Квалитет на воздухот според часовни мерења од пет метал-оксидни сензори и референтни анализатори: Италијански урбан контекст (2004–2005)

Проект по Податочно рударство

Лина Панева

Скопје, 2025

Содржина

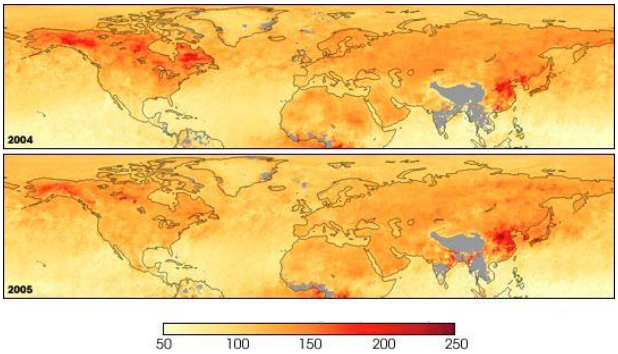
Вовед

Загадувањето на воздухот претставува еден од најсериозните еколошки и здравствени предизвици во современото општество. Урбанизацијата, сообраќајот, индустриските активности и затоплувањето на домовите значително придонесуваат за влошување на квалитетот на воздухот, особено во густо населени подрачја. Изложеноста на високи концентрации на **штетни гасови**, како јаглерод моноксид (CO), азотни оксиди (NOx), азот диоксид (NO_2) и бензен, директно се поврзува со појава на респираторни и кардиоваскуларни заболувања, намалување на квалитетот на живот и предвремена смртност. Поради тоа, континуираното следење и анализа на загадувањето на воздухот е клучно за заштита на јавното здравје и за изработка на ефикасни политики за животна средина.

Податочното множество **Air Quality**, достапен преку UCI Machine Learning Repository, претставува значајна основа за вакви анализи. Тој содржи повеќе од **9.300** часовни мерења собрани во периодот од март 2004 до февруари 2005 година во урбана средина во Италија. Податоците ги комбинираат референтните мерења од сертифициран анализатор со податоци од **пет метал-оксидни сензори**, како и метеоролошки параметри – температура, релативна и апсолутна влажност. Ова овозможува не само проценка на квалитетот на воздухот, туку и анализа на односот помеѓу загадувачите и климатските услови, како и развој на предиктивни модели со примена на статистички и машинско-учечки методи.

Особено значење има фактот што ваквите сензори често се соочуваат со проблеми како постепено поместување на сигналот со текот на времето и меѓусебна чувствителност на различни гасови, што ја нагласува важноста од истражувања во областа на калибрација на сензори и предвидување на загадувањето врз основа на сензорски податоци.

Во рамки на овој проект, **Air Quality податочното множество** ќе се искористи за претпроцесирање и чистење на податоците, визуелизација, како и примена на техники од податочно рударство, како што се анализа на корелациис, регресија, кластеризација, со цел да се идентификуваат скриени шеми и трендови во податоците. Со комбинирање на методи од машинско учење и класичка статистика, ќе се добијат информации кои можат да помогнат во оценка на квалитетот на воздухот, подобрување на точноста на сензорите и поддршка при креирање политики за одржлив урбан развој и заштита на животната средина.



Слика 1. Загаденост на воздухот на глобално ниво 2004-2005

Преглед на податоци

Податоците од множеството **AirQuality** се собрани во периодот од март 2004 до февруари 2005 година во **урбана средина во Италија**, и истото содржи **9.357 редици** и 15 колони, односно толку набљудувања и атрибути. Податоците се резултат на мерења од **референтни анализатори**, кои даваат точни концентрации на загадувачи, како и од **теренски сензори**, кои реагираат на присуството на специфични гасови. Дополнително, податочното множество вклучува и **метеоролошки параметри** како температура, релативна и апсолутна влажност, кои овозможуваат анализа на зависностите помеѓу загадувачите и климатските услови.

Типови на променливи

- Датум и време:
  - Date *(Date)*
  - Time *(Categorical)*
- Референтни мерења (точни концентрации):
  - CO(GT) – јаглерод моноксид (mg/m³) *(Integer)*
  - NMHC(GT) – неметански јаглеводороди (µg/m³) *(Integer)*
  - C6H6(GT) – бензен (µg/m³) *(Continuous)*
  - NOx(GT) – азотни оксиди (ppb) *(Integer)*
  - NO2(GT) – азот диоксид (µg/m³) *(Integer)*
- Сензорски мерења:
  - PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O_3) – часовни просечни реакции на метал-оксидни сензори, номинално насочени кон одредени загадувачи *(Categorical)*
- Метеоролошки променливи:
  - T (температура, °C) *(Continuous)*
  - RH (релативна влажност, %) *(Continuous)*
  - AH (апсолутна влажност) *(Continuous)*

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
0	2004-03-10	18:00:00	2.6	1360.00	150	11.881723	1045.50	166.0	1056.25	113.0	1692.00	1267.50	13.60	48.875001	0.757754
1	2004-03-10	19:00:00	2.0	1292.25	112	9.397165	954.75	103.0	1173.75	92.0	1558.75	972.25	13.30	47.700000	0.725487
2	2004-03-10	20:00:00	2.2	1402.00	88	8.997817	939.25	131.0	1140.00	114.0	1554.50	1074.00	11.90	53.975000	0.750239
3	2004-03-10	21:00:00	2.2	1375.50	80	9.228796	948.25	172.0	1092.00	122.0	1583.75	1203.25	11.00	60.000000	0.786713
4	2004-03-10	22:00:00	1.6	1272.25	51	6.518224	835.50	131.0	1205.00	116.0	1490.00	1110.00	11.15	59.575001	0.788794

Слика 2. Претстава на првите 5 записи од податочното множество AirQuality

Предизвици

Во податочното множество постојат вредности кои бараат соодветна обработка и филтрирање пред анализа, што така бара внимателен пристап при претпроцесирањето.

Оваа секција обезбедува целосна слика на податочното множество и ги поставува основите за **претпроцесирање**, **визуелизација и процесирање**, кои ќе следат во следните делови на проектот.

Претпроцесирање на податоците

Претпроцесирање на податоците е почетна и многу важна фаза во **анализа на податоци и машинско учење**. Целта е суровите податоци да се подготват за понатамошна обработка и моделирање. Во овој процес се вклучуваат активности како **чистење на податоците**, **трансформација**, **енкодирање на категоријални променливи** итн. Со правилно претпроцесирање се подобрува **квалитетот на податоците и точноста на моделите**.

Недостасувачки вредности: приказ и обработка

Командата *df.info()* е команда од **pandas** библиотеката во Python и се користи за прикажување на основни информации за *DataFrame*. При нејзино извршување врз податочното множество се добиваат информации за број на редови и колони, имиња на колоните, типови на податоци и број на ненедостасувачки вредности (Слика 3).

```
RangeIndex: 9357 entries, 0 to 9356
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Date            9357 non-null   object
 1   Time            9357 non-null   object
 2   CO(GT)          9357 non-null   float64
 3   PT08.S1(CO)     9357 non-null   int64
 4   NMHC(GT)        9357 non-null   int64
 5   C6H6(GT)        9357 non-null   float64
 6   PT08.S2(NMHC)   9357 non-null   int64
 7   NOx(GT)         9357 non-null   int64
 8   PT08.S3(NOx)    9357 non-null   int64
 9   NO2(GT)         9357 non-null   int64
 10  PT08.S4(NO2)    9357 non-null   int64
 11  PT08.S5(O3)     9357 non-null   int64
 12  T               9357 non-null   float64
 13  RH              9357 non-null   float64
 14  AH              9357 non-null   float64
dtypes: float64(5), int64(8), object(2)
```

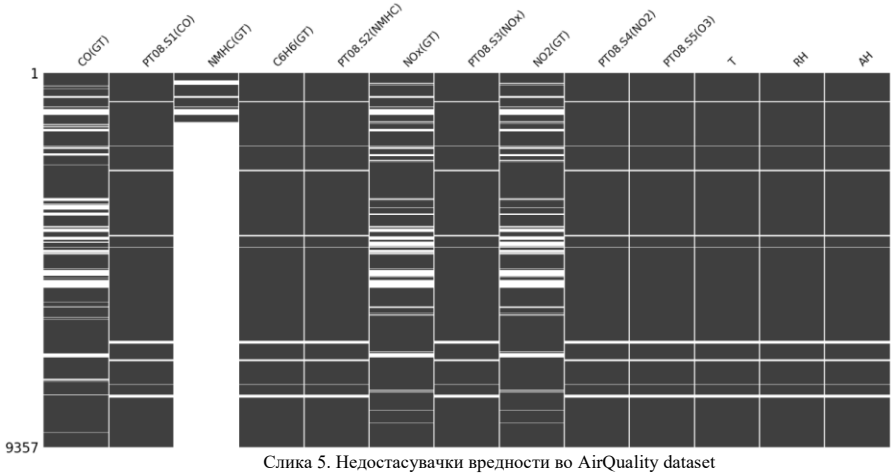
Слика 3. Резултат од df.info()

Од излезот може да се согледа дека има три типа на податоци и дека **нема недостасувачки вредности**. Сепак при детална анализа на множеството, се забележува дека сепак таквите вредности постојат, само се означени со **-200.0** и **-200** наместо *NaN* (како што е и наведено во описот на множеството). Начин за справување со ова е само тие вредности да се заменат со *NaN* вредности од библиотеката **numpy**. По оваа постапка, може јасно да се согледа присуството на недостасувачки вредности, чиј број е прикажан на Слика 4. Вкупниот број на истите е **16.701**.

