Анализа података- предикција загађења ваздуха PM 2.5 граду Шен-јанг

Алекса Шкрбић, IN29/2018, aleksa.shrbic@gmail.com

I Увод

Извештај се бави анализом података везаних за концентрацију PM 2.5 честица (када су повишене смањују видљивост и узрокују замагљење у ваздуху). То су честице са пречником мањим од 2.5 микрометара, што значи да су око 3% пречника људске длаке. Настају под утицајем рада фабрика (најчешће у области металургије), моторних возила, пожара итд. Због своје мале величине, људи и животиње их лако уносе дисајним путевима, веома брзо и лако долазе до плућа и узрокују кардиоваскуларне и пулмоналне болести. Анализом нама доступних података који утичу на концентрацију ових честица у ваздуху, могуће је креирати модел за њихову предикцију, што може резултирати смањењу ових честица и омогућити безбеднији и квалитетнији живот.

II База података

База података садржи податке о 52584 узорака и 17 обележја. Постоји 11 нумеричких обележја, а то су: редни број мерења, концентрација PM 2.5 честица на три 3 различите локације, температура росе, влажност ваздуха, дневна температура, ваздушни притисак, брзина ветра, падавине на сат и укупне падавине. Ови параметри су мерени сваког сата, у временском интервалу од 5 година (2010. - 2015) .
Поред нумеричких, постоји и 6 категоричких обележја. Њих чине година, месец, дан, сат, сезона (годишње доба), правац ветра.

III Анализа података

Анализом података уклоњена су обележја за измерену количину PM 2.5 честица на следећим локацијама: Тајухан и Ксајохен (по услову задатка, мада су оба обележја имала више од 50% недостајућих вредности).
За обележје ,‘‘PM_US Post‘‘ односно концентрацију PM 2.5 честица на овој локацији недостајало је око 58% вредности обележја(што би било 3 године и 3 месеца), па је самим тим нелогично попуњавати ове вредности, поготово ако знамо да се концентација ових честица мења из године у годину, тако да су овде недостајући подаци елиминисани. Последња 2 обележја падавине на сат (‘‘precipitation‘‘) и укупне падавине (‘‘Iprec‘‘), поседују око 5% недостајућих вредности (након уклањања недостајућих вредности обележја ‘‘PM_US Post‘‘).Ове и остале недостајуће вредности других обележја попуњене су првом претходном вредношћу тога обележја. Након “data cleaninga”, тј чишћења података, сада имамо 21679 узорака и 14 обележја.

А. Категоричка обележја

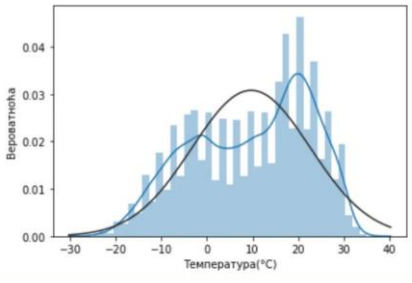
Једино категоричко обележје у бази које садржи ненумеричке вредности обележја, јесте правац ветра (“cbwd”), тако да је ту извршено превођење у нумеричке вредности, тако што је за сваки правац ветра, постављен број од 1-5.

Б. Анализа температуре

Годишње доба	Минимална температура	Просечна температура	Максимална температура
Пролеће	-14°C	12,26°C	35°C
Лето	10°C	23,29°C	35°C
Јесен	-18°C	10 °C	31°C
Зима	-25°C	-7,13 °C	15°C

Табела1: Приказ минималних, просечних и максималних температура у граду Шенг-Јанг, током годишњих доба.

Из табеле 1, можемо видети да је у граду Шенг-јанг, заступљена континентална клима, где су зиме изразито хладне, лета блага и умерено топла уз повремен пораст температуре и преко 30°C, док су пролећа и јесени прохладни.



Слика1: Поређење расподеле просечних дневних температура са функцијом нормалне расподеле

Можемо уочити да је расподела просечних дневних температура спљоштенија у односуна ф-ју нормалне расподеле, а такође уочавамо негативан коефицијент асиметрије који износи -0,39, тј благо искривљене у десно.

