**УНИВЕРЗИТЕТ “ГОЦЕ ДЕЛЧЕВ” – ШТИП**

**ФАКУЛТЕТ ЗА ИНФОРМАТИКА**

****

**СЕМИНАРСКА РАБОТА ПО ПРЕДМЕТОТ**

**Машинско учење**

**ТЕМА: Предвидување на квалитетот на воздухот со машинско учење**

**Изработил : Столе Ристов**

**Јуни, 2023,**

**Штип**

# Contents

[Contents 2](#_Toc137378816)

[Вовед 3](#_Toc137378817)

[Цели и истражувачки прашања 4](#_Toc137378818)

[Топлотната мапа на корелација 5](#_Toc137378819)

[Pair Plot 6](#_Toc137378820)

[График на распределба 7](#_Toc137378821)

[Box plot 8](#_Toc137378822)

[Scatter plot за Линеарна Регресија 9](#_Toc137378823)

[Random Forest 10](#_Toc137378824)

[Споредба помеѓу Општини во Скопје 11](#_Toc137378825)

[Заклучок 14](#_Toc137378826)

# 

# Вовед

Во оваа теза, ги истражуваме предвидувањата за квалитетот на воздухот во градот Скопје и со посебен фокус на применетото машинско учење. Се имплементираат ветувачки пристапи за машинско учење од литературата и се оценуваат нивните перформанси. Презентираното истражување има за цел да ги подобри предвидувањата за квалитетот на воздухот и да го зголеми знаењето за загадувањето на воздухот во градот Скопје. Скопје е град со зголемена загаденост на воздухот, со особено голем проблем во зимските месеци. Ова предизвикува сериозни последици по здравјето на жителите и зафаќа различни аспекти на животот во градот. Преку прецизни предвидувања за нивоата на загаденост на воздухот, јавноста и владите можат да преземат соодветни мерки и одлуки, како што се превентивни активности за намалување на загадувањето, чистење на прашина и обесхрабрување на активности на отворено, како и ублажување на штетните последици од аерозагадувањето. Во ова прво поглавје, даваме преглед на предизвиците поврзани со загадувањето на воздухот во градот Скопје и го образложуваме мотивот зад ова истражување. Исто така, објаснуваме целта на истражувањето и методот што се користи во процесот на анализа и предвидување. На крајот, претставуваме резиме на придонесите на ова истражување и наведуваме контурите на структурата на целата теза. Со ова истражување се надеваме да донесеме поблиска историја и разбирање за загадувањето на воздухот во градот Скопје. Преку анализата на различни фактори и атрибути, како што се метеоролошки услови, население и антропогените активности, се надеваме да го идентификуваме главниот извор на загадување и да го разбереме неговиот влијание врз квалитетот на воздухот.

# Цели и истражувачки прашања

Целта е да се користат најсовремените резултати од истражувањето со посебен фокус на методите за машинско учење за предвидување на квалитетот на воздухот, да се проценат и применат идеите и алгоритмите од овие студии во контекст на предвидување временски серии. Крајната цел на предвидувањето за квалитетот на воздухот е да им овозможи на националните и глобалните носители на одлуки, заедниците и поединците да проактивно преземање мерки за намалување на здравствените опасности предизвикани од аерозагадувањето.

**A magnifying glass on a clipboard with graphs and charts

Description automatically generated with medium confidence**

# Топлотната мапа на корелација

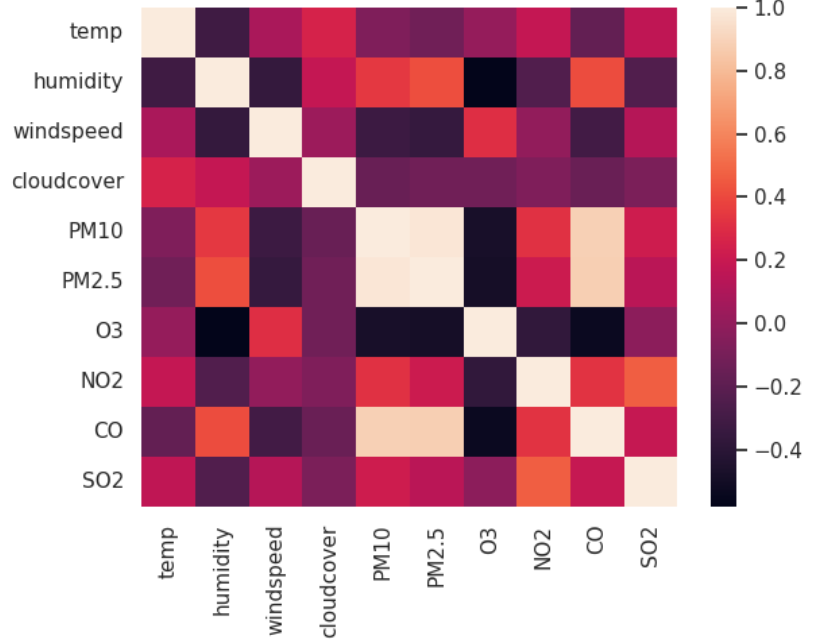
Корелацијата го мери степенот на линеарна врска помеѓу две променливи и може да земе вредности од -1 до 1.

