" s vrijednostima od 0 do 5, a funkcijom *cat.codes* smjer vjetra od 0 do 16 prema stranama svijeta. Nadalje, u analizu se uključuju i temporalni podaci. Korištenjem funkcije *One-Hot Encoding*, temporalni podaci (datum) prevode se u binarni oblik (0 i 1). Na taj način generiraju se kolone za pojedini dan u tjednu (od ponedjeljka do nedjelje), mjesec (od siječnja do prosinca), godišnje doba (proljeće, ljeto, jesen, zima) i blagdane. U tim kolonama, broj 1 označava da se radi o konkretnom danu, mjesecu ili blagdanu, dok broj 0 označava da to nije slučaj. Ovaj pristup omogućuje modelima strojnog učenja da bolje razumiju i koriste temporalne informacije za analizu i predviđanje. U analizu su uključeni blagdani prikazani u tablici 20.

Tablica 20. Datumi blagdana u Republici Hrvatskoj.

Datum	Blagdan
01.01	Nova godina
06.01	Sv. tri kralja
pomični	Uskrs
pomični	Uskrsni ponedjeljak
01.05	Praznik rada
pomični	Tijelovo
22.06	Dan antifašističke borbe
05.08	Dan pobjede i domovinske zahvalnosti
15.08	Velika Gospa
08.10	Dan neovisnosti

01.11	Svi sveti
18.12	Dan sjećanja na žrtve Domovinskog rata
25.12	Božić
26.12	Sveti Stjepan

Određivane su korelacije između značajki i ciljane varijable, računanjem Pearsonovog koeficijenta korelacije. Na kraju je korištena funkcija *pd.merge* iz Pandas knjižice kako bi se spojili podaci prema zajedničkom ključu (indeksi *Dataframe*-a, podatkovnog okvira, koji sadrže datume), pojedinačno za lokacije ZG1, ZG2 i ZG3 s podacima iz lokacije Maksimir. Za potrebe strojnog učenja korištene su sve značajke. Prije treniranja modela one su se skalirale (normalizirale) korištenjem funkcije *MinMaxScaler*, kako bi značajke imale isti raspon vrijednosti, od 0 do 1.

3.5. Razvoj modela

Optimalni hiperparametri za modele određivani su funkcijom *GridSearchCV*. U tablici 21 prikazani su hiperparametri između kojih *GridSearchCV* pronalazi najbolje kombinacije. Korištena su dva modela te se usporedila njihova djelotvornost. To su već spomenuti SVM i RF klasifikacijski modeli. Modeli su opisani u poglavljima 2.4.1. i 2.4.2. Nakon određivanja optimalnih parametara model se trenira na skupu za obuku. Testni skup se koristi za konačnu procjenu performansi modela.

Tablica 21. Rasponi parametara modela.

		AQI PM_{10}	AQI $PM_{2,5}$
Model	**Parametar**	**Vrijednosti**	**Vrijednosti**
RF	*n_estimators*	100, 200, 300	100, 200, 300, 400, 500
	max_depth	None, 10, 20	None, 10, 20, 40
	min_samples_split	2, 5, 10	2, 5, 10, 20
	min_samples_leaf	1, 2, 4	2, 5, 10, 20
SVM	*C*	0,1, 1, 10	0,1, 1, 10
	Gamma	*scale, auto*, 0,001, 0,01	*scale, auto*, 0,001, 0,01
	Kernel	*linear, poly, rbf, sigmoid*	*linear, poly, rbf, sigmoid*

Za validaciju modela korišteni su kriterij vrednovanja modela za klasifikaciju (tablica 8) te makro-uprosječivanje. Uz to, izrađene su matrice konfuzije i grafovi uspoređenih stvarnih i predviđenih vrijednosti AQI.

4. EKSPERIMENTALNI DIO

Za 3 lokacije, ZG1, ZG2 i ZG3 korišten je isti algoritam optimizacije hiperparametara i razvoja modela. Krajnji podaci korišteni za strojno učenje prikazani su u tablici 22 i 23.

Tablica 22. Broj i vrsta varijabli korištenih za predviđanje AQI na temelju koncentracije PM_{10} na 3 lokacije (ZG1, ZG2 i ZG3).

Lokacije	Meteorološki podaci mjereni na postaji	Meteorološki podaci s postaje Maksimir	Temporalni podaci	Period	Broj podataka jedne varijable	Ukupni broj varijabli
ZG1	Smjer i brzina vjetra*, temperatura, relativna vlažnost zraka, koncentracija PM_{10}, AQI	Smjer i brzina vjetra*, temperatura, relativna vlažnost zraka, tlak zraka, količina oborina	Blagdani, godišnje doba, dan u tjednu, mjeseci, godina.	21.4.2016 – 31.12.2020.	1716	41
ZG2	Smjer i brzina vjetra*, temperatura, relativna vlažnost zraka, koncentracija PM_{10}, AQI			12.05.2016. – 20.06.2020.	1501	41
ZG3	Smjer i brzina vjetra*, koncentracija PM_{10}, AQI			21.04.2016 – 31.12.2020.	1716	39

* brzina vjetra po Beaufortovoj ljestvici i u ms^{-1}, smjer vjetra prema stranama svijeta i u stupnjevima

* AQI je ciljana varijabla, dok su ostale nezavisne

Tablica 23. Prikaz broja i vrsta varijabli korištenih za predviđanje AQI na temelju koncentracije $PM_{2,5}$ na lokaciji ZG3.

Lokacije	Meteorološki podaci mjereni na postaji	Meteorološki podaci s postaje Maksimir	Temporalni podaci	Period	Broj podataka jedne varijable	Ukupni broj varijabli
ZG3	Smjer i brzina vjetra*, temperatura, relativna vlažnost zraka, koncentracija $PM_{2,5,}$ AQI	Smjer i brzina vjetra*, temperatura, relativna vlažnost zraka, tlak zraka, količina oborina	Blagdani, godišnje doba, dan u tjednu, mjeseci, godina.	09.03.2021. – 31.12.2023.	663	41

* brzina vjetra po Beaufortovoj ljestvici i u ms^{-1}, smjer vjetra prema stranama svijeta i u stupnjevima

* AQI je ciljana varijabla, dok su ostale nezavisne

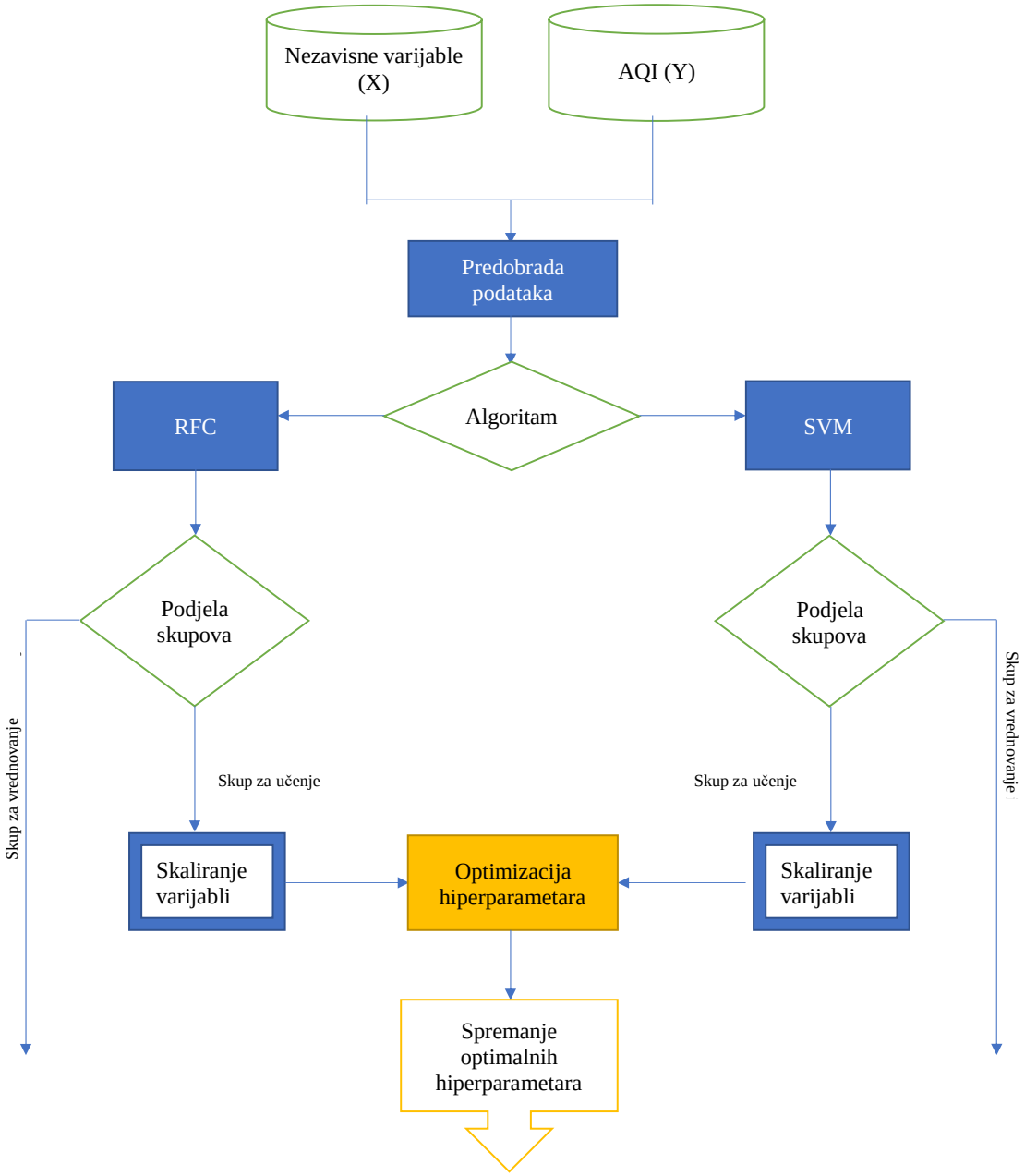
4.1. Određivanje optimalnih hiperparametara i razvoj modela

Optimalni hiperparametri SVM i RF modela određeni su *GridSearchCV* algoritmom.

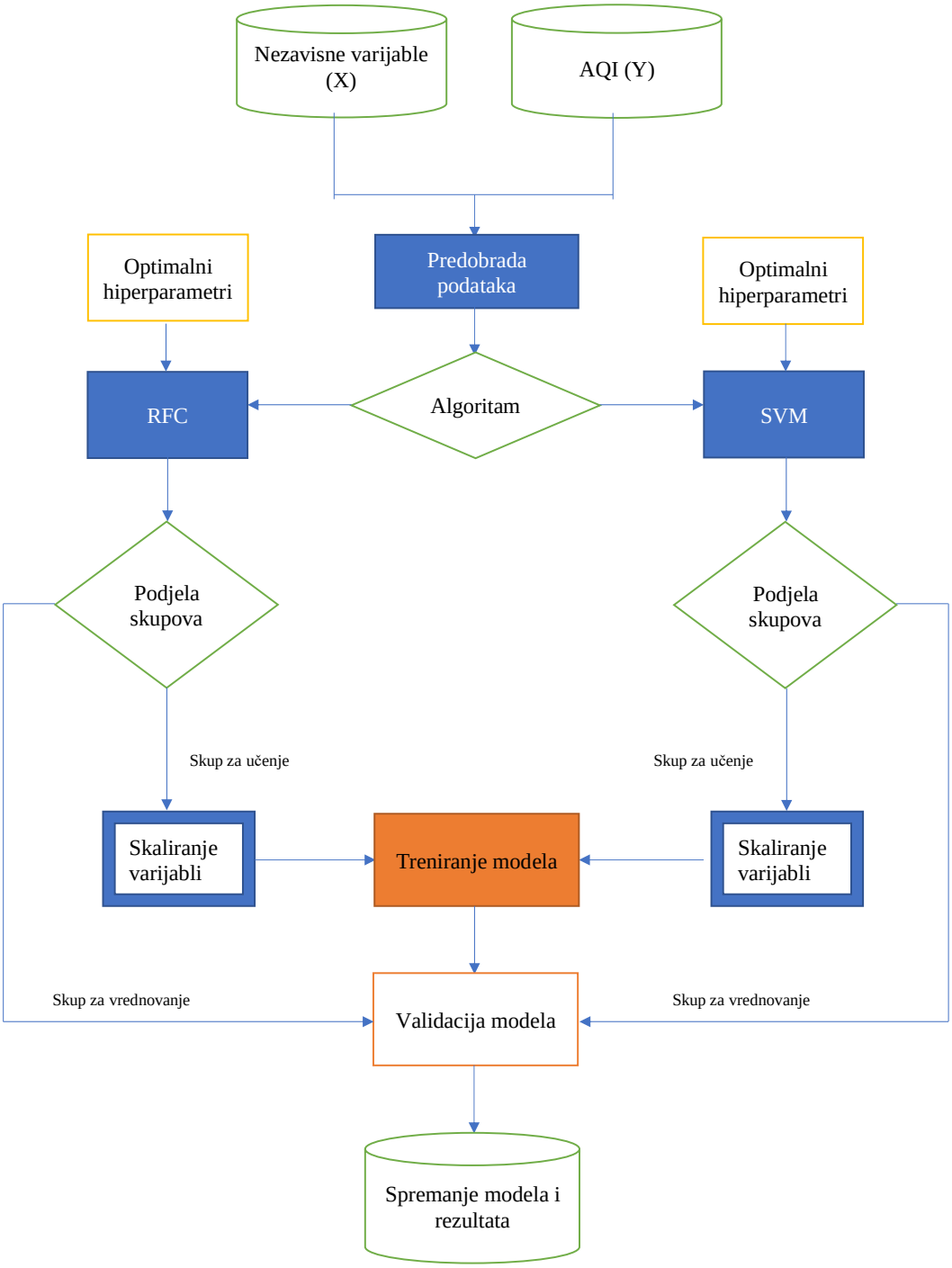
Prvi korak je definiranje ciljane varijable AQI (*y*) i ostalih značajki, odnosno nezavisnih varijabli (*X*). Zatim se provodi skaliranje značajki *MinMaxScaler* naredbom.

Na slici 22 prikazana je shema postupka određivanja hiperparametara. Podaci se dijele u 2 seta, za treniranje i za testiranje, pomoću *train_test_split* funkcije. Cijeli skup podataka podijeljen je na trening (80 %) i test (20 %) skup.

Za potrebe pronalaženja optimalnih parametara modela podaci iz skupa za obuku se *k-fold* unakrsnom validacijom dijele na 5 jednakih skupova. Svaki skup sadrži drugačije pomiješane podatke kako bi se osigurala robusnost i izbjeglo pretreniranje modela. Unakrsna validacija se provodi 5 puta, pri čemu svaki put jedan od 5 skupova služi kao skup za validaciju dok ostalih 4 služe za obuku modela. Optimalni parametri služe za formiranje modela za predviđanje AQI PM_{10}. Na slici 23 prikazana je shema razvoja obje vrste modela.