CO(GT)	1683
PT08.S1(CO)	366
NMHC(GT)	8443
C6H6(GT)	366
PT08.S2(NMHC)	366
NOx(GT)	1639
PT08.S3(NOx)	366
NO2(GT)	1642
PT08.S4(NO2)	366
PT08.S5(O3)	366
T	366
RH	366
AH	366

Слика 4. Број на недостасувачки вредности во AirQuality dataset

За појасна претстава на недостасувачките вредности, истите можат да се визуализираат. Еден начин е преку **матрица од библиотеката missingno**, каде **белите простори** ги означуваат овие вредности (Слика 5).



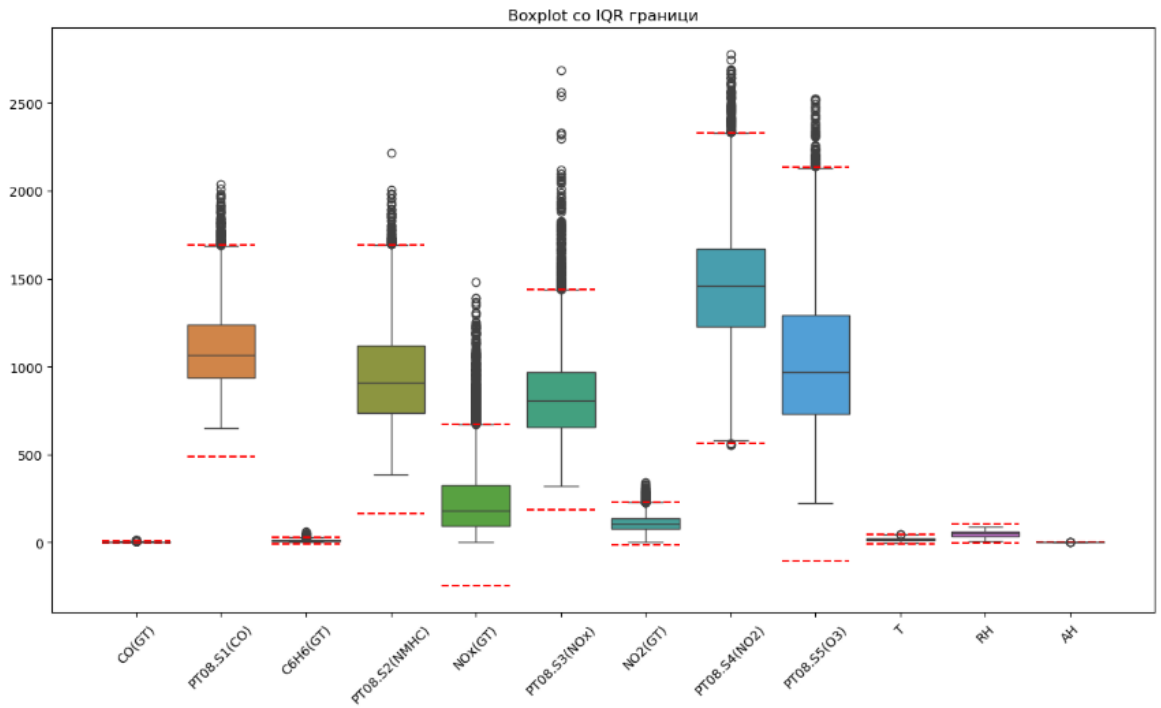
Слика 5. Недостасувачки вредности во AirQuality dataset

Како што може да се согледа, во колоната **NMHC(GT)** бројот на недостасувачки вредности е многу голем. Примената на **праг од 80% недостасувачки вредности** овозможува **отстранување на колони** кои содржат недоволно информации. На овој начин се подобрува **квалитетот на податочното множество** и се овозможува поефикасно **претпроцесирање и анализа**.

Пополнувањето на недостасувачките вредности се извршува со **интерполација**, што претставува метод за пресметка на недостасувачките податоци врз основа на постоечките вредности во соседството. Во овој проект е применет **временски метод на интерполација,** кој ја користи **временската зависност на податоците** за точно предвидување на недостасувачките вредности. Со тоа се овозможува **зачувување на трендовите и динамиката на мерењата** во податочното множество.

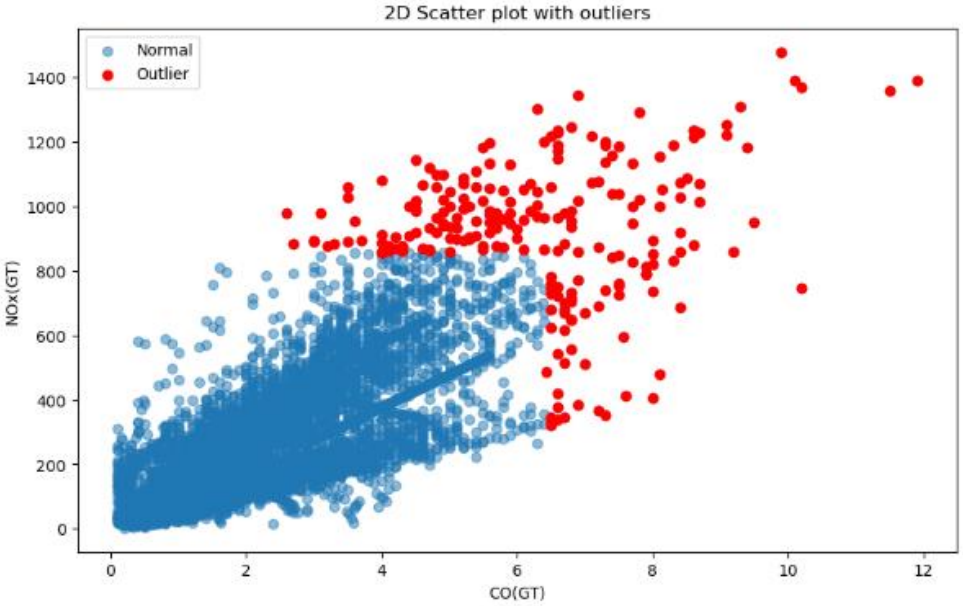
Екстремни вредности (Outliers): приказ и обработка

За согледување на **екстремните вредности (outliers)** во податочното множество е избран **boxplot**. Овој вид на визуелизација ги прикажува **медијаната**, **квартилите и распонот на податоците** (означен со црвени испрекинати линии), а вредностите кои се надвор од интеркватилниот распон се истакнуваат како точки, што овозможува лесно **идентификување на екстремните вредности**. Boxplot е интуитивен и ефикасен метод за анализа на една или повеќе променливи, овозможувајќи брзо согледување на потенцијални аномалии.

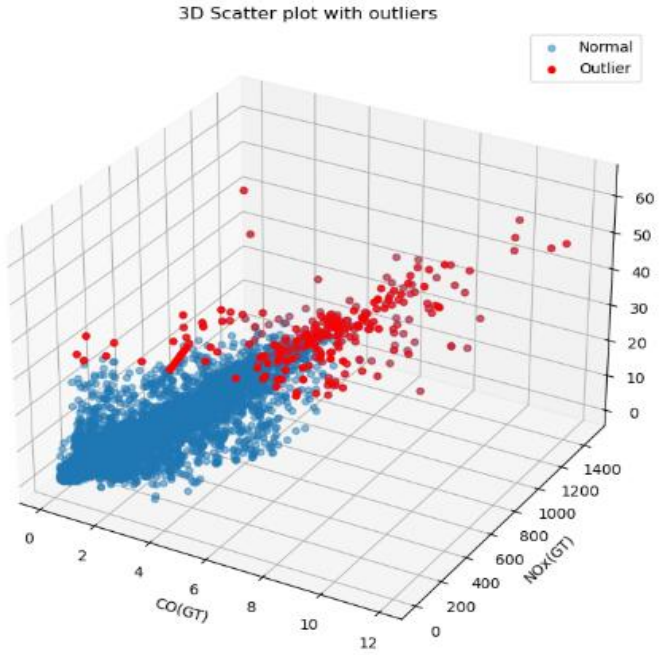


Слика 6. BoxPlot врз AirQuality dataset

Бројот на екстремните вредности може да се согледа користејќи го **z-score** од библиотеката **stats**. Тоа претставува мерка на тоа колку **стандардни девијации** е една вредност оддалечена од средната вредност на податоците. Со други зборови, го покажува колку една вредност се разликува од „просекот“ на распределбата. Вредности со **z-score поголем од 3** (или помалку од -3) се сметаат за екстремни, бидејќи се значително оддалечени од просечните вредности. Во анализираното податочно множество такви вредности има **747**.