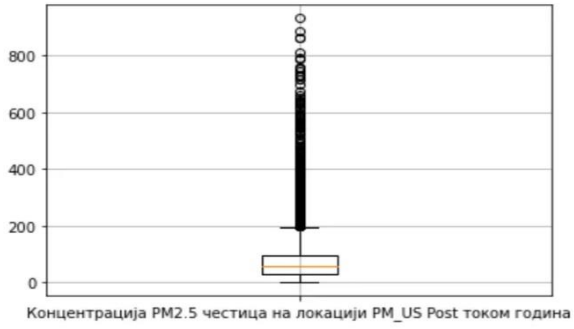
В. Основне статистике нумеричких обележја

Из статистичких величина обележја, можемо видети да за скоро све вредности обележја постоје изузетно високе вредности које се ретко појављују, аутлајери(Outlieri). За ''сумњиве вредности неких обележја'', најбоље би било консултовати се са стручњацима. На пример, ако посматрамо обележје Јачина ветра ("Iws"), вредности, тј брзине ветра се углавном налазе у интервалу од 3 до 24 ms^-1. Међутим, приметно је да постоје брзине од преко 400 ms^-1 (1440 kmh^-1), што је брзина за око 5 пута јача од брзине урагана '''Катрина'' који је 2005. погодио САД. Тако да можемо претпоставити да је у питању грешка приликом уноса, мада свакако би се требало консултовати са стручњаком. Обележја везана за падавине, брзину ветра, ваздушни притисак као и PM2.5 честице имају леве асиметричне расподеле, док обележја везана за температуру и влажност ваздуха имају десне асиметричне расподеле.

Г. Анализа концентрације PM2.5 честица на локацији PM_US Post

Интерквартилни опсег	31- 97
Медијана	57
Максимална вредност	932

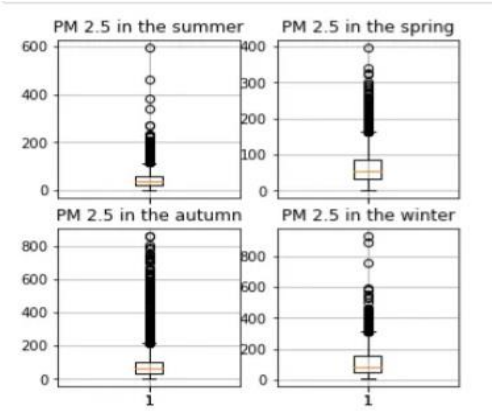
Табела 2. Статистиче величине обележја PM2.5



Слика 2. Box-plot за концентрацију PM2.5 честица на локацији PM_US Post у периоду од 2013-2015. године.

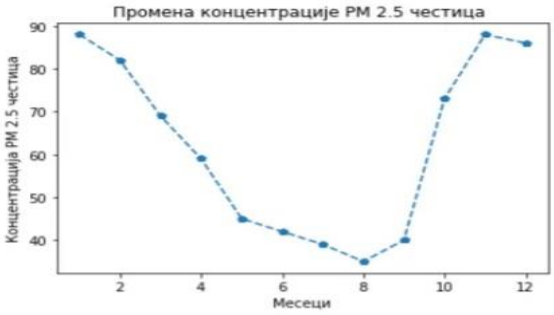
Из овог boxplot-a , да су аутлајери заступљени у великој мери(постоје на високим, а не постоје на ниским вредностима). Медијана износи 57*μg(m^3)^-1* за концентрацију ових честица, а горњи квартил је 97. Забрињавајуће на овој слици је то, да постоје вредности и преко 800 μg(m^3)^-1,што је иразито висока бројка*, а* максимум *је* 882.

Г. 1. Анализа концентрације PM2.5 честица по годишњим добима



Слика3. Концентрација PM2.5 честица током годишњих доба

На boxplotovima најбоље се види медијана, интерквартилни опсег као и аутлајери. Можемо уочити да је највећа концентрација ових честица измерена зими, због недостатка Сунчеве светлости и велике потрошње фосилних горива (видели смо да су зими овде изразито хладне). Најнижа количина PM2.5 честица заступљена је током лета износи око 40 μg(m^3)^-1, из разлога што су лета блага са просечном температуром од око 23°C . Мада сличну медијану овог обележја, има и пролеће. У сваком годишњем добу постоје аутлајери, а најзаступљенији су у јесен, док је највећа вредност од 932 μg(m^3)^-1 , забележена у зиму као последица велике потрошње фосилних горива.