Slika 22. *Algoritam odabira hiperparametara modela*



Slika 23. *Algoritam treniranja modela*

5. REZULTATI I RASPRAVA

5.1. Pearsonov koeficijent korelacije

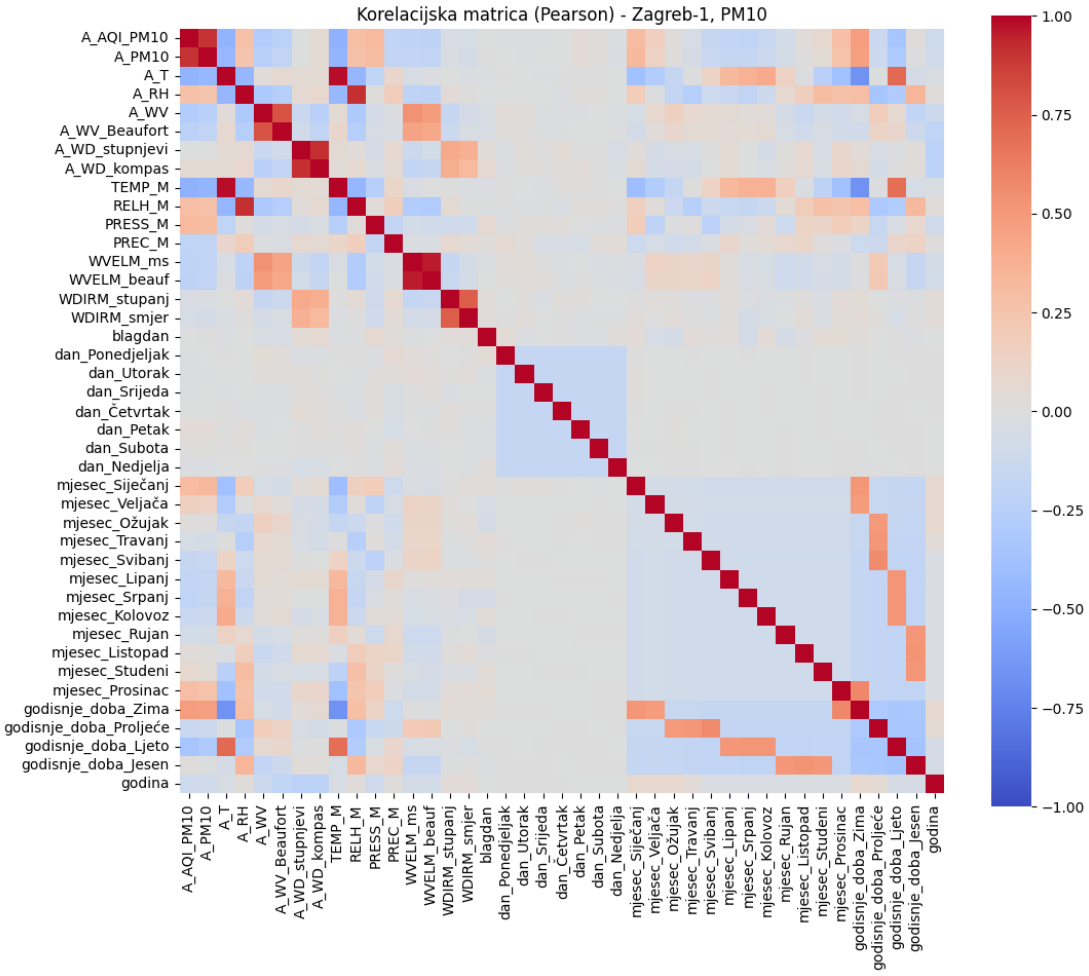
Grafovi na slikama pokazuju Pearsonovu korelaciju između varijabli, a u daljnjem tekstu opisana je korelacija varijabli s ciljanom varijablom AQI za tri različite lokacije (ZG1, ZG2 i ZG3).

Pearsonova korelacija prikazuje linearne odnose između varijabli. Koeficijent korelacije može biti pozitivan (crvena boja, vrijednost 1, povećanje jedne varijable uzrokuje povećanje druge) ili negativan (plava boja, vrijednost -1, povećanje jedne varijable uzrokuje smanjenje druge).[101] Što je intenzitet boje jači to je jača korelacija. Pregled značenja vrijednosti koeficijenta dan je u tablici 24.

Tablica 24. Vrijednost Pearsonovog koeficijenta korelacije i njegovo značenje.

Pearsonov koeficijent korelacije (r)	**Snaga**	**Smjer**
$r > 0,5$	Snažna	Pozitivna
$0,3 < r < 0,5$	Umjerena	Pozitivna
$0 < r < 0,3$	Slaba	Pozitivna
$r = 0$	Nema	Nema
$0 > r > -0,3$	Slaba	Negativna
$-0,3 > r > -0,5$	Umjerena	Negativna
$r < 0,5$	Snažna	Negativna

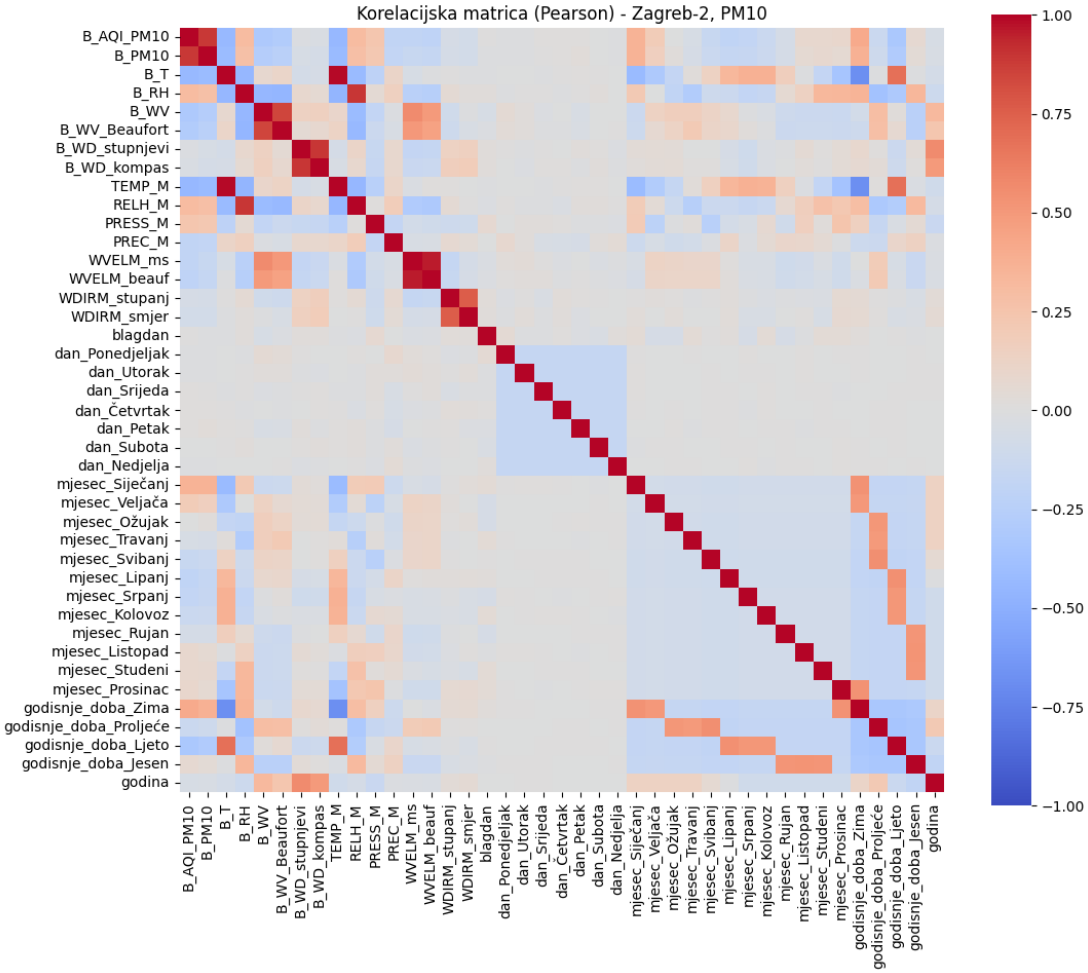
Važno je napomenuti da nadalje opisane korelacije ne znače nužno uzročno-posljedičnu vezu između varijabli. Na povećanje koncentracije PM utječu karakteristične meteorološke situacije, gustoća prometa u određenom razdoblju te spomenuta pojava temperaturne inverzije kao i korištenje izvora grijanja tijekom zimskih mjeseci.



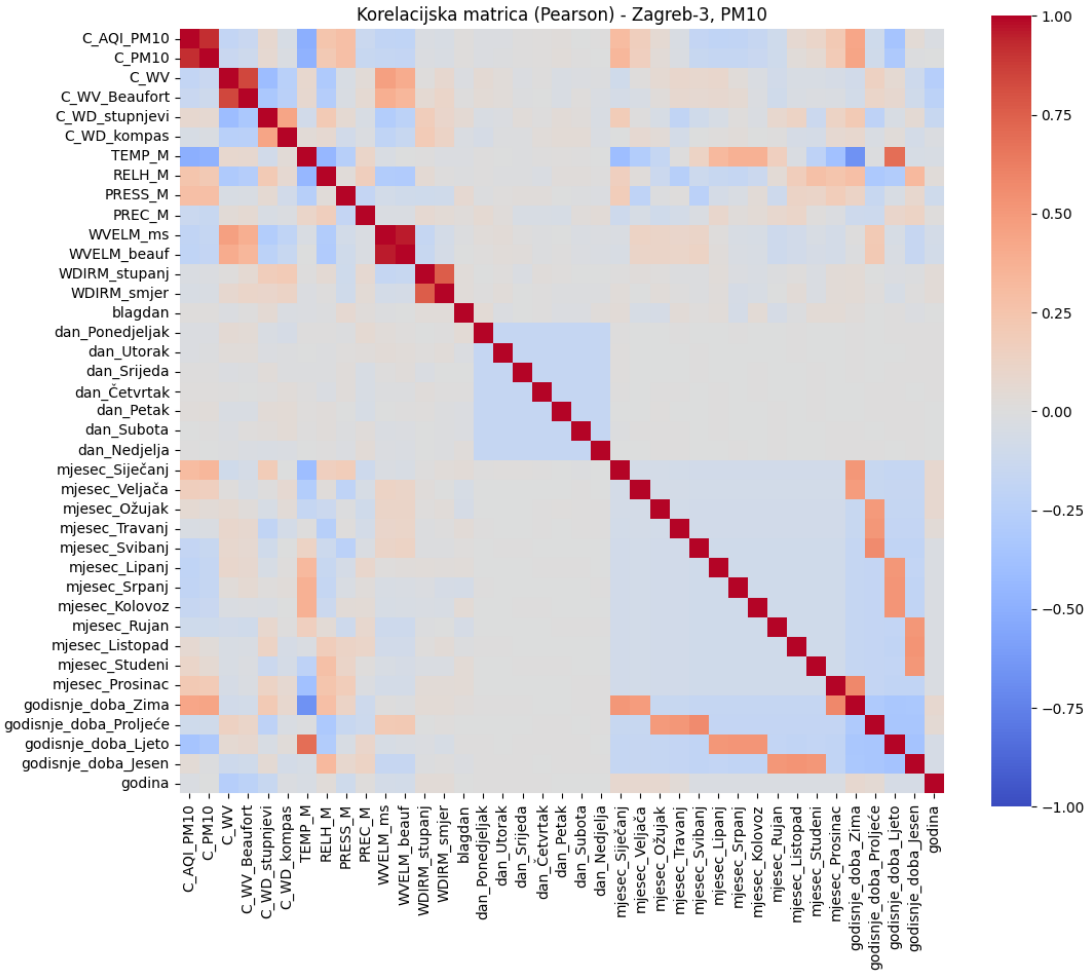
Slika 24. *Pearsonove korelacijske matrice varijabli za lokaciju Zagreb-1*

Na slici 24 prikazane su grafički vrijednosti Pearsonove korelacije između različitih varijabli na lokaciji Zagreb-1 koristeći toplinsku mapu (engl. *heathmap*). Koncentracija PM_{10} pokazuje jaku pozitivnu korelaciju s AQI, što je očekivano jer PM_{10} čestice direktno utječu na računanje indeksa kvalitete zraka. Temperatura općenito ima umjereno negativnu korelaciju, što znači da veće temperature neizravno smanjuju koncentracije PM_{10} čestica i poboljšavaju kvalitetu zraka jer se ljeti ne koristi grijanje i ne dolazi do temperaturne inverzije. Dakle, viša temperatura često znači bolju disperziju onečišćenja zbog pojačanog kretanja zraka. Smjer vjetra (u stupnjevima i stranama svijeta) na lokaciji Maksimir pokazuje vrlo slabu negativnu korelaciju, što bi trebalo biti intuitivno jer se koncentracija onečišćenja mjeri na drugoj lokaciji, no smjer vjetra mjeren na lokaciji Zagreb-1 također pokazuje vrlo slabu negativnu korelaciju, stoga se može zaključiti da smjer vjetra nema značajan utjecaj na indeks kvalitete zraka. S druge strane, brzina vjetra utječe na kvalitetu zraka, što dokazuje negativna korelacija s koncentracijom PM. Dakle vjetrovi pomažu u raspršivanju PM_{10} čestica ovisno o jačini, a

manje o smjeru. Atmosferski tlak pokazuje slabu pozitivnu korelaciju s koncentracijama PM_{10}, što potvrđuje da se u stabilnim vremenskim uvjetima povećava koncentracija čestica. Relativna vlažnost zraka mjerena na lokaciji Maksimir i ZG1 pokazuju slabo pozitivnu korelaciju. Oborine imaju vrlo slabu negativnu korelaciju, što bi značilo da u određenoj mjeri utječu na uklanjanje PM_{10} iz zraka. Na slici se prema intenzitetu boje može zaključiti da najvišu korelaciju s PM_{10} i AQI imaju meteorološke varijable mjerene na lokaciji mjerenja koncentracije PM_{10}, vrlo slična korelacija vidljiva je s meteorološkim varijablama mjerenim na postaji Maksimir. Sukladno meteorološkim uvjetima godišnja doba i mjeseci također pokazuju korelaciju s AQI, dok dani u tjednu i blagdani nisu pokazali bitan utjecaj na AQI. Vrlo slični rezultati su dobiveni za lokacije Zagreb-2 i Zagreb-3. Pearsonove korelacijske matrice za preostale dvije lokacije prikazane su na slikama 25 i 26.