Слика 7. 2D Scatter plot на CO(GT) и NOx(GT) со идентификувани outliers



Слика 8. 3D catter plot на CO(GT) , NOx(GT) и C6H6(GT) со идентификувани outliers

Иако се идентификуваат **екстремни вредности** во податочното множество, тие не се отстрануваат**.** Причината е што во мерењата на **гасови и загадувачи** е очекувано да постојат високи или ниски концентрации во одредени моменти, што се смета за **природна варијабилност на податоците.** Отстранувањето на овие вредности би можело да ги изгуби важните информации за **екстремни услови во урбаната средина**, кои се релевантни за анализа на загадувањето.

Уникатни вредности и дупликати

Уникатни вредности претставуваат број на различни вредности присутни во секоја колона на податочното множество. Со анализа на **Слика 9** се согледува процентуалното застапување на уникатните вредности во секоја колона.

```
CO(GT): 15.34% unique values
PT08.S1(CO): 14.27% unique values
C6H6(GT): 8.19% unique values
PT08.S2(NMHC): 17.04% unique values
NOx(GT): 24.89% unique values
PT08.S3(NOx): 16.60% unique values
NO2(GT): 16.77% unique values
PT08.S4(NO2): 20.53% unique values
PT08.S5(O3): 21.99% unique values
T: 8.52% unique values
RH: 11.67% unique values
AH: 75.27% unique values
```

Слика 9. Процентуално застапување на уникатни вредности по колона во AirQuality dataset

Дупликати се редови кои се идентични во сите колони. При анализата на бројот на дупликати се согледува дека податочното множество **не содржи дуплирани редови**, што овозможува точна и непречена понатамошна анализа.

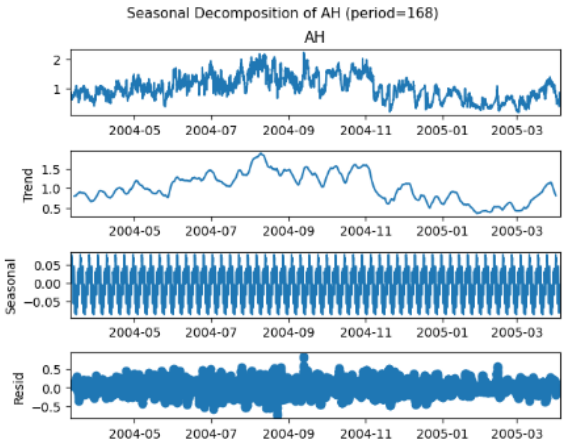
Визуелизација на податоците

Визуелизација на податоци претставува процес на **графичко претставување на информации** со цел полесно и поефикасно нивно разбирање. Таа е **клучна алатка во анализата на податоци**, бидејќи овозможува откривање на **трендови**, **шеми и аномалии** кои тешко би се воочиле преку бројки или табели. Со визуелниот приказ комплексните податоци стануваат поразбирливи, а **комуникацијата на резултатите е поедноставна и достапна** и за луѓе без техничка позадина. **Визуелизацијата** не само што го подобрува

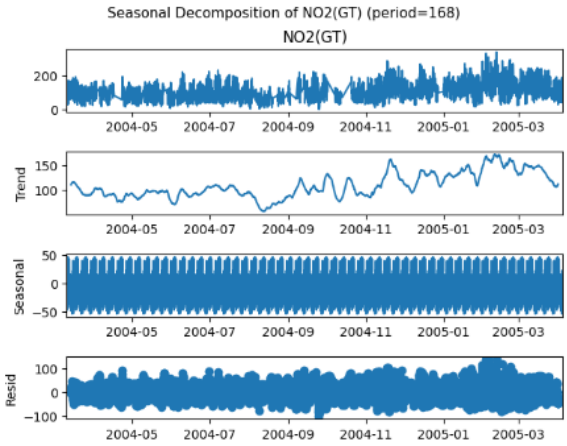
процесот на анализа, туку и го поддржува **донесувањето на информирани одлуки** врз основа на јасни и прегледни информации.

Анализа на временска серија

За подетално проучување на податоците е применета **сезонска декомпозиција на временски серии,** метод *seasonal_decompose* што е дел од библиотеката **statsmodels** во Python и овозможува разделување на податоците на **тренд, сезоналност и остаток**. Со овој пристап временската серија не се разгледува како еден единствен сигнал, туку се дели на **долгорочен тренд** кој ја опишува насоката на движење на податоците, **сезонска компонента** која ги прикажува повторливите шеми во рамките на дефиниран период и **остаток** кој ги претставува нерегуларните и случајни варијации. На овој начин податоците стануваат полесни за интерпретација бидејќи може да се разграничи што е **трајна промена**, што е **сезонски образец** и што претставува **шум**, а резултатите се визуелно прикажани со **графици** за појасно разбирање на целокупното однесување на серијата. Примери за сезонска декомпозиција на променлива се прикажани на **Слика 10** и **Слика 11**, каде при примената на декомпозицијата е дефиниран **период од 168 часа**, што одговара на повторување на сезонските шеми во рамките на **неделна структура** (часовни мерења со неделна повторливост).



Слика 10. Сезонска декомпозиција на променлива AH

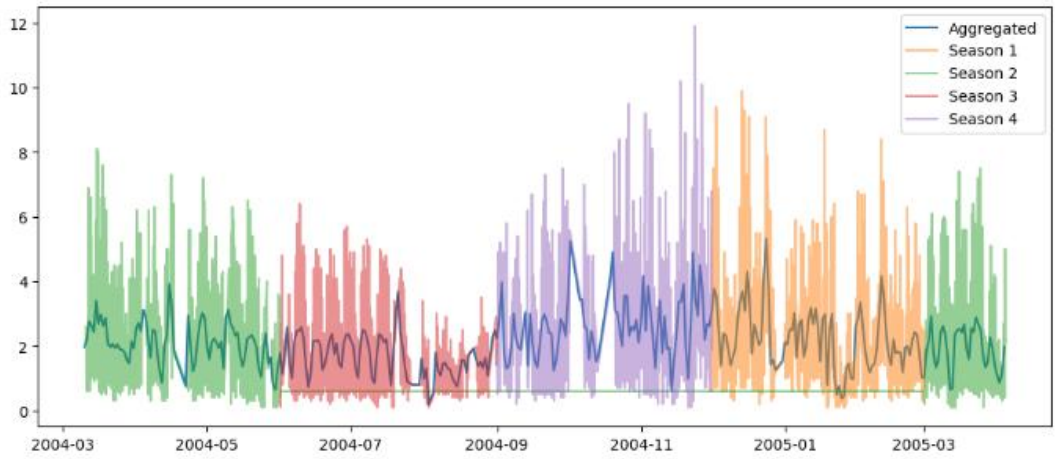


Слика 11. Сезонска декомпозиција на променлива NO2

Симпсонов парадокс

Симпсонов парадокс е статистички феномен при кој **тренд или корелација** која се забележува во повеќе поединечни групи може да се промени кога групите се комбинираат во целокупен сет на податоци. Овој парадокс се јавува кога постои **скриена променлива или фактор** кој влијае на резултатите, што води до суптилни и понекогаш **измамнички заклучоци** при анализата на агрегирани податоци. На пример, при разгледување на серија податоци поделена на различни сезони, можеби ќе се забележи **позитивен тренд** во секоја

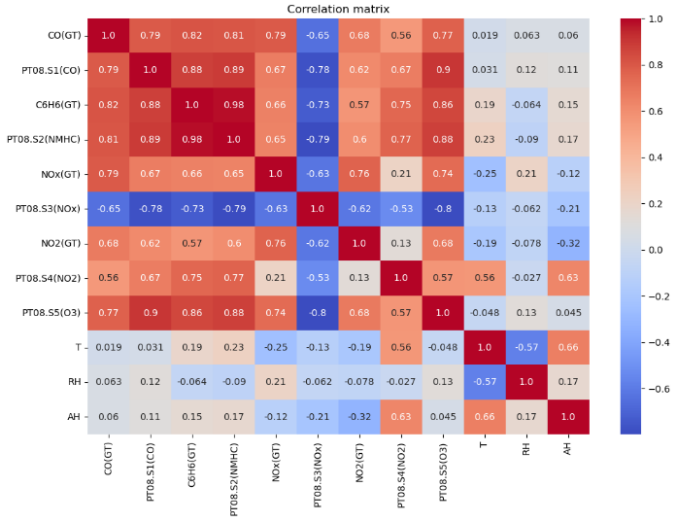
поединечна сезона, додека агрегираниот податок може да покаже **спротивен тренд**. Поради тоа, е важно да се **анализираат податоците на сегментиран начин**, како што е прикажано на графикот, за да се разликуваат ефектите на различните групи и да се избегнат **погрешни заклучоци** кои произлегуваат од агрегирање на податоците.