Слика 4. Медијана обележја PM2.5 честица по месецима

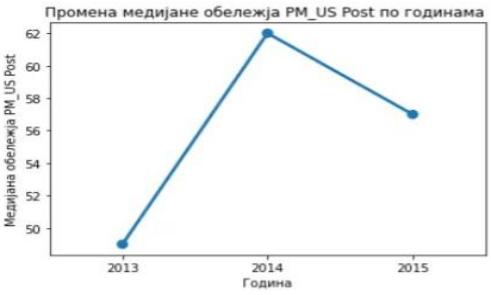
На 4. слици још боље можемо уочити понашање ових честица кроз одређене годишње периоде. Јасно се види како је током зиме присутна највећа концентрација ових честица и како лагано опада до пролећа, па све до августа када достиже свој глобални минумум. Потом, како креће јесен и приближавамо се зимским месецима, ф-ја креће да расте до свог глобалног максимума.

Г. 2. Анализа концентрације PM2.5 честица по годинама

Статистчке величине обележја	2013	2014	2015
IQR опсег	27-86	35-101	31-98
Медијана	49	62	57
Максимум	583	725	932

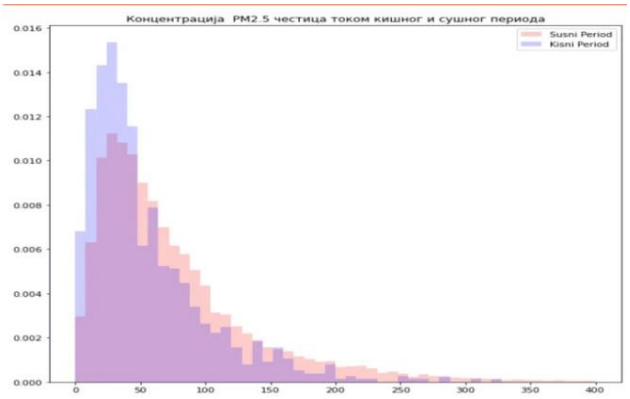
Табела 3. Статистичке вредности PM2.5 честица током година

Посматрајући ову табелу и слику 5 (уколико изузмемо 2013-ту годину) из разлога што нам фале подаци за прва 3 месеца, када је концентрација ових честица највиша због зимског периода , уочавамо да медијана опада, што може указивати на бољи квалитет ваздуха.Међутим ако посматрамо IQR опсег видећемо да су вредности поприлично сличне, чак се и највећа вредност овога обележја налази налази у 2015-тој (расте по годинима) што нам баш и не даје оптимистичне податке у вези са загађењем ваздуха.



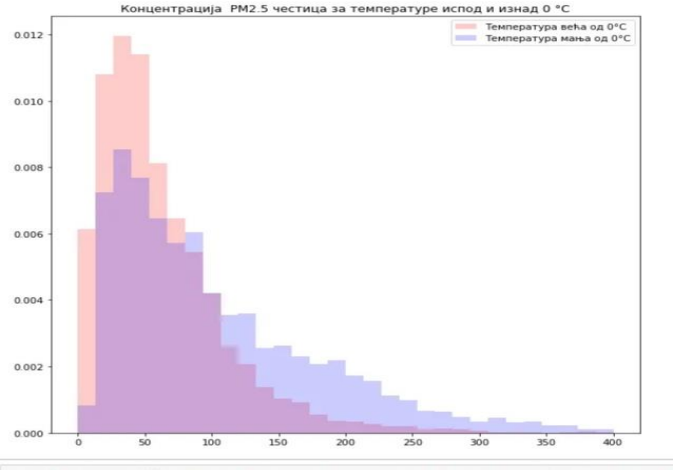
Слика 5. Промена медијане обележја PM_US Post по годинама.