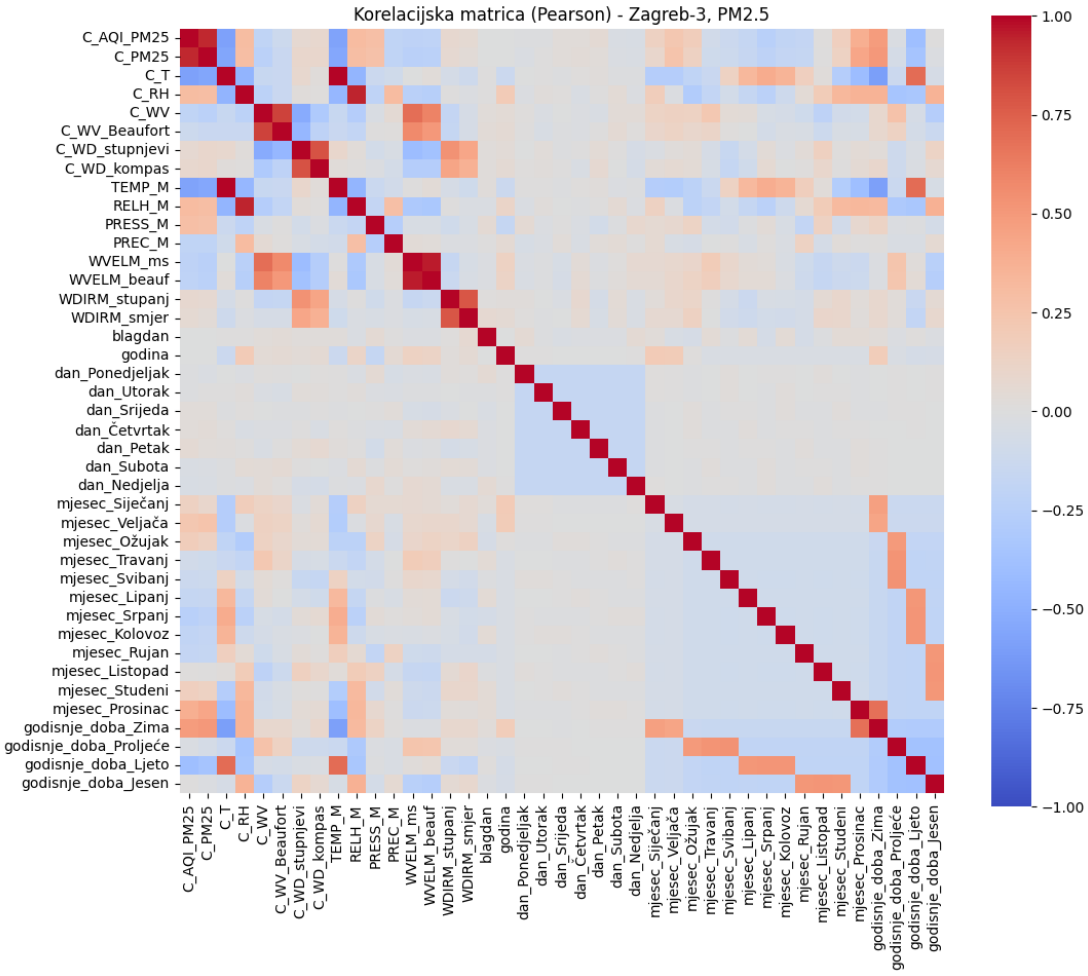


Slika 25. *Pearsonove korelacijske matrice varijabli za lokaciju Zagreb-2*



Slika 26. *Pearsonove korelacijske matrice varijabli za lokaciju Zagreb-3*

Na slici 27 se prema intenzitetu boje može zaključiti da najvišu negativnu korelaciju s $PM_{2,5}$ i AQI imaju meteorološke varijable temperatura i brzina vjetra mjerene na lokaciji mjerenja koncentracije $PM_{2,5}$ (ZG3), neznatno manja negativna korelacija je uočena s istim varijablama na lokaciji Maksimir. Oborine također pokazuju negativnu korelaciju s AQI. Relativna vlažnost i tlak pokazuju blago pozitivnu korelaciju s $PM_{2,5}$ i AQI na obje lokacije. Smjer vjetra nema značajan utjecaj na AQI. Godišnja doba i mjeseci također pokazuju korelaciju s AQI, dok dani u tjednu i blagdani nisu pokazali bitan utjecaj na AQI.



Slika 27. *Pearsonove korelacijske matrice varijabli za lokaciju Zagreb-3 ($PM_{2,5}$)*

5.2. Optimalni hiperparametri modela

Optimalni hiperparametri korišteni za izradu modela, dobiveni funkcijom *GridSearchCv*, prikazani su u tablici 25.

Tablica 25. Određeni optimalni parametri za modele RF i SVM za sve 3 lokacije.

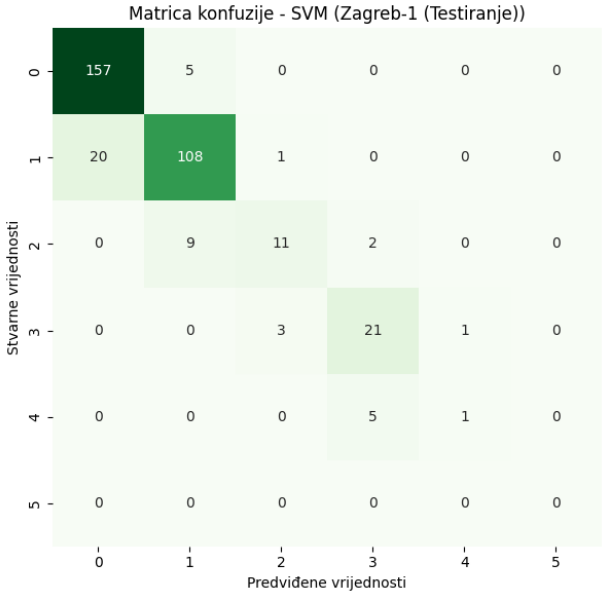
		PM_{10}			**$PM_{2,5}$**
Model	Parametar	Zagreb-1	Zagreb-2	Zagreb-3	Zagreb-3
RF	*n_estimators*	100	300	100	500
	max_depth	20	None	None	None
	min_samples_split	2	2	5	5
	min_samples_leaf	1	1	1	1
SVM	*C*	10			
	gamma	*scale*			
	kernel	*linear*			

5.3. Rezultati i validacija modela

Izrađeni su SVM i RF klasifikacijski modeli s parametrima prikazanim u tablici 25. Rezultati predviđanja i vrednovanja modela za sve 3 lokacije prikazani su u sljedećim poglavljima te je na kraju uspoređena uspješnost dvaju modela. Modeli su trenirani i validirani s podacima prikazanim u tablicama 22 i 23.

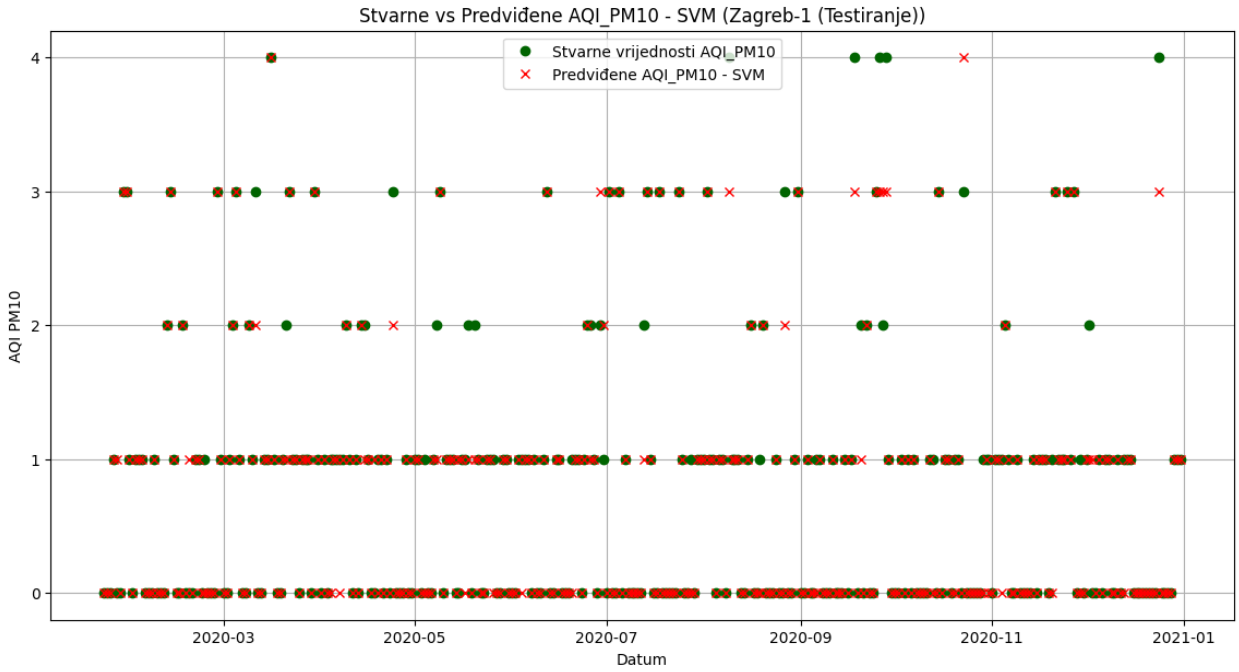
5.3.1. SVM algoritam

Prikazani su rezultati predikcije AQI na temelju koncentracije PM_{10}, SVM modelom na testnom skupu podataka. Slike 28, 30 i 32 prikazuju konfuzijske matrice ispravno i pogrešno predviđenih podataka. Grafovi na slikama 29, 31 i 33 prikazuju usporedbu stvarnih vrijednosti AQI PM_{10} s predviđenim vrijednostima koristeći SVM model na lokaciji ZG1, ZG2 (od siječnja do prosinca 2020.) i ZG3 (od kolovoza 2019. do lipnja 2020. godine) za testni skup podataka. Zelene točke predstavljaju stvarne vrijednosti AQI, a crveni križići predstavljaju predviđene vrijednosti AQI od strane SVM modela.

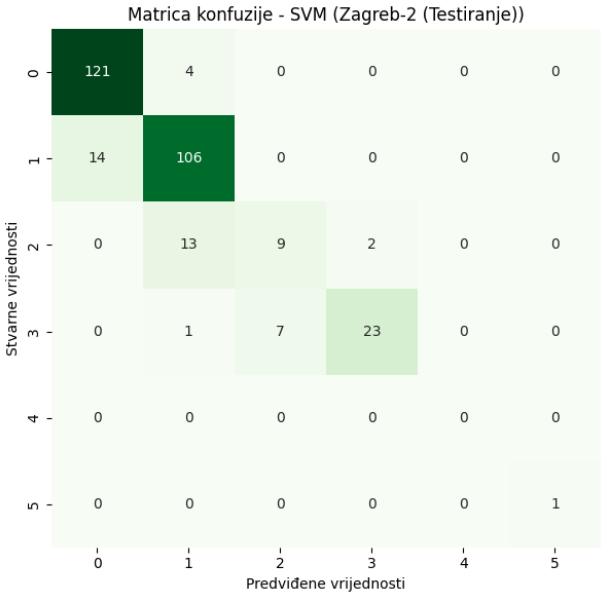


Slika 28. *Matrice konfuzije za testni skup podataka za lokaciju Zagreb-1*

Matrica konfuzije za testni skup podataka prikazuje da je model točno predvidio 157 slučaja za klasu 0, dok je 5 slučaja pogrešno klasificirano. Za klasu 1 točno je predviđeno 108 slučaja, a krivo 21. Za klasu 2, model je točno predvidio 11 slučaja, dok je 11 puta došlo do pogreške. Klasa 3 je točno predviđena 21 put, dok je 4 puta pogrešno klasificirana. Klasa 4 je točno predviđena samo jedanput, a pogrešno 5 puta. U skupu ne postoje podaci za klasu 5. Na slici 29 rezultati su prikazani grafički, korišten je skup podataka od siječnja do prosinca 2020. godine.

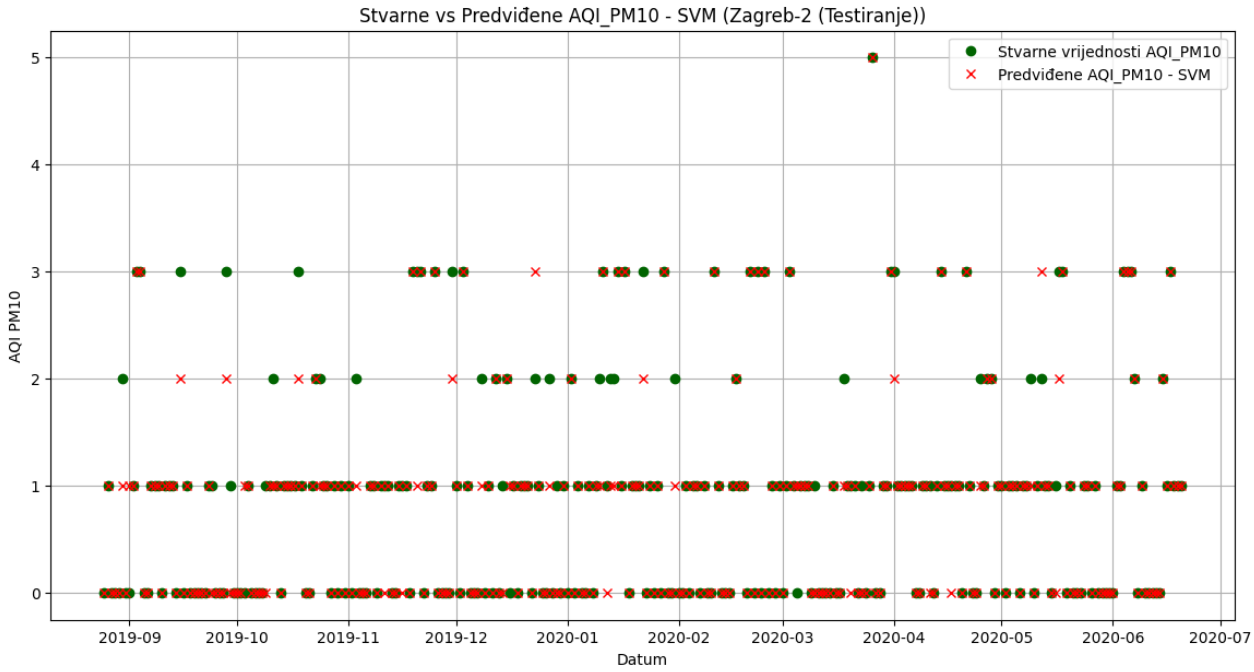


Slika 29. *Prikaz stvarnih vrijednosti i modelom predviđenih vrijednosti AQI na lokaciji Zagreb-1 za testni skup podataka za razdoblje od siječnja do prosinca 2020.*

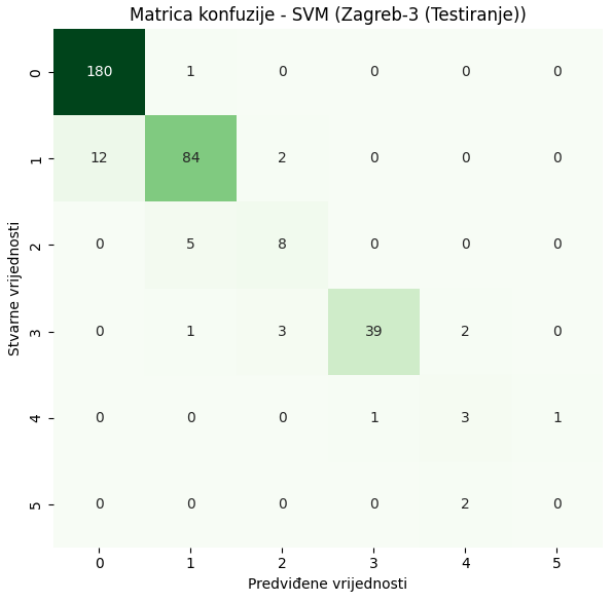


Slika 30. *Matrice konfuzije za testni skup podataka za lokaciju Zagreb-2*

Matrica konfuzije za testni skup podataka prikazuje da je model točno predvidio 121 od 125 slučaja za klasu 0. Za klasu 1, model je točno predvidio 106 od 120 slučaja. Za klasu 2, model je točno predvidio 9 slučaja, dok je 15 puta došlo do pogreške u predviđanju. Klasa 3 je točno predviđena 23 puta, dok je 8 puta pogrešno klasificirana. U skupu ne postoje podaci za klasu 4. Klasa 5 je točno predviđena za jedan postojeći podatak u skupu. Na slici 31 rezultati su prikazani grafički, korišten je testni skup za razdoblje od kolovoza 2019. godine do lipnja 2020. godine.