Слика 12. Визуализација на Симпсонов парадокс врз променливата CO(GT) од AirQuality Dataset

Корелациска анализа

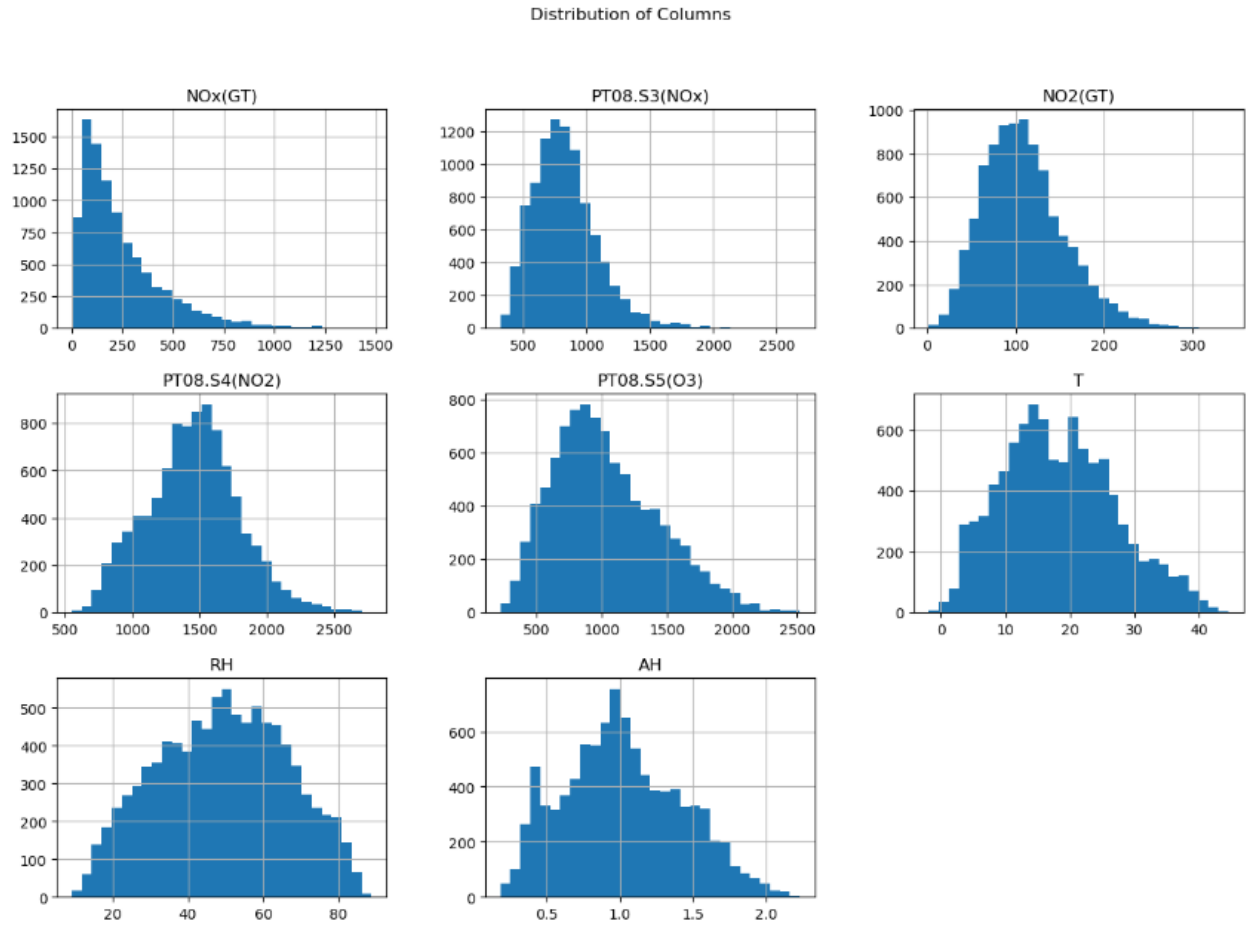
Корелациска анализа претставува **статистичка метода** која ја оценува **зависноста помеѓу две променливи**, прикажувајќи ја **силата и насоката на нивната меѓусебна поврзаност.** Во рамките на спроведената анализа беа пресметани корелациите меѓу сите променливи со користење на функцијата *corr()*, при што колони со **висока меѓусебна корелација поголема од 0.8** беа **отстранети од податочниот сет**. Овој пристап овозможува **намалување на дуплирањето на информацијата** и **минимизирање на мултиколинеарноста**, што е клучно за стабилноста и интерпретабилноста на понатамошните модели. Со елиминацијата на високо корелираните променливи, податоците стануваат почисти и појасни, овозможувајќи **поточна и доверлива анализа и моделирање**.



Слика 13. Корелациска матрица на променливите од AirQuality dataset

Анализа на распределба/дистрибуција

Анализата на распределба претставува **важен чекор во истражувачката анализа на податоците**, бидејќи овозможува увид во тоа како се распоредени вредностите на различните променливи, дали постојат **екстремни вредности** и колку се концентрирани околу **средната вредност**. Во рамките на овој проект, беше направена **распределбена анализа на колоните** во податочниот сет по **елиминацијата на високо корелираните променливи**, со цел да се добие **почист и поинформативен сет за понатамошната анализа**. На слика 14 е прикажана дистрибуцијата на останатите колони, што овозможува визуелна проценка на нивните карактеристики и варијабилност.



Слика 14. Распределба на променливи од AirQuality dataset

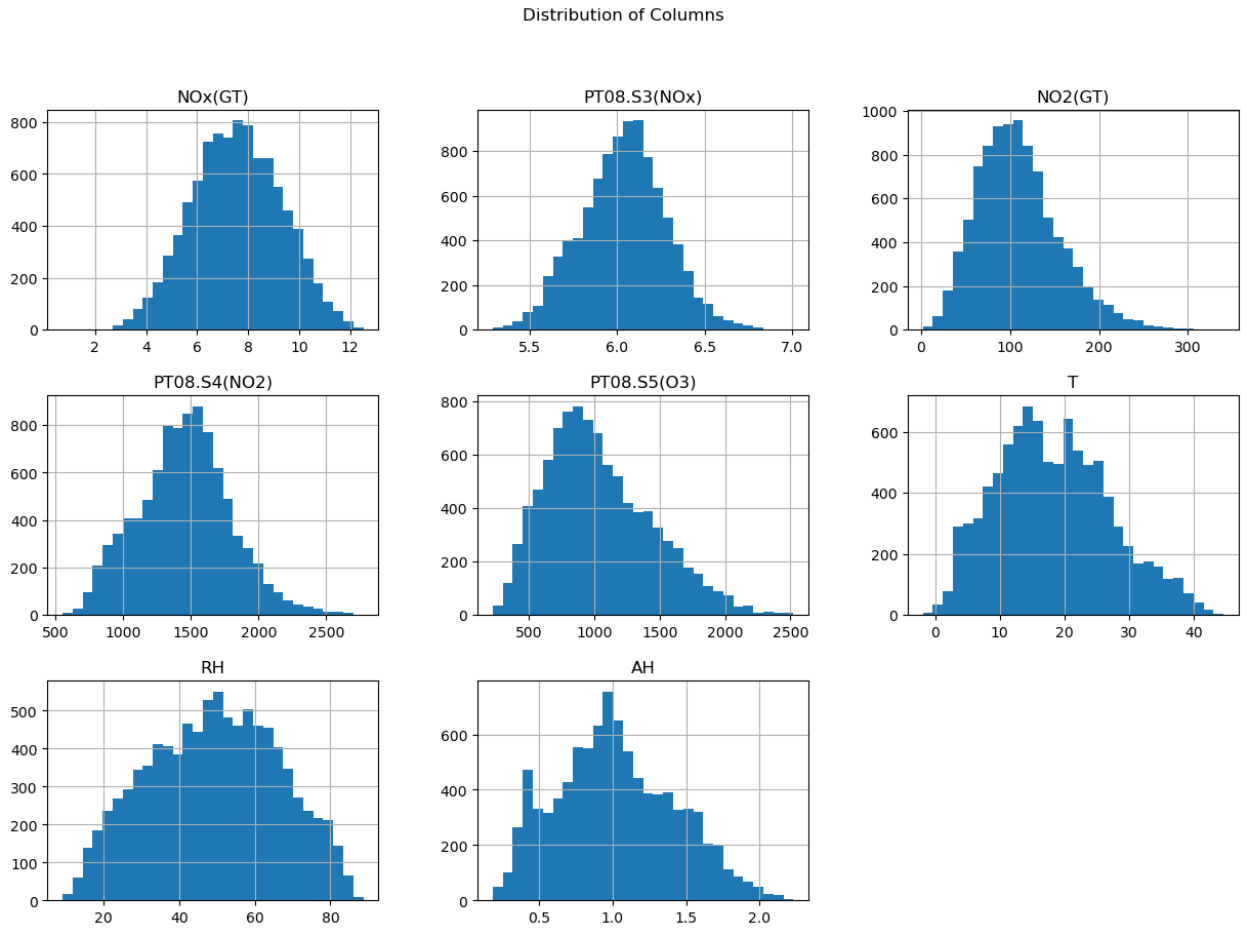
Box-Cox

Во текот на **анализа на распределба на податоците**, беше идентификуван проблем со **асиметрија (*skewness*)** кај одредени променливи. *Skewness* е мера која покажува колку **распределбата на податоците е несиметрична** околу својата **средна вредност**. Високи позитивни или негативни вредности на *skewness* можат да ја нарушат **стабилноста и точноста** на статистичките модели и техниките за машинско учење, вклучувајќи и **PCA**,

бидејќи овие методи претпоставуваат дека **податоците се приближно нормално распределени**.