Г .3. Утицај кише и температуре на концентрацију PM2.5 честица



Слика 6. Утицај падавина, тј кишног и сушног периода на концентрацију PM2.5 честица

Такође разматран је утицај како кишног тако и сушног периода на концентрацију PM2.5 честица током година. Уколико је присутан сушни период (укупне падавине су једнаке нули), концентрација ових честица је најчешће око 30 μg(m^3)^-1, са мањом вероватноћом између 100 и 200 μg(m^3)^-1 и са готово занемарљивом вероватноћом од преко 200 μg(m^3)^-1 и уочава се нормална расподела. Током кишног периода такође је присутна нормална расподела са благим искривљењем у десно. Концентрација PM2.5 честица је између 10 и 40 μg(m^3)^-1, а приметна је и мања вероватноћа између 100-150 μg(m^3)^-1, а све преко тога је занемарљиво.



Слика 7. Утицај температуре на концентрацију PM2.5 честица (када је температура мања, односно већа од 0 °C)

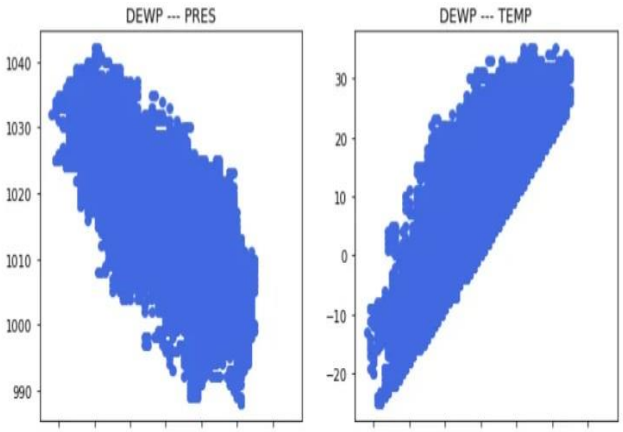
Са слике 7, где смо разматрали утицај позитивне и негативне температуре, можемо видети да се добија сличан график претходноме, односно да график ''Утицаја падавина'' кореспондентира са графиком ''Утицаја температуре''.

Д. Корелација између обележја

Слика 8. Топлотна мапа (heatmap) која показује корелацију између нумеричких обележја

Применом функције **corr** утврђени су парови обележја и њихове корелације, односно међусобне зависности.

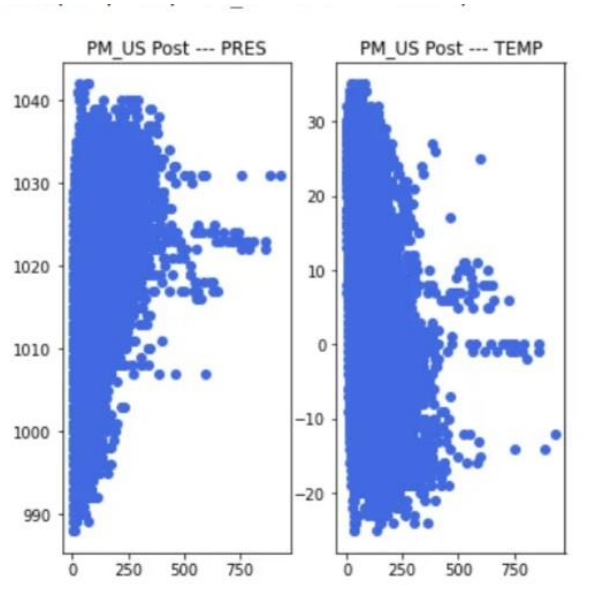
Највећа корелација јавља се међу обележјима температуре росе и дневне температуре (0,88) и она је позитивна. Такође приметне су корелације између температуре росе и ваздушног притиска (0,78 и то је негативна корелација), а може се видети и корелација између падавина на сат и укупних падавина.



Слика 9. Корелација температуре росе са ваздушним притиском (слика лево) и корелација са дневном температуром (слика десно)

Са слике лево видимо да порастом ваздушног притиска опада температура росе (кондензације), а са слике десно да температура росе тј кондензације расте заједно са порастом дневном температуром.

Са топлотне мапе (слика 8), видимо да је концентрација PM2.5 честица у највећој корелацији са дневном температуром (TEMP) , ваздушним притиском (PRES) и температуром росе (DEWP). Овакав корелација сматра се слабом, тј незнатном линеарном корелацијом.