Кога корелацијата е 1, тоа покажува дека две променливи се позитивно корелирани, што значи дека кога едната променлива расте, и другата обично расте. На пример, ако имаме променливата X која претставува ниво на загаденост на воздухот и променливата Y која претставува број на респираторни заболувања, ако корелацијата помеѓу нив е 0.8, тоа значи дека кога има поголемо ниво на загаденост на воздухот, постои поголем ризик од заболувања на респираторниот систем. Ако имаме променливата Z која претставува количина на штетни гасови и променливата W која претставува број на случаи на алергии, ако корелацијата помеѓу нив е -0.6, тоа значи дека кога има поголема количина на штетни гасови, може да има намалување на случаите на алергии.

Кога корелацијата е -1, тоа покажува дека две променливи се негативно корелирани, што значи дека високите вредности на PM10 и PM2.5 се негативно корелирани со ниските вредности на NO2 и CO. Односно, кога концентрацијата на честички во воздухот е висока, концентрацијата на азотен диоксид и јаглероден моноксид обично е ниска и обратно.

Кога корелацијата е 0, тоа покажува дека две променливи не се корелирани и нема линеарна врска помеѓу нив. Тоа не значи дека нема други видови на врски помеѓу променливите, само што не постои линеарна зависност помеѓу нив.

Топлотната мапа преку боите и интензитетот на боите ни овозможува да визуелизираме и прегледаме оваа корелација помеѓу променливите во податочната структура. Вака можеме да идентификуваме кои променливи се силно корелирани, слабо корелирани или некорелирани, што ни помага во анализата.



# Pair Plot

Pair Plot, познат и како Scatter Matrix, е графички метод за прикажување на врска помеѓу повеќе променливи во еден дијаграм. Оваа визуелизација е корисна за анализа на корелацијата и распределбата на податоците помеѓу променливите.Во случајот на загадувањето на воздухот во Скопје, можеме да искористиме Pair Plot за да истражиме врските помеѓу различни аспекти на загадувањето на воздухот како PM10, PM2.5, Брзина на воздхот, Температурата. На главната дијагонала на графиконот, ќе бидат прикажани хистограми кои ни даваат информација за распределбата на секоја променлива поединечно. Со тоа можеме да видиме колку често се појавуваат одредени вредности и како се распределуваат.На горниот и долните правоаголници на графиконот, ќе бидат прикажани scatter plots кои ни даваат информација за врската помеѓу две променливи. Преку овие графици, можеме да видиме дали постои некоја видлива корелација или зависност помеѓу различните аспекти на загадувањето на воздухот.

**A picture containing diagram, line, electric blue, blue

Description automatically generated**

# График на распределба

Поминувањето на графикот на распределба го прикажува количеството на PM10 емисии на хоризонталната оска (PM10 Emissions) и вертикалната оска (Probability Density). Во овој случај, вертикалната оска го претставува процентот или веројатноста за појава на одредена вредност на емисиите на PM10.Графикот ни дава представа за распределбата на PM10 емисиите во воздухот на градот Скопје. Можеме да забележиме како емисиите на PM10 се концентрираат со пик на графикот. Ова може да ни даде информација за типичните нивоа на загадување на воздухот со PM10 честички во градот Скопје.

**A picture containing plot, diagram, line, text

Description automatically generated**

# Box plot

Овој график е боксплот (box plot) кој прикажува распределбата на емисиите на PM10 (честички со пречник до 10 микрометри) во воздухот на градот Скопје според различни услови или состојби (conditions).

Графикот е поделен на различни категории на услови (conditions) на хоризонталната оска, а емисиите на PM10 се прикажани на вертикалната оска. Во овој случај, боите на кутиите и линиите на графикот се дефинирани со палетата на бои "husl", што овозможува различните услови да бидат различно прикажани и разграничени.