Со пресметување на *skewness* за секоја колона, беа идентификувани **колоните со апсолутна вредност на *skewness* поголема од 1**, што укажува на **значителна асиметрија**. За овие колони беше применета **Box-Cox трансформација**, која ја **намалува асиметријата** и ја **приближува распределбата до нормална**. За колони со **негативни или нулти вредности**, прво беше направено **поместување на податоците** така што сите вредности стануваат **позитивни**, што е услов за **Box-Cox трансформацијата**.

По трансформацијата, проверката на *skewness* покажа дека **асиметријата е значително намалена** и сите **високо скривени колони сега имаат вредности на *skewness* близу до нула**. Ова овозможува понатамошна анализа и моделирање со подобра стабилност и доверливост на резултатите.

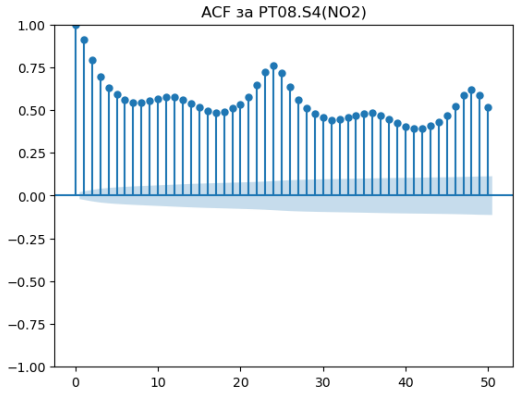


Слика 15. Распределба на променливи од AirQuality по Box-Cox трансформација dataset

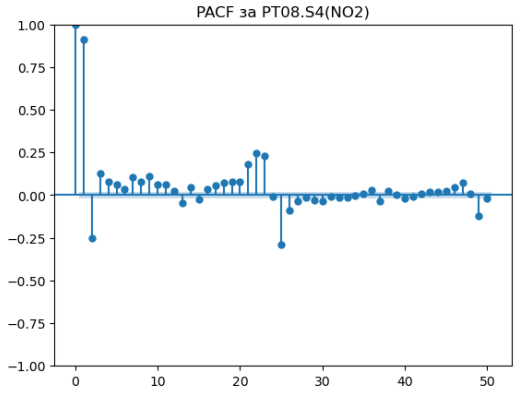
ACF (Autocorrelation Function) / PACF (Partial Autocorrelation Function)

За секоја променлива од податочниот сет беше извршена **анализа на автокорелација** со цел да се оцени **зависноста на сегашните вредности од претходните во временската**

серија. Со помош на **ACF** се прикажува колку сегашните вредности се поврзани со минатите вредности на различни задоцнувања, додека **PACF** ја прикажува корелацијата со претходните вредности исклучувајќи ја интерференцијата од помалите задоцнувања. Овие графици овозможуваат **идентификување на трендови, сезоналност и потенцијални редови на автокорелација**.



Слика 16. ACF за променлива PT08.S4 (NO2)



Слика 17. PACF за променлива PT08.S4 (NO2)

Principal Component Analysis (PCA)

Принципална компонентна анализа (PCA) претставува **статистичка техника за редукција на димензионалноста на податоците**, која ја намалува сложеноста на големиот број на променливи додека задржува што е можно повеќе од варијансата во оригиналните податоци. PCA ја трансформира оригиналната колекција на променливи во нов сет на независни променливи, наречени **главни компоненти (*principal components*)**, кои се **линеарни комбинации на оригиналните колони** и се уредени според тоа колку **варијанса во податоците објаснуваат.**

Во спроведената анализа, податоците во избраните релевантни колони беа **стандартизирани** со цел секоја променлива да има **иста скала**, што е клучно за **PCA**, бидејќи техниката е чувствителна на **различните мерни единици и варијанси.** Беше применета **PCA со праг од 95% на објаснета варијанса**, што значи дека се задржуваат онолку главни компоненти колку што е потребно за да се објасни 95% од варијансата во оригиналните податоци. На овој начин се намалува бројот на променливи, се елиминира шумот и се поедноставува понатамошната анализа или моделирање, додека се задржува најважната информација.

Datetime	PC1	PC2	PC3	PC4	PC5
2004-03-10 18:00:00	-0.010964	-0.409318	0.048978	1.134399	0.833332
2004-03-10 19:00:00	-1.119768	-0.473434	0.058705	1.037485	0.717972
2004-03-10 20:00:00	-0.591651	-0.755030	-0.177719	0.881355	0.955909
2004-03-10 21:00:00	-0.088301	-0.876133	-0.493947	0.832527	1.018797
2004-03-10 22:00:00	-0.649552	-0.924282	-0.467137	0.776146	1.063058
...	...	...	...	...	...
2005-04-04 10:00:00	2.651773	-0.496135	1.780904	-0.136524	-0.389453
2005-04-04 11:00:00	1.485162	-0.402537	2.177319	-0.546508	-0.301920
2005-04-04 12:00:00	1.070349	-0.235564	2.579728	-0.623024	-0.376675
2005-04-04 13:00:00	-0.093230	-0.443358	2.956288	-0.891453	-0.468742
2005-04-04 14:00:00	0.335619	-0.384910	3.030828	-0.849884	-0.381790

Слика 18. DataFrame со PCA компоненти од AirQuality dataset

Резултатот е нов DataFrame каде секоја колона претставува една **главна компонента**, со **индексите зачувани од оригиналниот податочен сет**.

По примена на PCA и добивање на главните компоненти, беше спроведена **проверка на мултиколинеарноста** користејќи го **Variance Inflation Factor (VIF)** за секоја компонента. **VIF** е мерка која покажува колку **варијансата на регресиските коефициенти се зголемува поради мултиколинеарност** помеѓу променливите. Ниска вредност на VIF (вообичаено под 5) укажува дека променливата не е мултиколерирана со другите, додека висока вредност покажува силна зависност од другите колони. Во случајот со добиените главни компоненти од PCA, **сите VIF вредности се 1,** што е очекувано и логично, бидејќи PCA создава **некорелирани (ортогонални) компоненти**.

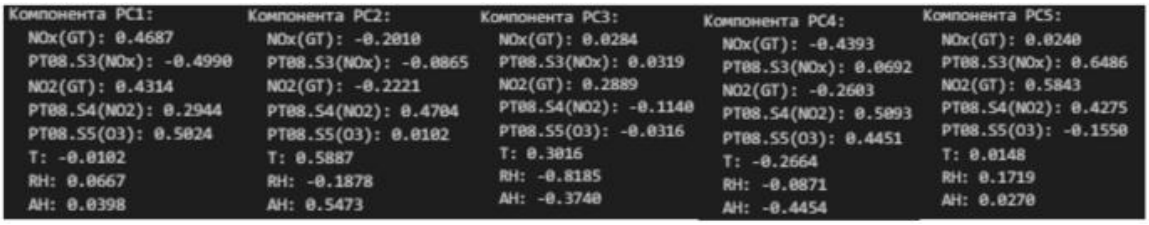
	Feature	VIF
0	NOx(GT)	83.872801
1	PT08.S3(NOx)	74.411707
2	NO2(GT)	27.568659
3	PT08.S4(NO2)	85.182631
4	PT08.S5(O3)	39.777527
5	T	73.980501
6	RH	71.359852
7	AH	68.577657

Слика 19. VIF вредности за променливи од AirQuality dataset

	Feature	VIF
0	PC1	1.0
1	PC2	1.0
2	PC3	1.0
3	PC4	1.0
4	PC5	1.0

Слика 20. VIF вредности за PCA компоненти

Со претставување на тежините на променливите кај секоја компонента се заклучува дека **PC1** е доминирана од O_3 и *NOx,* **PC2** од *T* и *AH*, а **PC3** од *RH*, додека **PC4** и **PC5** се поврзани со *NO2, NOx* и *RH*.



Слика 21. Вредности за тежини на променливи кај секоја PCA копмпонента

Примена на техники од податочно рударство

Анализата на временски серии се занимава со проучување на **податоци зависни од времето** со цел **идентификување на трендови, сезонски варијации и аномалии**. Алгоритмите за оваа анализа овозможуваат **моделирање на историските податоци за точно предвидување на идните вредности**. Различни пристапи се користат во зависност од природата на податоците, при што некои алгоритми се оптимизирани за **краткорочни и стабилни серии**, а други се наменети за **долгорочни, сезонски или комплексни податоци**. Примената на овие методи обезбедува поддршка во донесувањето одлуки и е корисна во области каде се следат динамични процеси.