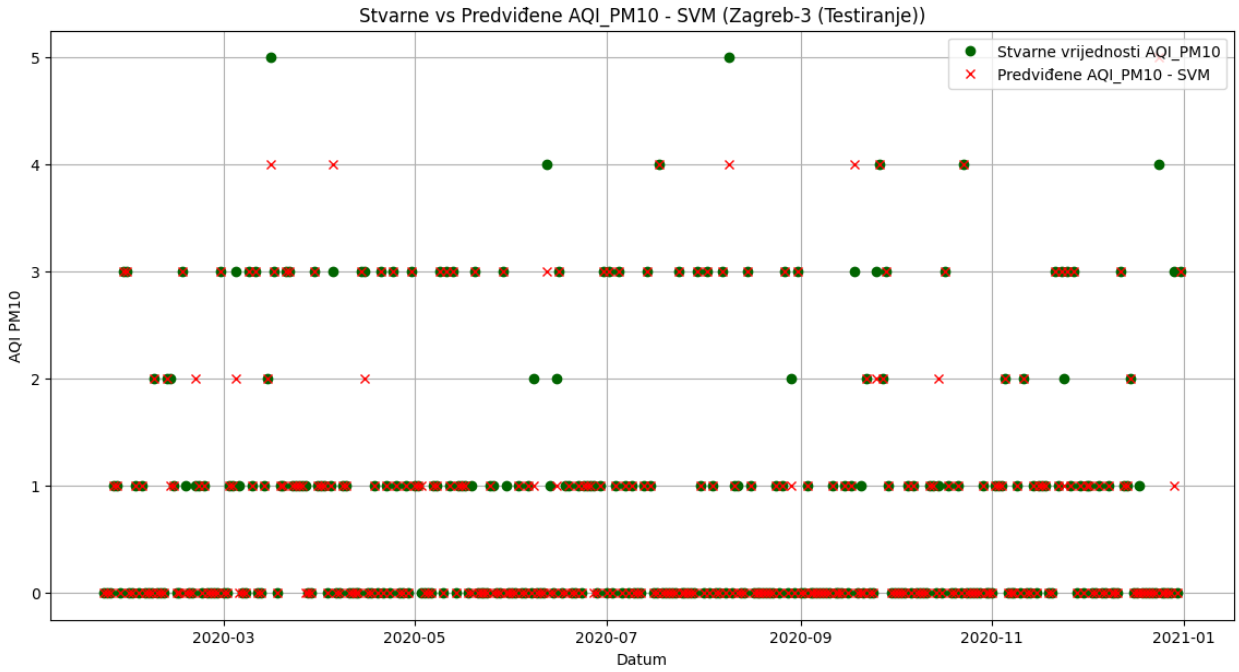


Slika 31. *Prikaz stvarnih vrijednosti i modelom predviđenih vrijednosti AQI na lokaciji Zagreb-2 za testni skup podataka za razdoblje od kolovoza 2019. do lipnja 2020. godine*



Slika 32. *Matrice konfuzije za testni skup podataka za lokaciju Zagreb-3*

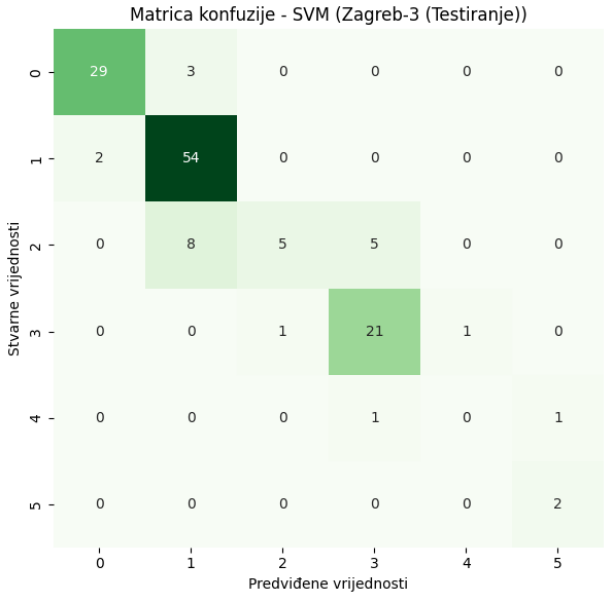
Za testni skup podataka, model je točno predvidio 180 od 181 slučaja za klasu 0. Za klasu 1, model je točno predvidio 84 slučaja, a 14 slučaja krivo. Za klasu 2, model je točno predvidio 8 slučaja od ukupnih 13. Klasa 3 je točno predviđena 39 od 45 puta. Model je točno predvidio 3 od 5 slučaja klase 4. Klasa 5 je krivo predviđena u oba slučaja. Rezultati su prikazani na slici 33, korišten je testni skup od siječnja do prosinca 2020. godine.



Slika 33. *Prikaz stvarnih vrijednosti i modelom predviđenih vrijednosti AQI na lokaciji Zagreb-3 za testni skup podataka za razdoblje od siječnja do prosinca 2021. godine*

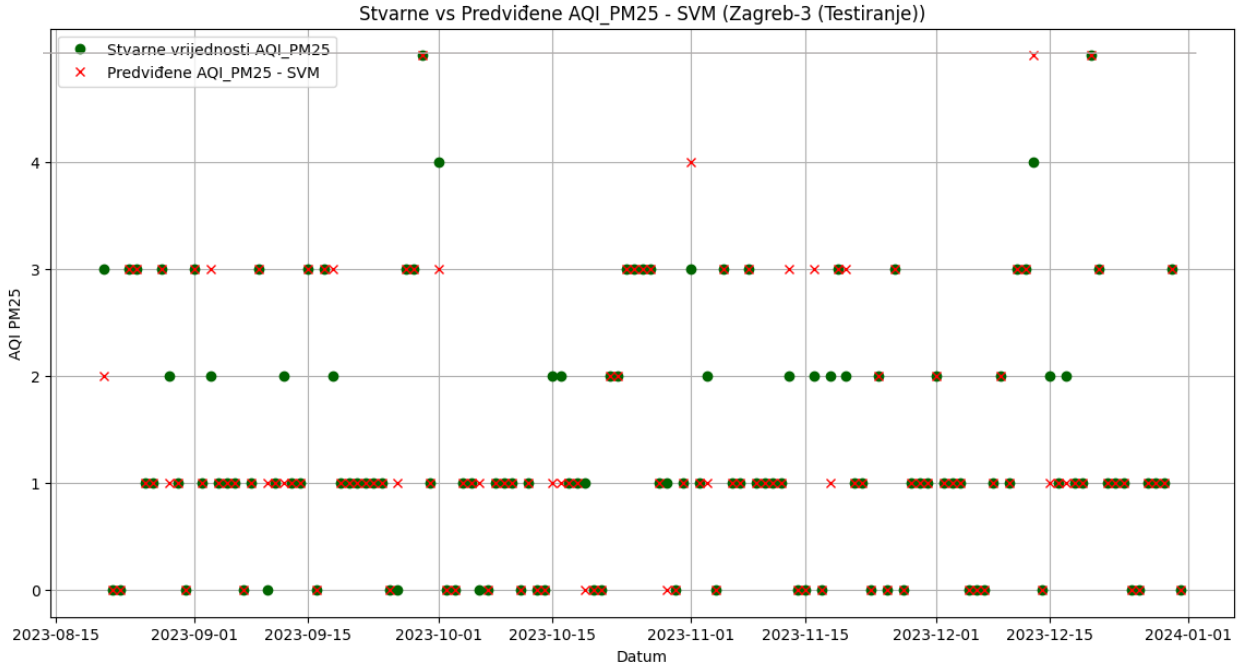
Iz rezultata se vidi da za klase 0 i 1, koje sadrže najviše podataka, model pokazuje visoku točnost u predikciji, što ukazuje na pouzdanost modela za ove specifične klase. Za klasu 3, koja je treća najzastupljenija klasa, dolazi do više pogrešaka u predviđanju. SVM model pokazuje poteškoće u predviđanju rijetkih (2) i vrlo rijetkih klasa (4 i 5), što ukazuje na potrebu za daljnjom optimizacijom ili balansiranjem podataka.

Također su prikazani rezultati predikcije AQI na temelju koncentracije $PM_{2,5}$, SVM modelom na testnom skupu. Slika 31 prikazuje konfuzijsku matricu ispravno i pogrešno predviđenih podataka. Graf na slici 31 prikazuje usporedbu stvarnih vrijednosti AQI na temelju $PM_{2,5}$ sa predviđenim vrijednostima, koristeći SVM model na lokaciji ZG3 za testni skup podataka. Zelene točke predstavljaju stvarne vrijednosti AQI, a crveni križići predstavljaju predviđene vrijednosti AQI od strane SVM modela.



Slika 34. *Matrice konfuzije za testni skup podataka za lokaciju Zagreb-3 ($PM_{2,5}$)*

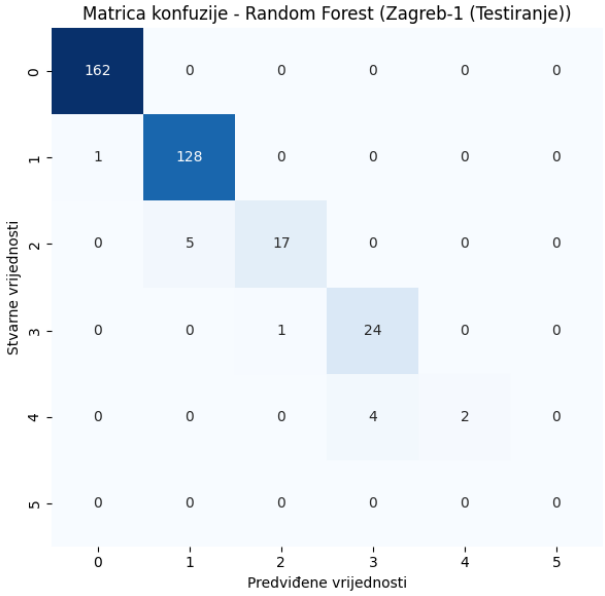
Za testni skup podataka, model je točno predvidio 29 od 31 slučaja za klasu 0. Za klasu 1 model je točno predvidio 54 od 56 slučaja. Za klasu 2, model je točno predvidio 5 slučaja, a pogrešno 13. Klasa 3 je točno predviđena 21 put, dok je 2 puta pogrešno klasificirana. Za klasu 4, model je pogriješio u predviđanju oba dva puta. Klasa 5 je točno predviđena 2 puta. Rezultati su prikazani na slici 35.



Slika 35. *Prikaz stvarnih vrijednosti i modelom predviđenih vrijednosti AQI na lokaciji Zagreb-3 za testni skup podataka za razdoblje od kolovoza do kraja prosinca 2023. godine*

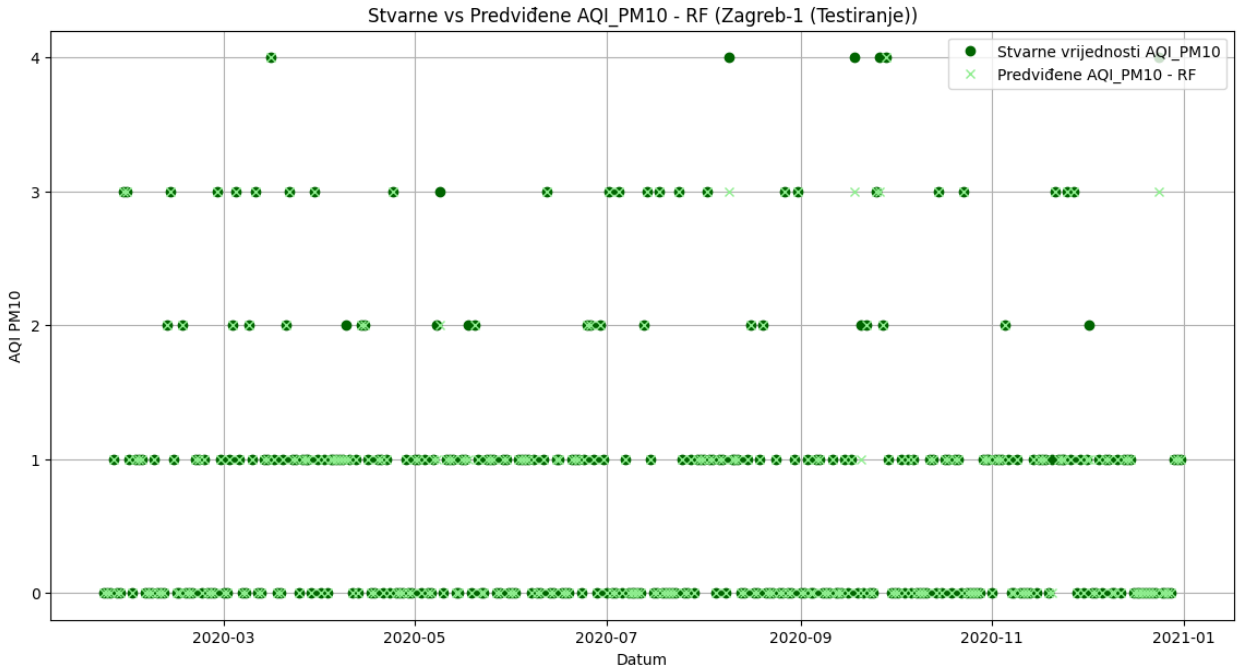
5.3.2. RF algoritam

Prikazani su rezultati predikcije AQI, RF klasifikacijskim modelom na testnom skupu. Slike 36, 38 i 40 prikazuju konfuzijske matrice ispravno i pogrešno predviđenih podataka. Grafovi na slikama 37, 39 i 41 prikazuju usporedbu stvarnih vrijednosti AQI PM_{10} s predviđenim vrijednostima koristeći SVM model na lokaciji ZG1, ZG2 i ZG3 za testni skup podataka. Testni skup podataka sastoji se od istih podataka koji su korišteni i za SVM model. Tamno zelene točke predstavljaju stvarne vrijednosti AQI, a svijetlo zeleni križići predstavljaju predviđene vrijednosti AQI od strane RF modela.