Слика 10. Корелација PM 2.5 честица са ваздушним притиском (слика лево) и дневном температуром (слика десно)

Из свега приложеног до сада, уочавамо да PM2.5 честице имају слабу корелацију са осталим обележјима, што нам даје да закључимо да атмосферски услови немају велики утицај на концентрацију ових честица у ваздуху. тј ове честице су више осетљиве на разне друге врсте загађивача ваздуха, као што су фосилна горива, керозин и остали издувни гасови.

IV Линеарна регресија

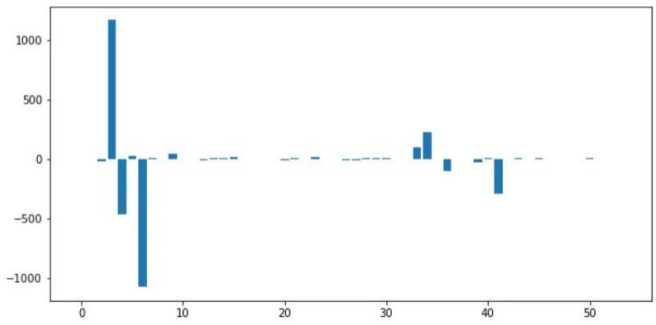
Да бисмо кренули са предикцијом концентрације PM 2.5 честица, почетни скуп узорака неопходно је поделити на два подскупа, један за обуку, а други за тестирање. Скуп за обуку садржи 90% узорака (на којима ћемо обучавати наш модел), док скуп за тестирање садржи 10% насумично одабраних узорака на којима ћемо тестирати успешност нашег модела. Пре но што смо поделили скуп узорака, из њега је било неопходно избацити категоричка обележја и вредности обележја која су већински близу нули. Од категоричких обележја одбачено је само обележје ''Year'',односно година, а од нумеричких (precipitation) Падавине на сат и укупне падавине (Iprec). Остала категоричка и нумеричка обележја нису одбачена из разлога зато што је оваква комбинација дала најбоље перформансе (од осталих испробаних) и омогућила да се добијају боље предиктивне вредности у односу на циљне.

За обуку модела коришћено је више приступа, међутим најбоље вредности приликом тестирања показала је Линеарна регресија са хипотезом интеракције и квадрата.

Мера успешности регресора	Израчуната вредност
Средња квадратна грешка	3359,06
Средња апсолутна грешка	37,25
Корен средње квадратне грешке	57,95
R^2 вредност	0,328
R^2 прилагођена вредност	0,326

Табела 4. Мера успешности линеарне регресије са хипотезом интеракције и квадрата

Из табеле 4, за нашу R^2 вредност можемо рећи да је она слаба, мада то смо и могли очекивати зато што PM2.5 нема јаку корелацију са осталим обележјима. Овакав модел покрива око 33% укупне варијансе. Испробана је "Lasso" и "Ridge" регресија у циљу постизања боље R^2 вредности, али безуспешно, јер није направљен помак у односу на линеарну регресију са хипотезом интеракције и квадрата, па је овај модел остао коначан.



Слика 11. Илустрација коефицијената линеарне регресије са хипотезом интеракције и квадрата.

Са слике 11, уочавамо да је већина коефицијената сличних вредности, али постоје и вредности коефицијената које досежу и до 1000 односно и до -1000, а они у великој мери утичу (негативно) на тест скуп.

V Закључак

Анализом базе података и посматрањем концентрације PM 2.5 честица у ваздуху, јасно је да није довољно посматрати само временске услове да бисмо дошли до закључка у вези са штетним материјама које се налазе у ваздуху. Можемо претпоставити да лагано уништавање озонског омотача и глобално загревање доводе до повећања броја ових честица. У уводу је напомењено да су ове честице изразито малих димензија и самим тим лако долазе до свих живих бића и изазивају разне врсте кардиоваскуларних и пулмоналних болести. Да бисмо добили боље резултате линеарне регресије, неопходно би било посматрати рад фабрика,метало индустрије као и сагоревање осталих штетних материја. Препорука за становнике Шенг-јанга била би да носе заштитне маске (највише зими, када је најјача концентрација ових честица), а особе које имају пулмоналне и кардиоваскуларне тегобе би требале да буду посебно обазриве и да се чувају нарочито у зимском периоду.