LSTM

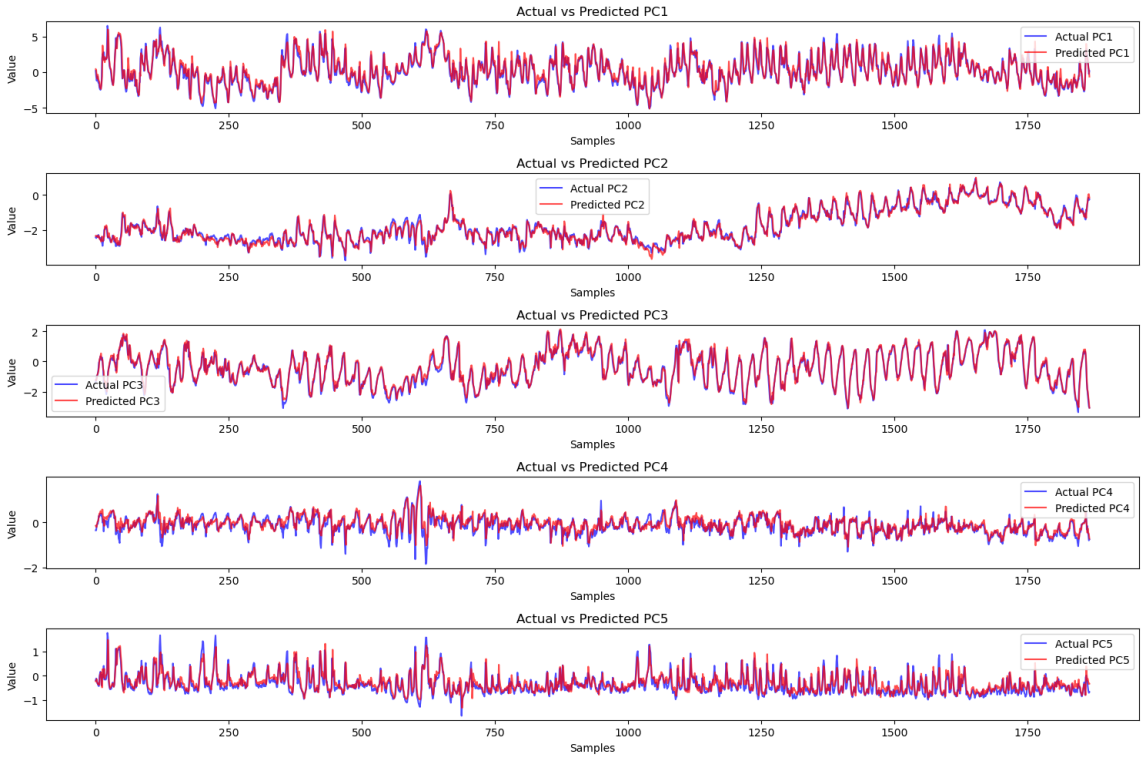
LSTM (Long Short-Term Memory) е тип на **рекурентни неурални мрежи** наменет за **анализа на секвенцијални и временски зависни податоци**. Неговата предност е **способноста да „памети“ информации** на подолг временски период преку механизми за контрола на протокот на информацијата. Со тоа се избегнува проблемот на класичните RNN кои имаат потешкотии при моделирање на подолги зависности. Ова ја прави **LSTM особено**

применлива кај временски серии каде што **тековните вредности зависат од историски трендови и повторливи обрасци**.

Во конкретниот модел се користат **24 претходни чекори како влез**, при што податоците се претставени преку главните компоненти добиени од PCA анализа (PC1–PC5). За секоја секвенца на влезни податоци, **таргет се тековните вредности на сите компоненти**. Податоците се поделени на **обука и тестирање во сооднос 80:20**, без мешање, за да се зачува хронолошката структура. Архитектурата на мрежата се состои од **еден LSTM слој со 50 неврони** и **излезен Dense слој со пет неврони** што одговараат на бројот на таргет променливи. Моделот е обучен со **Adam оптимизатор** и функција на грешка **MSE**.

Резултатите укажуваат на **добра способност** на моделот да ги следи и предвидува временските серии. Метричките вредности **MSE** од 0.1343, **RMSE** од 0.3664 и **MAE** од 0.2350 ја прикажуваат релативно ниската грешка, додека $\mathbf{R^2}$ од 0.8451 значи дека моделот објаснува околу **85%** од варијансата во податоците.

Визуелната споредба на реалните и предвидените вредности за секоја компонента дополнително потврдува дека **моделот успешно ги следи трендовите и варијациите во податоците**.



Слика 22. Визуализација на предвидени и точни вредности за секоја компонента со Unidirectional LSTM

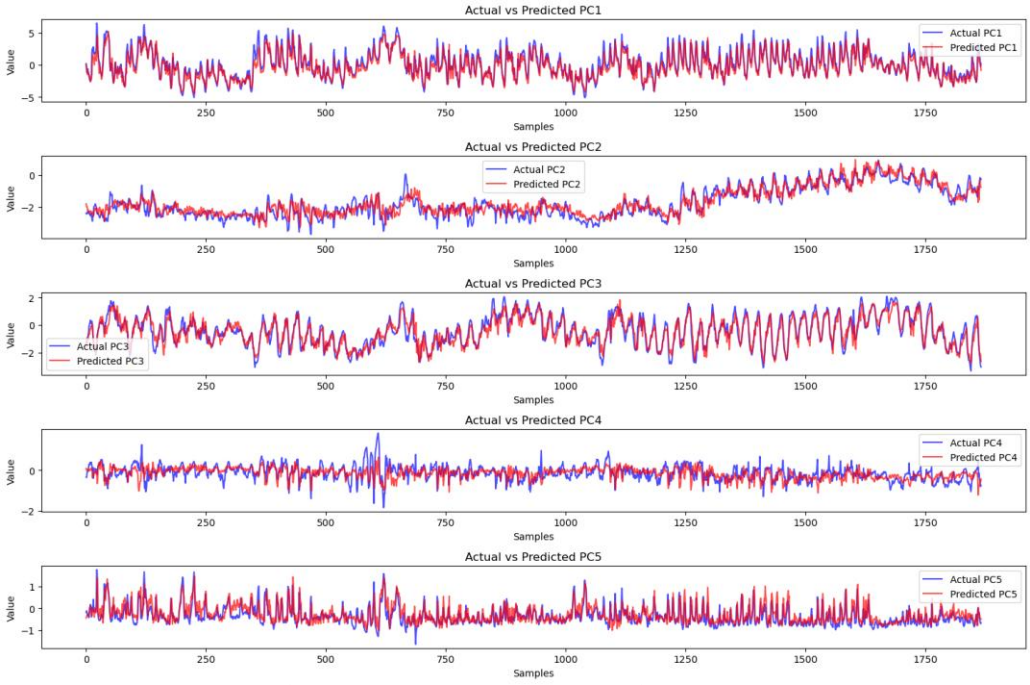
Bidirectional LSTM

Bidirectional LSTM претставува **проширување на класичната LSTM архитектура** кое овозможува **моделот да ги обработува податоците во два насоки** – од минатото кон иднината и обратно. Со ова моделот има пристап и до претходните и до идните контексти.

Во конкретната имплементација, за секоја од петте главни компоненти добиени со PCA анализа се креираат **lag карактеристики за претходните 24 чекори**. На тој начин, **секоја влезна секвенца го опфаќа историскиот контекст на податоците**. **Влезниот сет X** е организиран во **тридимензионална структура** (примероци × чекори × карактеристики), додека **таргет y** ги содржи **тековните вредности на сите компоненти**. Податоците се поделени на **обука и тест сет во сооднос 80:20**, без мешање, со цел зачувување на **временската структура**.

Мрежата е изградена од **еден Bidirectional LSTM слој со 50 неврони** и **активациона функција tanh**, проследен со **Dense слој кој има пет неврони**. За обука е користен **Adam оптимизатор** со функција на грешка **MSE**, процес изведен во **20 епохи** со **batch size 32**.

Оценувањето на перформансите покажува дека **моделот има умерена способност** за следење на динамиката во податоците. Добиваните вредности на метриките **MSE** 0.2071, **RMSE** 0.4551, **MAE** 0.3316 и **R^2** 0.6823 укажуваат на тоа дека **грешката е поголема во споредба со еднонасочниот LSTM модел**, а објаснетата **варијанса е намалена на околу 68%**. Ова значи дека иако Bidirectional пристапот овозможува поширок контекст, во овој случај **не придонесува кон зголемување на точноста** (како што може да се заклучи од визуалниот приказ на точните и предвидени вредности на Слика 22), веројатно поради природата на податоците и фактот што временските серии најчесто имаат **каузална структура** каде идните вредности не треба да влијаат на предвидувањата.