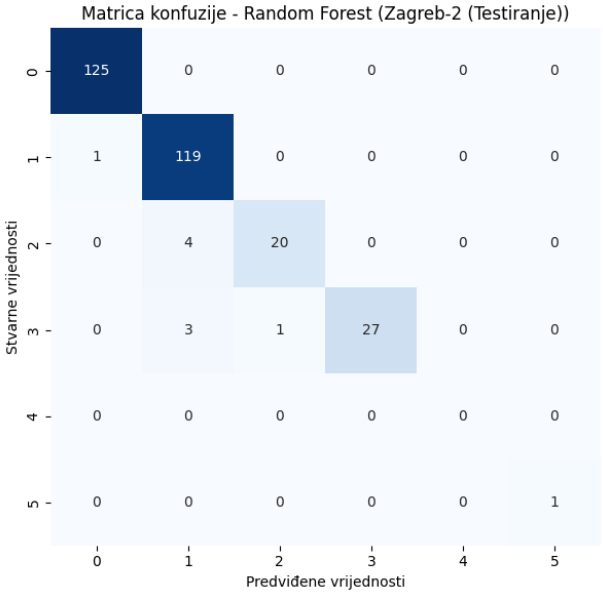


Slika 36. *Matrice konfuzije za testni skup podataka za lokaciju Zagreb-1*

Matrica konfuzije za testni skup podataka prikazuje da je model točno predvidio svih 162 slučaja za klasu 0. Za klasu 1 model je točno predvidio 128 od 129 slučaja. Za klasu 2, model je točno predvidio 17 od 22 slučaja. Klasa 3 je točno predviđena 24 puta, a do pogreške je došlo samo jedanput. Za klasu 4, model je točno predvidio 2 od 6 slučaja. U skupu ne postoje podaci za klasu 5. Isti rezultati prikazani su grafički na slici 34.

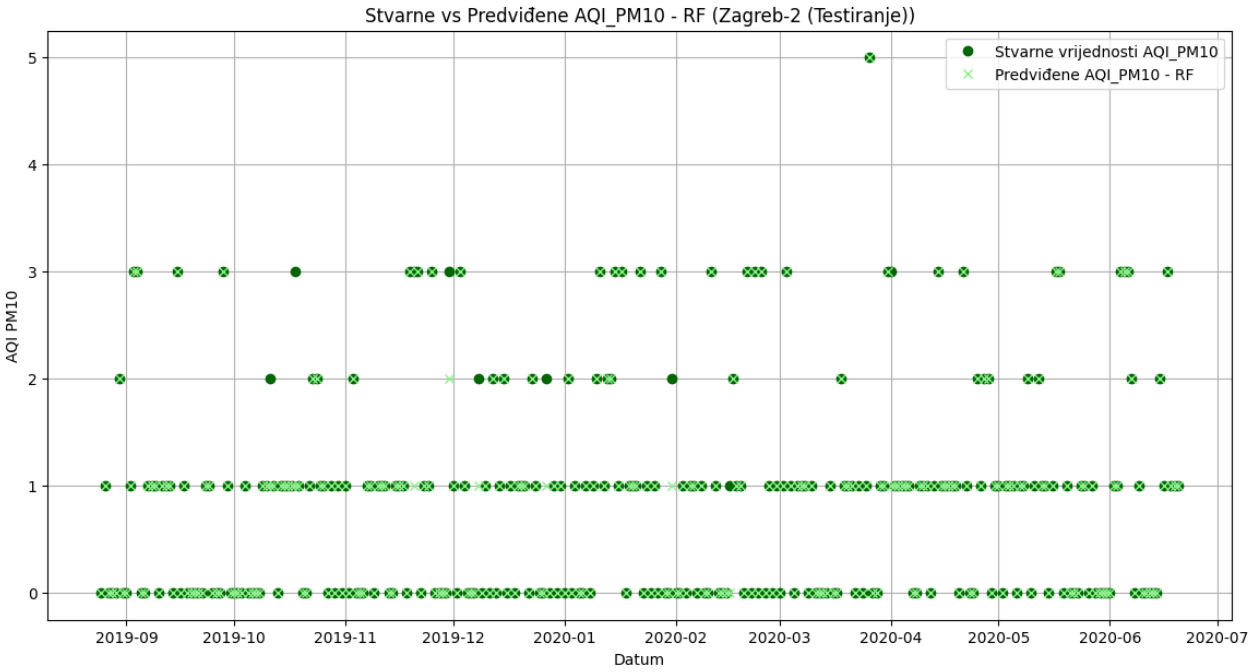


Slika 37. *Prikaz stvarnih vrijednosti i modelom predviđenih vrijednosti AQI na lokaciji Zagreb-1 za testni skup podataka za razdoblje od siječnja do prosinca 2020. godine*

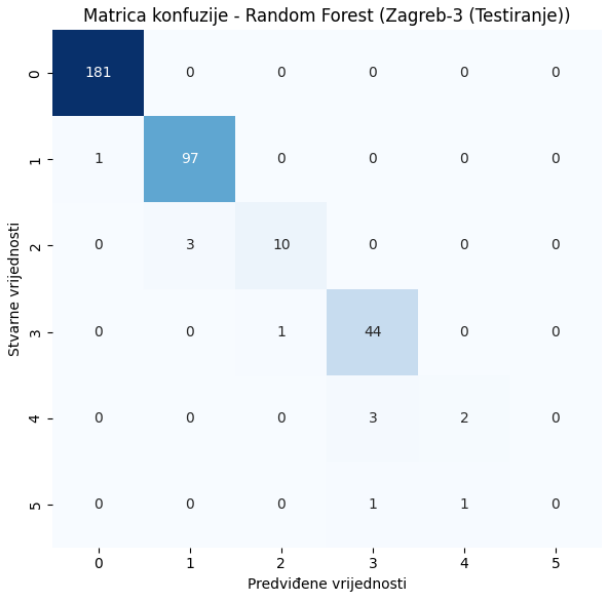


Slika 38. *Matrica konfuzije za testni skup podataka za lokaciju Zagreb-2*

Matrica konfuzije za testni skup podataka prikazuje da je model točno predvidio svih 125 slučaja za klasu 0. Za klasu 1, model je točno predvidio 119 od 120 slučaja. Za klasu 2, model je točno predvidio 20 od 24 slučaja. Klasa 3 je točno predviđena 27 puta, a 4 puta je došlo do pogreške. Za klasu 4 ne postoje podaci u skupu. Klasa 5 je jedanput točno predviđena. Rezultati predviđanja prikazani su grafički na slici 36.

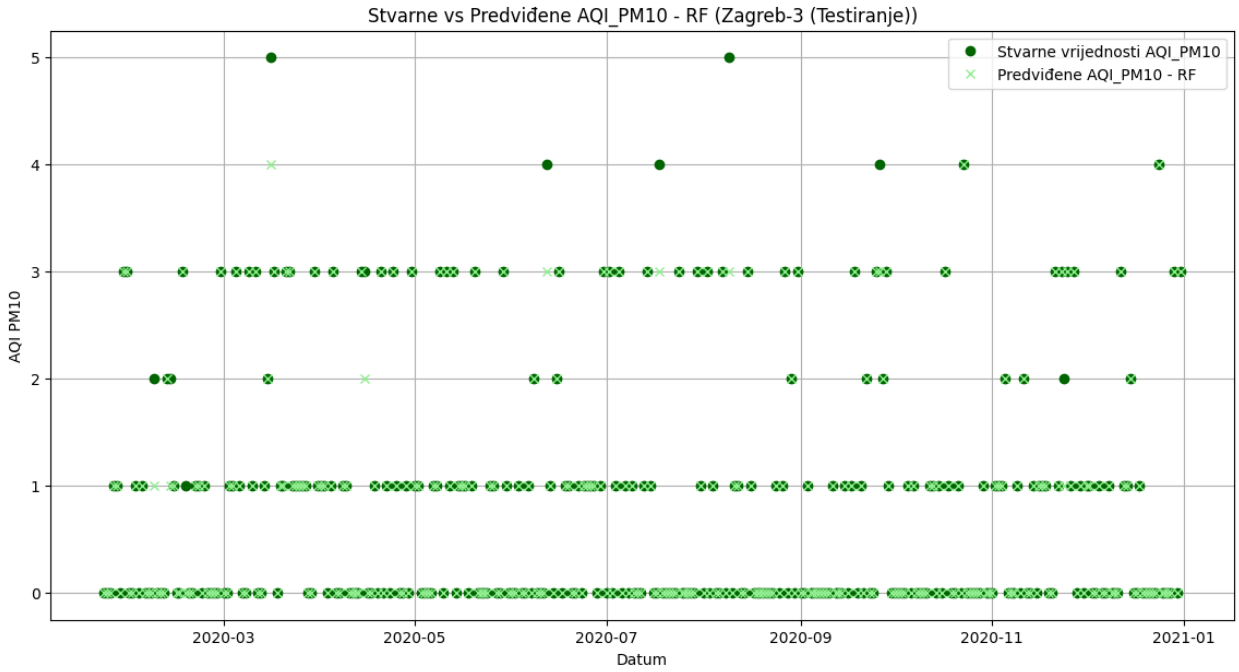


Slika 39. *Prikaz stvarnih vrijednosti i modelom predviđenih vrijednosti AQI na lokaciji Zagreb-2 za testni skup podataka za razdoblje od kolovoza do svibnja 2020. godine.*



Slika 40. *Matrica konfuzije za testni skup podataka za lokaciju Zagreb-3*

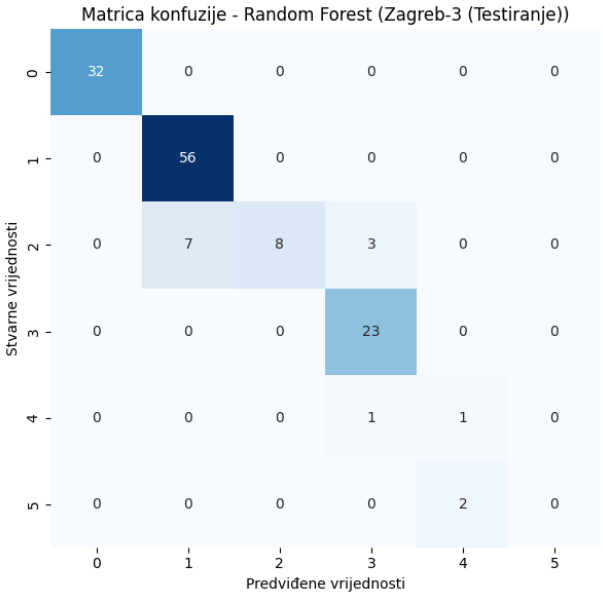
Matrica konfuzije za testni skup podataka prikazuje da je model točno predvidio svih 181 slučaja za klasu 0. Za klasu 1, model je točno predvidio 97 slučaja, a jednom je došlo do pogreške. Za klasu 2, model je točno predvidio 10 od 13 slučaja. Klasa 3 je točno predviđena 44 puta, dok je samo jednom došlo do pogreške. Za klasu 4, model je točno predvidio 2 od 5 slučaja. Klasa 5 nije točno predviđena, 2 puta je pogrešno klasificirana. Rezultati su prikazani grafički na slici 38.



Slika 41. *Prikaz stvarnih vrijednosti i modelom predviđenih vrijednosti AQI na lokaciji Zagreb-3 za testni skup podataka za razdoblje od siječnja do prosinca 2020. godine*

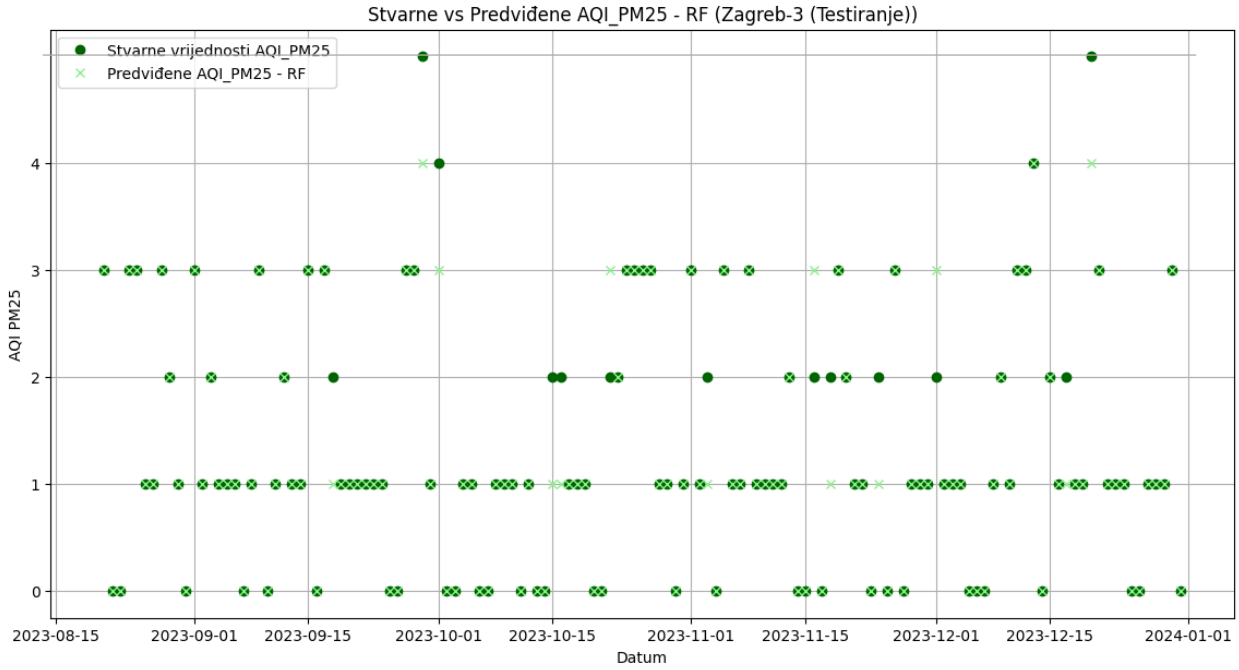
Iz rezultata se vidi da za klase 0 i 1, koje sadrže najviše podataka, model pokazuje visoku točnost u predikciji, što ukazuje na pouzdanost modela za ove specifične klase. RF model pokazuje bolje predikcije i nešto manje zastupljenih klasa (2 i 3) i onih vrlo rijetkih (4 i 5) u odnosu na SVM model.

Također su prikazani rezultati predikcije AQI na temelju koncentracije $PM_{2,5}$, RF modelom na testnom skupu. Slika 42 prikazuje konfuzijsku matricu ispravno i pogrešno predviđenih podataka. Graf na slici 40 prikazuje usporedbu stvarnih vrijednosti AQI $PM_{2.5}$ sa predviđenim vrijednostima, koristeći SVM model na lokaciji ZG3 za testni skup podataka. Tamno zelene točke predstavljaju stvarne vrijednosti AQI, a svijetlo zeleni križići predstavljaju predviđene vrijednosti AQI od strane RF modela.