Боксплотовите ни даваат информации за различните статистички величини на емисиите на PM10 во различните услови. За секоја категорија на услови, боксплотот ни покажува медијаната (линијата во средината на кутијата), интерквартилниот опсег (квартилите Q1 и Q3 кои ги ограничуваат кутијата) и потенцијални некорелирани вредности („промашки“ или „скали“) надвор од интерквартилниот опсег (тенките линии на графикот).

Овој график ни овозможува да споредиме емисиите на PM10 помеѓу различните услови и да идентификуваме можни разлики или варијации во нивото на загадување на воздухот. Можеме да забележиме дали има разлика во медијаната и интерквартилниот опсег на емисиите на PM10 помеѓу различните услови, што ни дава представа за нивото на загадување на воздухот во различни ситуации.

A picture containing text, diagram, screenshot, plot

Description automatically generated

# Scatter plot за Линеарна Регресија

На x-оската (horizontal axis) е прикажана вистинската вредност на загадувањето на PM10 (Measured), додека на y-оската (vertical axis) е прикажана предвидената вредност на загадувањето на PM10 (Predicted). Секој точка на графикот претставува едно мерење/инстанца во тест множеството.

Овој график служи за да го прикаже колку добро линеарната регресија предвидува вистинските вредности на загадувањето на PM10. Ако предвидените вредности се согласуваат со вистинските вредности, тогаш ќе имаме точки кои лежат приближно на дијагоналната линија. Ако има разлика помеѓу предвидените и вистинските вредности, тогаш точките ќе бидат распоредени околу дијагоналната линија со различни отклонувања.

Овој график овозможува да прецениме некои статистички метрики за оценка на моделот. Некои од метриките што се прикажани во овој пример се:

* Оценка на линеарната регресија:Го прикажува коефициентот на детерминација (R-squared) кој го мери објаснувањето на варијацијата на вистинските вредности со предвидените вредности. Вредноста блиску до 1 означува добро предвидување.
* Средна апсолутна грешка: Го пресметува просечниот апсолутен збир на разликите помеѓу вистинските и предвидените вредности. Што помалку е вредноста на оваа метрика, тоа покажува дека моделот има помалку грешки во предвидувањето на загадувањето на PM10.

A graph with blue dots

Description automatically generated with low confidence

# Random Forest

Случајни шумови е заслужен за споредба помеѓу вистинските вредности и предвидените вредности на загадувањето на PM10.Графикот е составен од точки кои претставуваат инстанците, каде што по X-оската се прикажани вистинските вредности, а по Y-оската се прикажани предвидените вредности од моделот. Ако точките се распоредени околу дијагоналната линија, тоа покажува дека предвидувањето од моделот е приближно точно. Сепак, ако точките се распоредени далеку од дијагоналната линија, тоа може да укажува на значителни грешки во предвидувањето.

A graph with blue dots

Description automatically generated with low confidence

# Споредба помеѓу Општини во Скопје

**A picture containing text, screenshot, plot, diagram

Description automatically generated**

Споредба на ПМ2.5 честици помеѓу општините Центар,ГазиБаба,Карпош и др.  
id е бројот на сите редови во дата базата која што има повеќе од 90 000 записи. A picture containing screenshot, plot, text, diagram

Description automatically generated

Споредба на NO2 од 2008 до 2018 од 90 000 записи

A picture containing text, screenshot, plot, diagram

Description automatically generated

Споредба на ПМ10 Честички од 2008 до 2018 на 90 000 записи

# Заклучок

Загадувањето на воздухот во градот Скопје е сериозен проблем со потенцијални негативни последици по здравјето и животната средина. За да се намали загадувањето, потребни се систематски напори и мерки за намалување на емисиите од различни извори, подобрување на атмосферските услови. Преку примена на алгоритми и модели за машинско учење, можеме да анализираме големи количества податоци за квалитетот на воздухот, атмосферски услови, емисии и други фактори кои влијаат врз загадувањето. Според овие податоци, моделите можат да ги предвидат нивоата на загадување во реално време или во иднина. Ова ни овозможува да преземеме превентивни мерки и да развиеме ефикасни стратегии за намалување на загадувањето.

Користена Литература

* <https://github.com/petarfinki/Skopje-Air-Quality-Prediction>
* <https://www.kaggle.com/datasets/cokastefan/pm10-pollution-data-in-skopje-from-2008-to-2018>