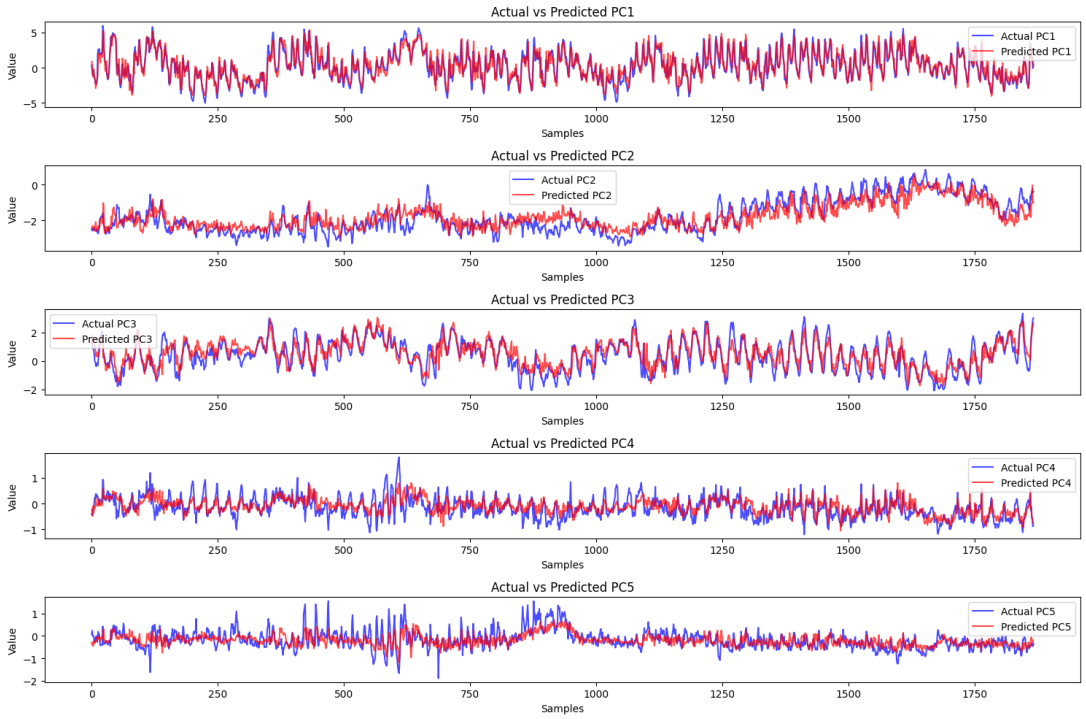
Слика 23. Визуализација на предвидени и точни вредности за секоја компонента со Bidirectional LSTM

GRU (RNN) Multivariate

GRU (Gated Recurrent Unit) е тип на **рекурентни неурални мрежи** наменет за анализа на секвенцијални и временски зависни податоци. Неговата **главна предност е поедноставена архитектура со помал број на параметри** бидејќи користи само **две порти update** и **reset** за контрола на протокот на информациите што го прави поефикасен за тренирање при поголеми или посложени податоци и при тоа задржува способност за моделирање на долгорочни зависности.

Во конкретниот модел се користат **24 претходни временски чекори како влез**. За секоја променлива се создаваат **лагови од 1 до 24** со што се добиваат секвенци кои ја претставуваат динамиката низ време. **Влезната матрица X** е обликувана во **3D тензор** со димензии: број на примероци, број на чекори 24 и број на променливи**. Таргет вредностите y** се **тековните непоместени вредности** на сите променливи. Податоците се поделени на **обука и тест во сооднос 80:20** без мешање за да се зачува хронолошкиот редослед. Архитектурата на мрежата се состои **од еден GRU слој со 50 неврони** и **активација tanh** и **излезен Dense слој** со број на неврони еднаков на бројот на таргет променливи. Моделот е обучен со **Adam оптимизатор** и функција на грешка **MSE**. Метричките резултати покажуваат добра способност на моделот да ги предвидува временските серии со **MSE** 0.2872, **RMSE** 0.5359, **MAE** 0.3912 и **R^2** 0.6124 што значи дека моделот **објаснува над 60% од варијансата во податоците**.

Визуелната споредба на предвидените и реалните вредности покажува дека моделот **успешно ги следи трендовите и динамиката** иако постојат **одредени отстапувања** во делови со повисока варијабилност.

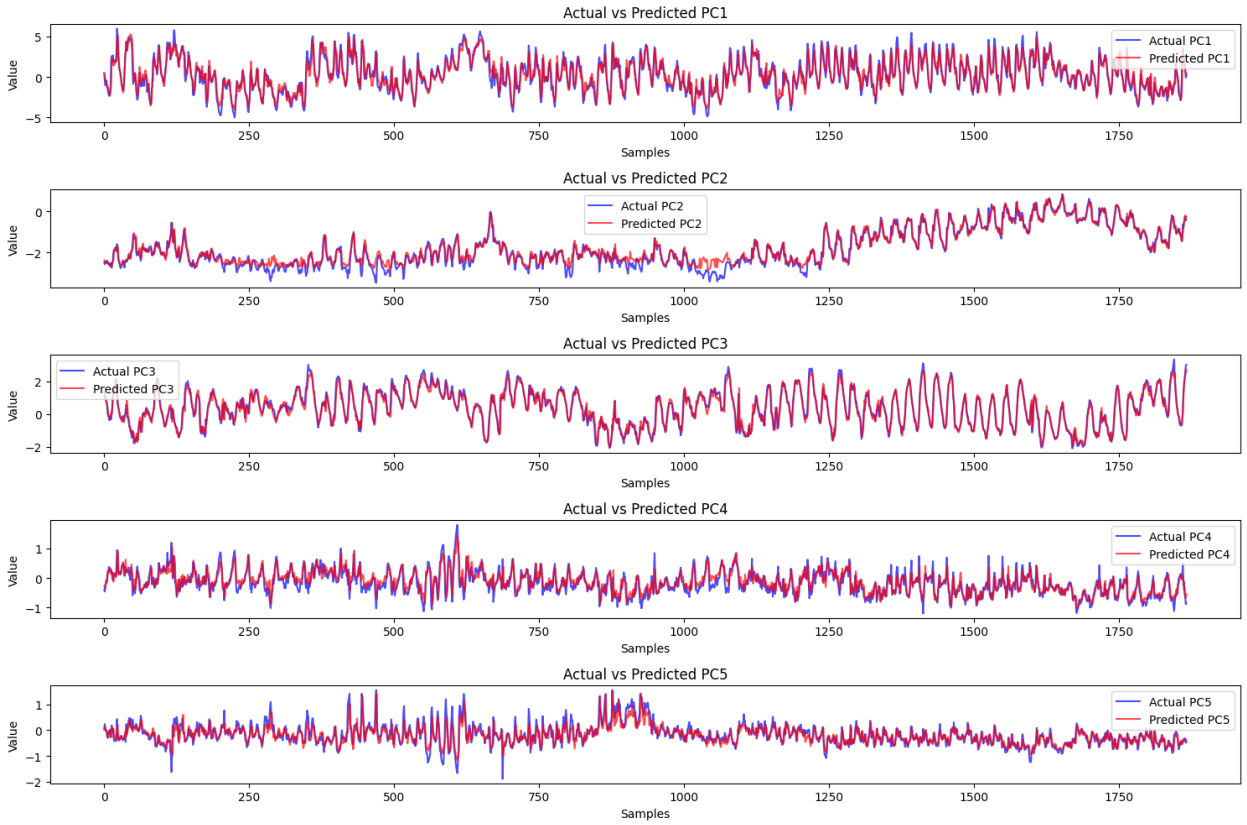


Слика 24. Визуализација на предвидени и точни вредности за секоја компонента со GRU

XGBoost

XGBoost е **алгоритам за градење на градиенти заснован на дрвја на одлука** кој е развиен за да обезбеди **висока точност, брзо тренирање и добра способност за справување со комплексни податоци**. Во контекст на временски серии, иако **XGBoost не е рекурентен модел**, можеме да го користиме **со лагови на претходни временски чекори** за да се моделира зависноста од минатите вредности.

Во конкретниот модел се користат **24 претходни чекори како влез**, при што за секоја променлива се создаваат **лагови од 1 до 24** и се формираат **функции кои ја претставуваат динамиката на времето**. **Влезната матрица X** ги содржи сите овие лагови додека **таргет вредностите y** ги претставуваат тековните вредности на сите променливи, што го прави моделот **multivariate**. Податоците се поделени на **обука и тест во сооднос 80:20** без мешање за да се зачува хронолошкиот редослед. Моделот се состои од X**GBRegressor со 200 дрвја**, **learning rate 0.1** и **максимална длабочина на дрвјата 5**, додека **MultiOutputRegressor** овозможува **истовремено предвидување на повеќе таргет променливи**. Метричките резултати укажуваат на висока точност со **MSE** 0.1512, **RMSE** 0.3888, **MAE** 0.2472 и **R^2** 0.8307 што значи дека моделот објаснува **над 83% од варијансата** во податоците и **успешно ги следи трендовите и динамиката** на временските серии иако не користи рекурентна структура.



Слика 25. Визуализација на предвидени и точни вредности за секоја компонента со XGBoost

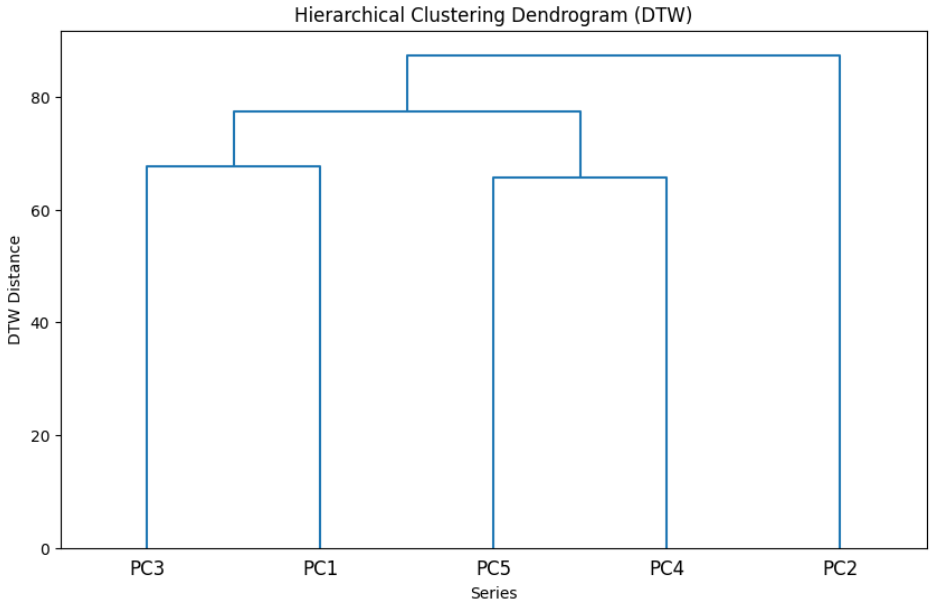
Dynamic Time Warping (DTW)

Dynamic Time Warping (DTW) е **техника за мерење на сличноста меѓу временски серии** која овозможува **еластично усогласување на времето**, така што серии со слични обрасци, но одложени или скалирани во времето, можат да се идентификуваат како слични. Во овој пример, секоја главна компонента од PCA анализата се третира како временска серија.