Slika 42. *Matrica konfuzije za testni skup podataka za lokaciju Zagreb-3 ($PM_{2,5}$)*

Matrica konfuzije za testni skup podataka prikazuje da je model točno predvidio svih 32 slučaja za klasu 0. Za klasu 1 model je točno predvidio svih 56 slučaja. Za klasu 2, model je točno predvidio 8 od 18 slučaja, 10 puta je došlo do pogreške. Klasa 3 je točno predviđena 23 puta. Za klasu 4, model je točno 1 od 2 slučaja. Klasa 5 je krivo predviđena dva puta. Rezultati su prikazani grafički na slici 43.



Slika 43. *Prikaz stvarnih vrijednosti i modelom predviđenih vrijednosti AQI na lokaciji Zagreb-3 za testni skup podataka u razdoblju od kolovoza do prosinca 2023. godine*

5.3.3. Usporedba SVM i RF modela

Točnost modela ukazuje na ispravnost modela u predviđanju klasa. Osjetljivost modela ukazuje na to koliko dobro model identificira stvarne pozitivne instance. Preciznost pokazuje koliko su predviđanja modela točna kada model predviđa određenu klasu. Viša preciznost modela znači da je manje sklon davanju lažno pozitivnih predikcija. Viši F1 rezultat označuje bolju ravnotežu između preciznosti i osjetljivosti, što ga čini pouzdanijim.

Kriterij vrednovanja na testnom skupu podataka za sve 3 lokacije i oba modela prikazani su u tablicama 26, 28 i 30, a pojedinačni rezultati za svaku klasu prikazani su u tablicama 27, 29 i 31. Ovi rezultati se odnose na predviđanje AQI na temelju koncentracije PM_{10}.

Tablica 26. Ukupni klasifikacijski kriterij vrednovanja za oba modela (lokacija ZG1).

Metrika	RF	SVM
Točnost	0,97	0,87
Preciznost	0,95	0,75
Osjetljivost	0,81	0,66
F1 rezultat	0,85	0,68

Model slučajnih šuma pokazuje višu ukupnu točnost (0,97) u usporedbi sa SVM modelom (0,87). RF model također pokazuje višu preciznost (0,95) u odnosu na SVM model (0,75). F1 rezultat je također bolji kod RF modela (0,85) u usporedbi sa SVM modelom (0,68) jer su i preciznost i osjetljivost RF modela (0,81) puno više od preciznosti i osjetljivosti (0,66) SVM modela.

Tablica 27. Kriteriji vrednovanja modela RF i SVM za pojedinu klasu na lokaciji Zagreb-1.

Klasa	RF model			SVM model			
	Preciznost	Osjetljivost	F1-rezultat	Preciznost	Osjetljivost	F1-rezultat	Podrška
0	1,00	1,00	1,00	0,89	0,97	0,93	162
1	0,96	0,99	0,98	0,89	0,84	0,86	129
2	0,94	0,77	0,85	0,73	0,50	0,59	22
3	0,86	0,96	0,91	0,75	0,84	0,79	25
4	1,00	0,33	0,50	0,50	0,17	0,25	6
5	/	/	/	/	/	/	0

makro prosjek	0,95	0,81	0,85	0,75	0,66	0,68	344
težinska srednja vrijednost	0,97	0,97	0,96	0,86	0,87	0,86	344

RF model ima savršene vrijednosti kriterija za klasu 0 s preciznošću, osjetljivošću i F1 rezultatom od 1,00. SVM model također pokazuje dobre rezultate, ali nešto niže (preciznost 0,89, osjetljivost 0,97, F1 rezultat 0,93). Klasa 0 je vrlo dobro zastupljena u podacima, stoga je i model dobro istreniran za njezino predviđanje.

Klasu 1, RF model predviđa gotovo savršeno (0,96, 0,99, 0,98), dok je SVM model lošiji u predviđanju ove klase (0,89, 0,84, 0,86).

RF model pokazuje relativno visoku učinkovitost za klasu 2 s preciznošću, osjetljivošću i F1 rezultatom od 0,94, 0,77 i 0,85. SVM model ima puno niže vrijednosti metrike (0,73, 0,50, 0,59). RF model je bolji u prepoznavanju ove klase, koja je manje zastupljena (svega 22 podatka) od klase 0 i 1.

Klasa 3 je također manje zastupljena (25 podataka) od klase 0 i 1, ipak SVM model pokazuje bolje rezultate metrike (0,75, 0,84, 0,79) nego za klasu 2. Međutim, RF model ima i dalje bolje rezultate s preciznošću, osjetljivošću i F1 rezultatom od 0,86, 0,96 i 0,91.

RF model pokazuje visoku preciznost (1,00) za klasu 4, ali vrlo nisku osjetljivost (0,33), što rezultira F1 rezultatom od 0,50. SVM model ima puno niže metrike (preciznost 0,50, osjetljivost svega 0,17 i F1 rezultat 0,25). RF model je bolji u preciznom predviđanju klase 4, ali može propustiti neke instance, dok SVM model vrlo loše predviđa ovu klasu.

U skupu za vrednovanje ne postoje kategorije klase 5, stoga ne postoje ni kriteriji vrednovanja za njezino predviđanje.

Tablica 28. Ukupni klasifikacijski kriterij vrednovanja za oba modela (lokacija ZG2).

Metrika	RF	SVM
Točnost	0,97	0,86
Preciznost	0,98	0,85
Osjetljivost	0,94	0,79
F1 rezultat	0,96	0,81

RF model pokazuje višu ukupnu točnost (0,97) u usporedbi sa SVM modelom (0,86). RF model također pokazuje višu preciznost (0,98) u odnosu na SVM model (0,85).

Osjetljivost RF modela (0,94) je također viša u odnosu na SVM model (0,79). F1 rezultat RF modela je viši (0,96) u usporedbi sa rezultatom SVM modela (0,81).

Tablica 29. Kriterij vrednovanja modela RF i SVM za pojedinu klasu na lokaciji Zagreb-2.

Klasa	RF model			SVM model			
	Preciznost	**Osjetljivost**	**F1-rezultat**	**Preciznost**	**Osjetljivost**	**F1-rezultat**	**Podrška**
0	1,00	1,00	1,00	0,90	0,97	0,93	125
1	0,94	0,99	0,97	0,85	0,88	0,87	120
2	0,95	0,83	0,89	0,56	0,38	0,45	24
3	1,00	0,87	0,93	0,92	0,74	0,82	31
4	/	/	/	/	/	/	0
5	1,00	1,00	1,00	1,00	1,00	1,00	1
makro prosjek	0,98	0,94	0,96	0,85	0,79	0,81	301
težinska srednja vrijednost	0,97	0,97	0,97	0,86	0,86	0,86	301

RF model ima savršene vrijednosti kriterija za klasu 0 s preciznošću, osjetljivošću i F1 rezultatom od 1,00. SVM model također pokazuje dobre rezultate, ali nešto niže (preciznost 0,90, osjetljivost 0,97, F1 rezultat 0,93). Klasa 0 je dobro zastupljena u podacima (125 instanci), stoga je i model dobro istreniran za njezino predviđanje.

RF model gotovo savršenu metriku za klasu 1 (preciznost, osjetljivost i F1 rezultat redom 0,94, 0,99 i 0,97) jer je također dobro zastupljena u skupu podataka (120 instanci), SVM model pokazuje nešto niže rezultate (preciznost, osjetljivost i F1 rezultat redom 0,85, 0,88 i 0,87).

RF model ima višu preciznost, osjetljivost i F1 rezultat (redom 0,95, 0,83 i 0,89) za klasu 2 u usporedbi sa SVM modelom (0,56, 0,38 i 0,45).

RF model ima višu preciznost, osjetljivost i F1 rezultat (redom 1,00, 0,87 i 0,93) za klasu 3 u usporedbi sa SVM modelom (0,92, 0,74 i 0,82). Ipak su rezultati općenito bolji u odnosu na metriku za klasu 2 jer postoji 31 instanca klase 3 u podacima.

Ne postoje podaci za klasu 4 u skupu podataka.

RF model i SVM model pokazuju savršene metrike u predviđanju klas 5 iako u skupu podataka postoji samo jedna instanca iste.

Tablica 30. Ukupni klasifikacijski kriterij vrednovanja za oba modela (lokacija ZG3).

Metrika	RF	SVM
Točnost	0,97	0,91
Preciznost	0,91	0,65
Osjetljivost	0,69	0,66
F1 rezultat	0,71	0,81

RF model pokazuje višu ukupnu točnost (0,97) u usporedbi sa SVM modelom (0,91). RF model pokazuje višu preciznost (0,91) u odnosu na SVM model (0,65). Osjetljivost je nešto manja za SVM model (0,66) u odnosu na RF model (0,69). Zato je i F1 rezultat nešto niži kod RF modela (0,71) u usporedbi sa SVM modelom (0,81).

Tablica 31. Kriterij vrednovanja modela RF i SVM za pojedinu klasu na lokaciji Zagreb-3.

Klasa	RF model			SVM model			
	Preciznost	Osjetljivost	F1-rezultat	Preciznost	Osjetljivost	F1-rezultat	Podrška
0	0,99	1,00	0,99	0,94	0,99	0,97	181
1	0,97	0,99	0,98	0,92	0,86	0,89	98
2	0,91	0,77	0,83	0,62	0,62	0,62	13
3	0,92	0,98	0,95	0,97	0,87	0,92	54
4	0,67	0,40	0,50	0,43	0,60	0,50	5
5	0,00	0,00	0,00	0,00	0,00	0,00	2
makro prosjek	0,74	0,69	0,71	0,65	0,66	0,65	344
težinska srednja vrijednost	0,96	0,97	0,97	0,91	0,91	0,91	344

RF model ima skoro savršene vrijednosti kriterija za klasu 0 s preciznošću, osjetljivošću i F1 rezultatom od 0,99, 1,00 i 0,99. SVM model također pokazuje dobre rezultate, ali nešto niže (preciznost 0,94, osjetljivost 0,99, F1 rezultat 0,98). Klasa 0 je najzastupljenija u podacima (181 instanca), stoga su oba modela dobro istrenirana za njezino predviđanje.

RF model ima visoku preciznost (0,97), osjetljivost (0,99) i F1 rezultat (0,98) za klasu 1. Preciznost SVM modela iznosi 0,92, dok osjetljivost i F1 rezultat iznose 0,86 i 0,89. Klasa 1 je također dobro zastupljena u podacima (98 instanci).

Klasa 2 ima samo 13 instanci u skupu podataka, što utječe na sposobnost predviđanja ove klase. Zato su i rezultati niži u odnosu na prve dvije klase. RF model i dalje ima bolje

vrijednosti preciznosti, osjetljivosti i F1 rezultata (redom 0,91, 0,77 i 0,83) od SVM modela (0,62).

RF model pokazuje vrlo visoku učinkovitost za klasu 3 s preciznošću, osjetljivošću i F1 rezultatom od 0,92, 0,98 i 0,95. SVM model ima nešto niže metrike (0,97, 0,87 i 0,92). RF model je bolji u prepoznavanju ove klase, koja je također vrlo zastupljena u podacima (45 instanci).

RF model pokazuje nisku preciznost (0,67) i osjetljivost (0,40), što rezultira F1 rezultatom od 0,50 za klasu 4. SVM model ima niže vrijednosti metrike (preciznost 0,43, osjetljivost 0,60, F1 rezultat 0,50). RF model je bolji u preciznom predviđanju klase 4, ali može propustiti neke instance. Postoji samo 5 instanci u skupu podataka, stoga dolazi do pogreške u predviđanju ove klase.

Klasa 5 se pojavljuje u skupu samo u 2 instance. Oba modela nisu predvidjela te instance, stoga su rezultati metrika 0.

Rezultati analize za sve tri lokacije u gradu Zagrebu (Zagreb-1, Zagreb-2, Zagreb-3) pokazuju slične performanse Random Forest (RF) i Support Vector Machine (SVM) modela. Općenito, RF model pokazuje bolje rezultate u predviđanju klasa ciljane AQI varijable u usporedbi sa SVM modelom, iako SVM model također pokazuje solidne performanse. RF model ima veću ukupnu točnost u predviđanju AQI klasa za sve tri lokacije u usporedbi sa SVM modelom. Ovo ukazuje na veću sposobnost RF modela da ispravno klasificira različite kategorije kvalitete zraka. Nadalje, RF model pokazuje konzistentno bolje vrijednosti F1 rezultata za većinu klasa, što ukazuje na njegovu veću uravnoteženost između preciznosti i osjetljivosti. RF model ima višu preciznost, što znači da je manje sklon davanju lažno pozitivnih predikcija u usporedbi sa SVM modelom. To je posebno važno za osiguranje pouzdanih predikcija koje se mogu koristiti za donošenje odluka o kvaliteti zraka.

Klase 0, 1 i 3, koje imaju više podataka se gotovo savršeno predviđaju RF modelom, dok SVM model pokazuje nešto niže ali i dalje vrlo dobre rezultate predikcije. U predviđanju rijetkih klasa (2, 4, 5) modeli rade više pogrešaka, zbog malog broja podataka nisu dobro istrenirani. Pokazalo se da je RF model bolji model za predviđanje rijetko zastupljenih klasa.

Dakle, RF model pokazuje konzistentno bolje performanse na svim lokacijama u Zagrebu u odnosu na SVM model. Zbog svoje veće točnosti, preciznosti i uravnoteženosti između osjetljivosti i preciznosti, RF model se može smatrati pouzdanijim izborom za predikciju kvalitete zraka u gradu Zagrebu. SVM model, iako manje učinkovit, može poslužiti kao solidna alternativa.

Kriterij vrednovanja na testnom skupu podataka za lokaciju Zagreb-3 i oba modela prikazani su u tablici 32, a pojedinačni rezultati za svaku klasu prikazani su u tablici 32. Ovi rezultati se odnose na predviđanje AQI na temelju koncentracije $PM_{2,5.}$

Tablica 32. Kriterij vrednovanja modela RF i SVM za pojedinu klasu na lokaciji Zagreb-3.

Metrika	RF	SVM
Točnost	0,90	0,83
Preciznost	0,68	0,67
Osjetljivost	0,66	0,68
F1 rezultat	0,65	0,64

Točnost te ostale metrike oba modela su niže u odnosu na metrike modela za predviđanje AQI PM_{10}. RF model pokazuje bolju točnost, preciznost i F1 rezultat, SVM ima bolju osjetljivost.

Tablica 33. Kriterij vrednovanja modela RF i SVM za pojedinu klasu na lokaciji Zagreb-3.

Klasa	RF modeli			SVM modeli			
	Preciznost	Osjetljivost	F1-rezultat	Preciznost	Osjetljivost	F1-rezultat	Podrška
0	1,00	1,00	1,00	0,94	0,91	0,92	32
1	0,89	1,00	0,84	0,83	0,96	0,89	56
2	1,00	0,44	0,62	0,83	0,28	0,42	18
3	0,85	0,50	0,40	0,78	0,91	0,84	23
4	0,33	0,00	0,00	0,00	0,00	0,00	2
5	0,00	0,00	0,00	0,67	1,00	0,80	2
makro prosjek	0,68	0,66	0,65	0,67	0,68	0,64	133
težinska srednja vrijednost	0,90	0,90	0,89	0,83	0,83	0,81	133

Klasa 1 koja je najzastupljenija prikazuje najviše vrijednosti kriterija, zatim slijedi klasa 0 (druga najzastupljenija), onda klasa 3. Klase 2, 4 i 5 imaju niske vrijednosti metrika vrednovanja jer nisu toliko zastupljene u podacima.

Ovi modeli pokazuju lošije rezultate predviđanja, zbog vrlo malog skupa podataka korištenih za njihovo treniranje. Također, ne mogu se smatrati valjanima jer korišteni podaci nisu validirani.

6. ZAKLJUČAK

U ovom radu prikazana je primjena metoda strojnog učenja u svrhu razvoja modela za procjenu indeksa kvalitete zraka za frakcije lebdećih čestica PM_{10} i $PM_{2,5}$ na tri lokacije u gradu Zagrebu. Korišteni su klasifikacijski modeli nasumičnih šuma i modeli potpornih vektora. Modeli su razvijeni u svrhu procjene kvalitete zraka na temelju ulaznih temporalnih i meteoroloških podataka s lokacije za koju je rađen model i lokacije Maksimir. Modeli su razvijeni za svaku od tri lokacije (Zagreb-1, Zagreb-2, Zagreb-3), što ukupno čini 4 modela za predviđanje AQI PM_{10}, jer je korišten isti model SVM za sve 3 lokacije. Dodatno su razvijena 2 modela za predviđanje AQI $PM_{2,5}$, međutim oni se trebaju uzeti s određenom rezervom jer prikupljeni podaci o masenim koncentracijama frakcija lebdećih čestica $PM_{2,5}$ nisu prošli proces validacije i korekcije, te je razvijen samo jedan model za predviđanje AQI na temelju koncentracije $PM_{2,5}$ na lokaciji ZG3. Za razvoj modela i primjenu algoritama RF i SVM korišten je programski jezik Python i njegove knjižnice.

Modeli RF i SVM razvijeni na lokaciji Zagreb-1 pokazali su visoku točnost predikcije, pri čemu je RF model pokazao bolju ukupnu točnost (97 %) u usporedbi sa SVM modelom (87 %). Slični rezultati dobiveni su i na lokacijama Zagreb-2 i Zagreb-3, gdje je RF model također nadmašio SVM model prema točnosti. Veća točnost RF modela znači da su ti modeli općenito bolji u predviđanju ispravnih klasa u većini situacija, čineći ih pouzdanijim za ukupnu klasifikaciju.

Za sve lokacije RF model nadmašuje SVM model u svim kriterijima vrednovanja. To znači da RF model bolje prepoznaje stvarne pozitivne primjere (viša osjetljivost) i postiže bolju ravnotežu između preciznosti (točnost u predviđanju pozitivnih primjera) i osjetljivosti (otkrivanje stvarnih pozitivnih slučajeva). Drugim riječima, RF model ne samo da identificira više stvarnih pozitivnih primjera nego također održava visoku točnost u svojim predviđanjima, što ga čini efikasnijim u usporedbi sa SVM modelom.

Oba modela pokazala su se uspješnima za predikciju češćih klasa (0 i 1), dok su imali više poteškoća s predikcijom rijetkih klasa (2, 3), a najviše poteškoća u predviđanju vrlo rijetkih klasa (4, 5).

Analiza matrica konfuzije i grafova stvarnih i predviđenih vrijednosti pokazala je da su RF modeli dosljedno bolji u predikciji kvalitete zraka u odnosu na SVM modele, time potvrđujući rezultate klasifikacijskih kriterija vrednovanja uspješnosti modela.

Treba uzeti u obzir da na kvalitetu zraka, odnosno koncentraciju PM_{10}, mogu utjecati mnogi čimbenici koji nisu uključeni u izradu modela. To su sekundarno stvaranje čestica i prijenos čestica na velike udaljenosti, sunčevo zračenje, gustoća prometa, doprinos resuspenzije cestovne prašine itd. Drugi ograničavajući čimbenik je učestalost podataka i metoda mjerenja. Ovaj rad se temelji na dnevnim prosjecima masene koncentracije čestica i dnevnim vrijednostima meteoroloških varijabli. Takve procjene predstavljaju grubi prikaz promjena koncentracije čestica i meteoroloških varijabli koje se događaju svakim satom.

Zaključno, modeli RF i SVM razvijeni u ovom radu mogu se uspješno primijeniti za predviđanje AQI na novim skupovima podataka iako se temelje na malom broju varijabli koje mogu utjecati na AQI i relativno malom broj dostupnih podataka. RF modeli, zahvaljujući svojoj visokoj točnosti mogu pružiti precizne i pouzdane predikcije, čime se omogućava bolje upravljanje kvalitetom zraka u gradu Zagrebu i općenito urbanim sredinama. SVM modeli, iako nešto manje učinkoviti, također mogu poslužiti u predikciji AQI kategorija. Za poboljšanje modela, osim korištenja više ulaznih varijabli i većeg broja podataka, može se koristiti tehnika SMOTE za generiranje vrijednosti manjinskih, rijetkih klasa. Na taj način bi se prevladao problem klasne neravnoteže i modeli bi bili dobro istrenirani za predviđanje svih kategorija.

7. POPIS SIMBOLA I OZNAKA

pH – mjera kiselosti (engl. *measure of acidity*)

Akronimi korišteni u radu:

AQG	Smjernice za kvalitetu zraka (engl. *Air Quality Guidelines*)
AQI	Indeks kvalitete zraka (engl. *Air Quality Indeks*)
AUC	Površina ispod ROC krivulje (engl. *Area under ROC curve)*
CCN	Oblačne kondenzacijske jezgre (engl. *Cloud condensation nuclei*)
CV	Unakrsna validacija (engl. *Cross validation*)
DHMZ	Državni hidrometeorološki zavod (engl. *Croatian Meteorological and Hydrological Service*)
DT	Stablo odluke (engl. *Decision Tree*)
FN	Netočno predviđene negativne instance (engl. *False Negative*)
FP	Netočno predviđene pozitivne instance (engl. *False Positive*)
FPR	Stope netočno predviđenih pozitivnih instanci (engl. *False Positive*) *Rate*)
HL	Huberova pogreška (engl. *Huber Loss*)
IMI	Institut za medicinska istraživanja i medicinu rada (engl. *Institute for Medical Research and Occupational Health*)
IQR	Interkvartilni raspon (engl. *Interquartile Range*)
LCL	Log Cosh gubitak (engl. *Log Cosh Loss*)
LDA	Linearna diskriminantna analiza (engl. *Linear discriminant analysis*)
MAE	Srednja apsolutna pogreška (engl. *Mean Absolute Error*)
MAPE	Srednja apsolutna postotna pogreška (engl. *Mean Absolute Percentage Error*)
MBE	Srednja pristranost pogreške (engl. *Mean Bias Error*)
ML	Strojno učenje (engl. *Machine Learning*)
MLH	Visina mješovitog sloja (engl. *Mixing Layer Height*)
MSE	Srednja kvadratna pogreška (engl. *Mean Squared Error*)
NaN	Nedostajuće vrijednosti (engl. *Not a Number*)
NRMSE	Normalizirani korijen srednje kvadratne pogreške (engl. *Normalized Root Mean Squared Error*)
OOB	Podaci isključeni iz skupa za treniranje RF (engl. *Out-of-bag data*)
PAN	Perokisacil nitrat (engl. *Peroxyacetyl nitrate*)

PCA	Analiza glavnih komponenti (engl. *Principal Component Analysis*)
PM	Lebdeće čestice (engl. *Particulate matter*)
PM_{10}	Lebdeće čestice aerodinamičnog promjera manjeg od 10 μm (engl. *Particulate matter with a diameter of 10 μm or less*)
$PM_{2,5}$	Lebdeće čestice aerodinamičnog promjera manjeg od 2,5 μm (engl. *Particulate matter with a diameter of 2,5 μm or less*)
PM_1	Lebdeće čestice aerodinamičnog promjera manjeg od 1 μm (engl. *Particulate matter with a diameter of 1 μm or less*)
PR	Krivulja preciznost-odziv (engl. *Precision-recall curve*)
QL	Kvantilna pogreška (engl. *Quantile Loss*)
RA	Relativna apsolutna pogreška (engl. *Relative Absolute Error*)
RFC	Klasifikacijski algoritam nasumičnih šuma (engl. *Random Forest Classification*)
RF	Algoritam nasumične šume (engl. *Random Forest algorithm)*
RH	Relativna vlažnost (engl. *Relative humidity*)
RMSE	Korijen srednje kvadratne pogreške (engl. *Root Mean Squared Error*)
RMSLE	Korijen srednje kvadratne logaritamske pogreške (engl. *Root Mean Squared Logarithmic Error*)
ROC	Krivulja operativnih karakteristika (engl. *Receiver Operating Characteristic*)
RRMSE	Relativni korijen srednje kvadratne pogreške (engl. *Relative Root Mean Squared Error*)
RSE	Relativna kvadratna pogreška (engl. *Relative Squared Error*)
SVM	Stroj potpornih vektora (engl. *Support Vector Machine algorithm*)
SVR	Regresijski algoritam potpornih vektora (engl. *Support Vector Regression*)
TN	Točno predviđene negativne instance (engl. *True Negative*)
TP	Točno predviđene pozitivne instance (engl. *True Positive*
TPR	Stopa stvarnih pozitivnih instanci (engl. *True Positive Rate*)
UHI	Urbanski toplinski otok (engl. *Urban Heat Island*)
VOC	Hlapljivi organski spojevi (engl. *Volatile organic compounds*)
WHO	Svjetska zdravstvena organizacija (engl. *World Health Organization*)
ZG1	Lokacija Zagreb-1 (engl. *Location in Zagreb-1*)
ZG2	Lokacija Zagreb-2 (engl. *Location in Zagreb-2*)
ZG3	Lokacija Zagreb-3 (engl. *Location in Zagreb-3*)

8. LITERATURA

1. Lee C, Lee K, Kim S, Yu J, Jeong S, Yeom J. Hourly Ground-Level PM2.5 Estimation Using Geostationary Satellite and Reanalysis Data via Deep Learning. *Remote Sensing*. 2021;13(11):2121. doi:10.3390/rs13112121

2. A Survey on Machine Learning-Based Performance Improvement of Wireless Networks: PHY, MAC and Network Layer. Accessed June 27, 2024. https://www.mdpi.com/2079-9292/10/3/318

3. Ravindiran G, Hayder G, Kanagarathinam K, Alagumalai A, Sonne C. Air quality prediction by machine learning models: A predictive study on the indian coastal city of Visakhapatnam. *Chemosphere*. 2023;338:139518. doi:10.1016/j.chemosphere.2023.139518

4. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Accessed June 29, 2024. https://www.who.int/publications/i/item/9789240034228

5. Harold F., Elizabeth J. Fechner, Hemond J. Fechner. *Chemical Fate and Transport in the Environment - Edition 4 - Elsevier Health Inspection Copies*. http://educate.elsevier.com/book/details/9780128222522

6. V. Tomašić, Tehnološki procesi u zaštiti zraka, Nastavni materijal, Fakultet kemijskog inženjerstva i tehnologije, Zagreb 2019.

7. Sun B, Fang C, Liao X, Guo X, Liu Z. The relationship between urbanization and air pollution affected by intercity factor mobility: A case of the Yangtze River Delta region. *Environmental Impact Assessment Review*. 2023;100:107092. doi:10.1016/j.eiar.2023.107092

8. Pénard-Morand C, Annesi-Maesano I. Air pollution: from sources of emissions to health effects. *Breathe*. 2004;1(2):108-119. doi:10.1183/18106838.0102.108

9. Kwak HY, Ko J, Lee S, Joh CH. Identifying the correlation between rainfall, traffic flow performance and air pollution concentration in Seoul using a path analysis. *Transportation Research Procedia*. 2017;25:3552-3563. doi:10.1016/j.trpro.2017.05.288

10. Jakovljević I, Sever Štrukil Z, Godec R, et al. Pollution Sources and Carcinogenic Risk of PAHs in PM1 Particle Fraction in an Urban Area. *Int J Environ Res Public Health*. 2020;17(24):9587. doi:10.3390/ijerph17249587

11. Almeida SM, Manousakas M, Diapouli E, et al. Ambient particulate matter source apportionment using receptor modelling in European and Central Asia urban areas. *Environmental Pollution*. 2020;266:115199. doi:10.1016/j.envpol.2020.115199

12. temperaturni gradijent - Hrvatska enciklopedija. Accessed June 30, 2024. https://www.enciklopedija.hr/clanak/temperaturni-gradijent

13. Mukherjee A, Agrawal M. A Global Perspective of Fine Particulate Matter Pollution and Its Health Effects. *Rev Environ Contam Toxicol*. 2018;244:5-51. doi:10.1007/398_2017_3

14. Grantz DA, Garner JHB, Johnson DW. Ecological effects of particulate matter. *Environ Int*. 2003;29(2-3):213-239. doi:10.1016/S0160-4120(02)00181-2

15. Brunekreef B, Holgate ST. Air pollution and health. *Lancet*. 2002;360(9341):1233-1242. doi:10.1016/S0140-6736(02)11274-8

16. Particulate matter (PM10 and PM2.5) - DCCEEW. Accessed June 30, 2024. https://www.dcceew.gov.au/environment/protection/npi/substances/fact-sheets/particulate-matter-pm10-and-pm25

17. Kvaliteta zraka u Republici Hrvatskoj. Accessed June 27, 2024. https://iszz.azo.hr/iskzl/help.htm

18. vjetar - Hrvatska enciklopedija. Accessed June 27, 2024. https://enciklopedija.hr/clanak/vjetar

19. Beaufortova ljestvica - Hrvatska enciklopedija. Accessed June 27, 2024. https://enciklopedija.hr/clanak/beaufortova-ljestvica

20. Kaluvagunta V, Musali K. Air quality monitoring at residential areas in and around Tirupati-a well-known pilgrimage site in India. *Indian Journal of Science and Technology*. 2011;4. doi:10.17485/ijst/2011/v4i11/30280

21. Understanding the impact of Wind Speed & Direction on Air Po. Accessed June 29, 2024. https://www.clarity.io/blog/air-quality-measurements-series-wind-speed-and-direction

22. How wind and weather affect air pollution. Accessed June 29, 2024. https://www.iqair.com/newsroom/wind-weather-air-pollution

23. Cuhadaroglu B, Demirci E. Influence of some meteorological factors on air pollution in Trabzon city. *Energy and Buildings*. 1997;25(3):179-184. doi:10.1016/S0378-7788(96)00992-9

24. temperatura zraka - Hrvatska enciklopedija. Accessed June 27, 2024. https://www.enciklopedija.hr/clanak/temperatura-zraka

25. US EPA O. Heat Island Effect. Published February 28, 2014. Accessed June 27, 2024. https://www.epa.gov/heatislands

26. Cao J, Zhou W, Zheng Z, Ren T, Wang WM. Within-city spatial and temporal heterogeneity of air temperature and its relationship with land surface temperature. *Landscape and Urban Planning*. 2021;206:103979. doi:10.1016/j.landurbplan.2020.103979

27. Guarnieri G, Olivieri B, Senna G, Vianello A. Relative Humidity and Its Impact on the Immune System and Infections. *Int J Mol Sci*. 2023;24(11):9456. doi:10.3390/ijms24119456

28. Murthy B, R L, Tiwari A, Rathod A, Singh S, Beig G. Impact of Mixing Layer Height on Air Quality in Winter. *Journal of Atmospheric and Solar-Terrestrial Physics*. 2019;197:105157. doi:10.1016/j.jastp.2019.105157

29. Vaishali, Verma G, Das RM. Influence of Temperature and Relative Humidity on PM2.5 Concentration over Delhi. *MAPAN*. 2023;38(3):759-769. doi:10.1007/s12647-023-00656-8

30. relativna vlažnost - Hrvatska enciklopedija. Accessed June 27, 2024. https://www.enciklopedija.hr/clanak/relativna-vlaznost

31. How Does Humidity Affect Air Quality? All You Need to Know - Airly WP | Air Quality Monitoring. Monitor in UK & Europe. Airly Data Platform and Monitors. Accessed June 29, 2024. https://airly.org/en/how-does-humidity-affect-air-quality-all-you-need-to-know/#humidity

32. Yang L, Qian F, Song DX, Zheng KJ. Research on Urban Heat-Island Effect. *Procedia Engineering*. 2016;169:11-18. doi:10.1016/j.proeng.2016.10.002

33. tlak - Hrvatska enciklopedija. Accessed June 29, 2024. https://www.enciklopedija.hr/clanak/tlak

34. anticiklona - Hrvatska enciklopedija. Accessed June 29, 2024. https://enciklopedija.hr/clanak/anticiklona

35. ciklona - Hrvatska enciklopedija. Accessed June 29, 2024. https://enciklopedija.hr/clanak/ciklona

36. oborina - Hrvatska enciklopedija. Accessed June 29, 2024. https://enciklopedija.hr/clanak/oborina

37. Meersens. The impact of weather on air quality. Published March 23, 2022. Accessed June 25, 2024. https://meersens.com/mpact-of-weather-on-air-quality/?lang=en

38. Statistički ljetopis Grada Zagreba. Accessed June 28, 2024. https://www.zagreb.hr/en/statisticki-ljetopis-grada-zagreba/1044

39. Zagreb - Hrvatska enciklopedija. Accessed June 29, 2024. https://www.enciklopedija.hr/clanak/66685

40. Passenger cars per 1 000 inhabitants reached 560 in 2022. Published January 17, 2024. Accessed July 21, 2024. https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20240117-1

41. Zagreb u brojkama, Grad Zagreb, Gradski ured za strategijsko planiranje i razvoj grada, 2023.

42. DHMZ - Državni hidrometeorološki zavod. Accessed June 29, 2024. https://meteo.hr/o_nama.php

43. Analyzing Air Pollutant Reduction Possibilities in the City of Zagreb. Accessed June 28, 2024. https://www.mdpi.com/2220-9964/11/4/259

44. Fugaš M, Gentilizza M, Valić F, Verhovnik S, Proučavanje onečišćenja atmosfere na području grada Zagreba, Arh. hig. rada, 965;161:227, https://hrcak.srce.hr/file/264163. (Accessed July 21, 2024.).

45. Fugaš M, Vađić V, Šega K, Hršak J, Kalinić N, Šišovič A, Air pollution studies, Arh. hig. rada tokikol., 1999;50(2):211-222, https://www.fda.gov/media/148472/download. (Accessed July 11, 2024.).

46. Fugaš M, Pregled aktinosti na očuvanju i poboljšanju kakvoće zraka u Republici Hrvatskoj, Zaštita zraka '97 1997:23-30, https://www.fda.gov/media/148472/download (Accessed July 21, 2024).

47. IMI i kvaliteta zraka – AIRQ. Accessed July 5, 2024. https://www.airq.hr/imi-i-kvaliteta-zraka/

48. Kvaliteta zraka u Republici Hrvatskoj, Izvješće o praćenju kvalitete zraka na teritoriju Republike Hrvatske za 2019, https://iszz.azo.hr/iskzl/godizvrpt.htm?pid=0&t=0 (Accessed July 21, 2024.).

49. Kvaliteta zraka u Republici Hrvatskoj i gradu Zagrebu | NZJZ Andrija Štampar. Accessed June 30, 2024. https://stampar.hr/hr/novosti/kvaliteta-zraka-u-republici-hrvatskoj-i-gradu-zagrebu

50. Jakovljević I, Štrukil ZS, Godec R, Davila S, Pehnec G. Influence of lockdown caused by the COVID-19 pandemic on air pollution and carcinogenic content of particulate matter observed in Croatia. *Air Qual Atmos Health*. 2021;14(4):467-472. doi:10.1007/s11869-020-00950-3

51. Lovrić M, Antunović M, Šunić I, et al. Machine Learning and Meteorological Normalization for Assessment of Particulate Matter Changes during the COVID-19 Lockdown in Zagreb, Croatia. *Int J Environ Res Public Health*. 2022;19(11):6937. doi:10.3390/ijerph19116937

52. McCarthy J. Artificial Intelligence, Logic and Formalizing Common Sense. In: Thomason RH, ed. *Philosophical Logic and Artificial Intelligence*. Springer Netherlands; 1989:161-190. doi:10.1007/978-94-009-2448-2_6

53. What Is Machine Learning (ML)? | IBM. Accessed June 19, 2024. https://www.ibm.com/topics/machine-learning

54. El Naqa I, Murphy MJ. What Is Machine Learning? In: El Naqa I, Li R, Murphy MJ, eds. *Machine Learning in Radiation Oncology: Theory and Applications*. Springer International Publishing; 2015:3-11. doi:10.1007/978-3-319-18305-3_1

55. Carbonell JG, Michalski RS, Mitchell TM. 1 - AN OVERVIEW OF MACHINE LEARNING. In: Michalski RS, Carbonell JG, Mitchell TM, eds. *Machine Learning*. Morgan Kaufmann; 1983:3-23. doi:10.1016/B978-0-08-051054-5.50005-4

56. Alpaydin E. *Introduction to Machine Learning*. 2nd ed. MIT Press; 2010.

57. Jamal S, Goyal S, Grover A, Shanker A. Machine Learning: What, Why, and How? In: ; 2018:359-374. doi:10.1007/978-981-13-1562-6_16

58. Aleksić D. Mogućnosti primjene metoda strojnog učenja u području telekomunikacija, 2021. https://urn.nsk.hr/urn:nbn:hr:119:003361

59. Badillo S, Banfai B, Birzele F, et al. An Introduction to Machine Learning. *Clin Pharmacol Ther*. 2020;107(4):871-885. doi:10.1002/cpt.1796

60. Shetty C, Shedole S, B J S, et al. A Machine Learning Approach for Environmental Assessment on Air Quality and Mitigation Strategy. *Journal of Engineering*. 2024;2024. doi:10.1155/2024/2893021

61. Grange SK, Carslaw DC, Lewis AC, Boleti E, Hueglin C. Random forest meteorological normalisation models for Swiss PM_{10} trend analysis. *Atmospheric Chemistry and Physics*. 2018;18(9):6223-6239. doi:10.5194/acp-18-6223-2018

62. What Is Random Forest? | IBM. Published October 20, 2021. Accessed June 19, 2024. https://www.ibm.com/topics/random-forest

63. RandomForestClassifier. scikit-learn. Accessed June 19, 2024. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

64. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

65. MyEducator - The Random Forest Algorithm. Accessed June 28, 2024. https://app.myeducator.com/reader/web/1421a/15/xu4rw/

66. 1.4. Support Vector Machines. scikit-learn. Accessed June 24, 2024. https://scikit-learn/stable/modules/svm.html

67. All You Need to Know About Support Vector Machines. Spiceworks Inc. Accessed July 5, 2024. https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/

68. Nastavni materijal: Šnajder J., Strojno učenje: 8. Stroj potpornih vektora, UNIZG FER, ak. god. 2020./2021.

69. Jovanovic I. METODA POTPORNIH VEKTORA S PRIMJENAMA U EKSTRAKCIJI INFORMACIJE.

70. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1007/BF00994018

71. Comprehensive Support Vector Machines Guide - Using Illusion to Solve Reality! | by Pranov Mishra | Analytics Vidhya | Medium. Accessed June 24, 2024. https://medium.com/analytics-vidhya/comprehensive-support-vector-machines-guide-using-illusion-to-solve-reality-ad3136d8f877

72. Bojanic D. STROJNO UČENJE PUTEM REGRESIJE I SVM.

73. Welcome to Python.org. Python.org. Published July 2, 2024. Accessed June 29, 2024. https://www.python.org/about/

74. The Python Tutorial. Python documentation. Accessed June 29, 2024. https://docs.python.org/3/tutorial/index.html

75. Python Libraries: Your Comprehensive Guide - Linux Dedicated Server Blog. Published September 7, 2023. Accessed June 29, 2024. https://ioflood.com/blog/python-libraries/

76. Descriptive Statistic - GeeksforGeeks. Accessed June 27, 2024. https://www.geeksforgeeks.org/descriptive-statistic/

77. MUI, prezentacija predobrada podataka, FKIT.

78. Hohnjec A. *Analiza podataka pomoću Python Pandas alata kroz primjer*. info:eu-repo/semantics/bachelorThesis. University of Rijeka. Department of Informatics; 2019. Accessed July 5, 2024. https://urn.nsk.hr/urn:nbn:hr:195:342603

79. What is a feature engineering? | IBM. Accessed June 19, 2024. https://www.ibm.com/topics/feature-engineering

80. Zekan M. PRIMJENA STROJNOG UČENJA U PREDVIĐANJU PONAŠANJA POTROŠAČA.

81. Feature Engineering: Scaling, Normalization, and Standardization - GeeksforGeeks. Accessed June 24, 2024. https://www.geeksforgeeks.org/ml-feature-scaling-part-2/

82. Cross Validation in Machine Learning. GeeksforGeeks. Published November 21, 2017. Accessed June 23, 2024. https://www.geeksforgeeks.org/cross-validation-machine-learning/

83. KFold. scikit-learn. Accessed June 24, 2024. https://scikit-learn/stable/modules/generated/sklearn.model_selection.KFold.html

84. K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries | IEEE Conference Publication | IEEE Xplore. Accessed June 24, 2024. https://ieeexplore.ieee.org/abstract/document/9033691

85. An Introduction to GridSearchCV | What is Grid Search | Great Learning. Accessed June 24, 2024. https://www.mygreatlearning.com/blog/gridsearchcv/

86. Jan Šnajder. Vrednovanje modela, Strojno učenje 1, UNIZG FER, ak. god. 2022./2023.predavanja, v1.3.

87. Kulkarni A, Chong D, Batarseh FA. 5 - Foundations of data imbalance and solutions for a data democracy. In: Batarseh FA, Yang R, eds. *Data Democracy*. Academic Press; 2020:83-106. doi:10.1016/B978-0-12-818366-3.00005-8

88. Evaluation of Regression Models: Model Assessment, Model Selection and Generalization Error. Accessed June 27, 2024. https://www.mdpi.com/2504-4990/1/1/32

89. Evaluation Metric for Regression Models - Analytics Vidhya. Accessed June 24, 2024. https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/

90. Blogger ML. 14 Loss functions you can use for Regression. Medium. Published January 21, 2023. Accessed July 5, 2024. https://medium.com/@mlblogging.k/14-loss-functions-you-can-use-for-regression-b24db8dff987

91. Galaxy Training Network. Published 51:21 . Accessed June 24, 2024. https://training.galaxyproject.org/training-material/404.html

92. Understanding Regression Analysis | SpringerLink. Accessed June 19, 2024. https://link.springer.com/book/10.1007/b102242

93. Linear Regression (Python Implementation). GeeksforGeeks. Published March 19, 2017. Accessed June 19, 2024. https://www.geeksforgeeks.org/linear-regression-python-implementation/

94. Applied Regression Analysis, 3rd Edition | Wiley. Wiley.com. Accessed June 27, 2024. https://www.wiley.com/en-us/Applied+Regression+Analysis%2C+3rd+Edition-p-9780471170822

95. Šimić I, Lovrić M, Godec R, Kröll M, Bešlić I. Applying machine learning methods to better understand, model and estimate mass concentrations of traffic-related pollutants at a typical street canyon. *Environmental Pollution*. 2020;263:114587. doi:10.1016/j.envpol.2020.114587

96. Gupta NS, Mohta Y, Heda K, Armaan R, Valarmathi B, Arulkumaran G. Prediction of Air Quality Index Using Machine Learning Techniques: A Comparative Analysis. *Journal of Environmental and Public Health*. 2023;2023(1):4916267. doi:10.1155/2023/4916267

97. Bhattacharya S, Shahnawaz S. Using Machine Learning to Predict Air Quality Index in New Delhi. Published online December 9, 2021. doi:10.48550/arXiv.2112.05753

98. Haq M. SMOTEDNN: A Novel Model for Air Pollution Forecasting and AQI Classification. *CMC*. 2021;71(1):1403-1425. doi:10.32604/cmc.2022.021968

99. Imam M, Adam S, Dev S, Nesa N. Air quality monitoring using statistical learning models for sustainable environment. *Intelligent Systems with Applications*. 2024;22:200333. doi:10.1016/j.iswa.2024.200333

100. iszz.azo.hr/iskzl/mreza.html?t=1). Accessed July 21, 2024. https://iszz.azo.hr/iskzl/mreza.html?t=1)

101. Turney S. Pearson Correlation Coefficient (r) | Guide & Examples. Scribbr. Published May 13, 2022. Accessed June 19, 2024. https://www.scribbr.com/statistics/pearson-correlation-coefficient/

9. ŽIVOTOPIS

Paola Klonkay, [redacted] Pohađala je X. Gimnaziju Ivan Supek, prirodoslovno-matematički smjer, gdje je maturirala 2017. godine. Iste godine upisuje Fakultet kemijskog inženjerstva i tehnologije, gdje 2022. godine pod vodstvom mentora prof. dr. sc. Petra Kassala stječe akademsku titulu Sveučilišni prvostupnik inženjer kemijskog inženjerstva s radom Stabilizacija ugljikovih nanocijevi poli(vinil butiralom). Diplomski rad izradila je pod mentorstvom izv. prof. dr. sc. Željke Ujević Andrijić na Zavodu za mjerenja i automatsko vođenje procesa i dr. sc. Silvija Davile te Marije Jelene Lovrić Štefiček na Institutu za medicinska istraživanja i medicinu rada na Zavodu za higijenu okoliša u 2024. godini.