Прво, податоците се нормализираат со **TimeSeriesScalerMeanVariance**, со што се отстранува ефектот на различни скали и средни вредности. Потоа се пресметуваат **DTW растојанија** меѓу сите серии и се користи **hierarchical clustering** со **средна врска** (average linkage) за групирање на сериите во **2 кластери**. Резултатите од кластеризацијата покажуваат следна распределба: **PC1, PC3, PC4 и PC5** припаѓаат на кластер 1, а **PC2** припаѓа на кластер 2.

Метриките на квалитет на кластеризацијата, **Silhouette score = 0.114** и **Davies-Bouldin индекс = 0.775**, укажуваат на умерена разделеност и компактност на кластерите. Со **само 5 серии**, и со DTW растојанија, очекувано е Silhouette score низок и DB индекс умерен. Ова значи: сериите се **делумно слични и делумно различни**, но **нема јасна поделба** на два добро дефинирани кластери.

Визуализацијата преку **dendrogram** ја прикажува **хиерархиската структура на сличностите меѓу сите серии**. Гранките ја претставуваат **DTW растојанијата**, додека сериите се означени на x-оска. Должината на гранките покажува колку се различни сериите или групите серии. Од **dendrogramot** може да се види кои серии се најслични и како **DTW** ги групира сериите **според нивната форма**, наместо според **апсолутните вредности**. Ова овозможува **визуелно идентификување на групи на серии со слична временска динамика и** разликување од серии **со различни обрасци**.



Слика 26. Дендрограм визуализација на кластери добиени со DTW

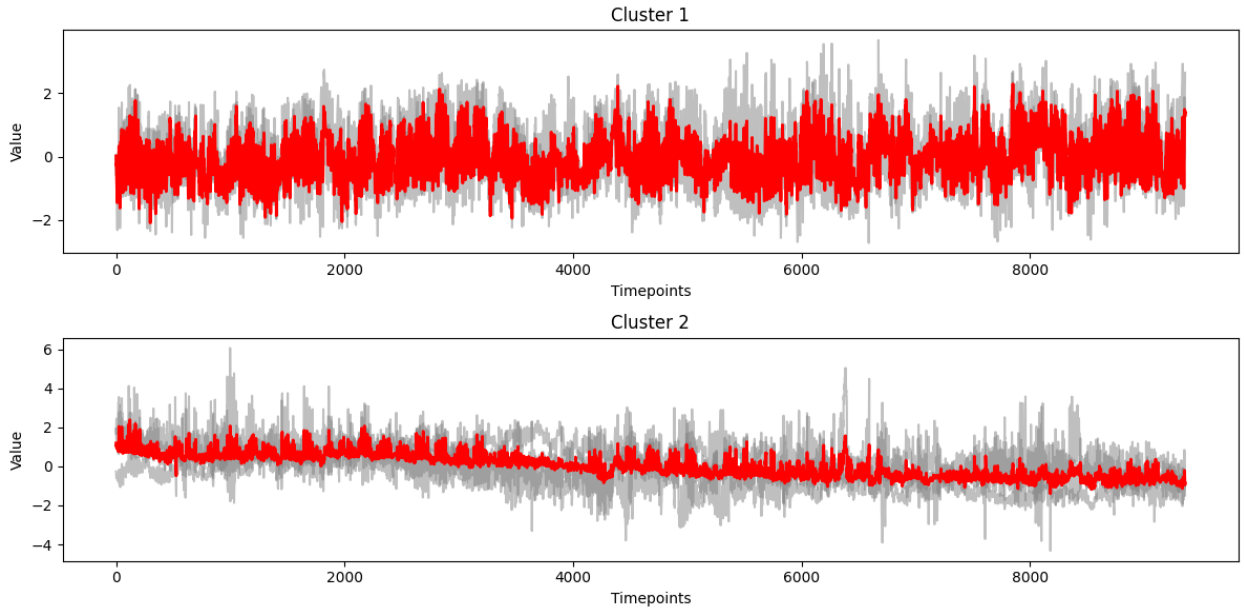
Time Series Spectral Clustering

Time Series Spectral Clustering е техника која ги групира временските серии врз основа на нивните **спектрални карактеристики**, како честотни и сезонски обрасци. Секоја **главна компонента** од PCA анализата се третира како временска серија, која прво се нормализира со **TimeSeriesScalerMeanVariance** за да се отстранат ефектите на различни **скали и средни вредности**. Потоа, секоја серија се трансформира во **спектрална репрезентација** преку **Fourier трансформација**, а **Spectral Clustering** се користи за **групирање на сериите во два кластери** според сличноста на нивните спектрални обрасци.

Резултатите од кластеризацијата покажуваат дека **PC1** и **PC3** припаѓаат на еден кластер, додека **PC2**, **PC4** и **PC5** се во другиот кластер. Визуализацијата ја прикажува распределбата на сериите во секој кластер со сиви линии за индивидуалните серии и црвена линија која ја претставува типичната спектрална динамика на кластерот.

Метриките за квалитет на кластерите покажуваат дека **Silhouette score** е 0.103, што укажува на тоа дека границите меѓу кластерите се релативно нејасни и дека дел од сериите можат да бидат слични на повеќе од еден кластер. **Davies-Bouldin индексот** од 1.307 исто така покажува умерена компактност и разделеност на кластерите. Овие резултати укажуваат дека кластерите ја доловуваат **општата сличност на сериите**, но постои **одредена преклопеност и варијабилност внатре во групите**.

Графикот ја прикажува распределбата на сериите во кластерите добиени со **Spectral Clustering**. Секој кластер е прикажан во **посебен subplot**, каде **сивите линии** ги претставуваат индивидуалните серии во кластерот, а **црвената линија** ја покажува централната траекторија или типичната динамика на кластерот. На овој начин, лесно се гледа кои серии имаат слични обрасци и како се групираат според спектралните карактеристики.



Слика 27. Визуализација на кластери добиени со TSSC

Резултати и дискусија

Заклучокот укажува дека **класичниот LSTM** најточно ги предвидува временските серии со најниска грешка и највисоко **R^2**, додека **Bidirectional LSTM** и **GRU** покажуваат повисока грешка и помала способност за моделирање на сложените времески зависимости. **XGBoost**, иако не е рекурентен модел, постигнува слично ниво на точност како **LSTM**, што го прави ефикасна и поедноставна алтернатива за мултиваријантни временски серии.

Кластеризацијата на временските серии, преку **Spectral Clustering** и хиерархиска кластеризација со **DTW**, овозможува идентификација на групи со слични обрасци. **Spectral Clustering** ја групира сериите според спектралните карактеристики и ја открива централната динамика на кластерите, додека хиерархиската кластеризација го илустрира нивното растојание и сличност преку **dendrogram**. Метриките за квалитет на кластерите укажуваат на умерена поделба и делумна преклопеност, што значи дека сериите делумно се слични меѓу кластери. Овие резултати се корисни за визуелна сегментација и анализа на временската динамика на сериите.

Анализата покажува дека **главните фактори** кои ја објаснуваат варијансата се концентрациите на гасовите *O_3*, *NOx* и *NO2*, како и *RH* и *AH*. **PC1** е доминирана од *O_3* и *NOx*, **PC2** од температура и *AH*, а **PC3** од *RH*, додека **PC4** и **PC5** се поврзани со *NO2*, *NOx* и *RH*. Гасовите *O_3*, *NOx* и *NO2* **се штетни** за респираторниот систем и здравјето, додека *RH* и *AH* **влијаат на нивните концентрации**, но сами по себе не се штетни.

Заклучок

Анализата ја откри динамиката и варијацијата на *загадувањето на воздухот*, како и скриените обрасци во податоците. Резултатите покажуваат дека **кластеризацијата и анализа на временски серии** овозможуваат подобро разбирање и предвидување на загадувањето. **Загадувањето на воздухот** е во голема мера предизвикано од **присуството на гасовите** *O_3*, *NOx* и *NO_2*, кои директно влијаат на здравјето на луѓето, особено на респираторниот систем. Релативната и апсолутната влажност (*RH* и *AH*) ја менуваат **концентрацијата на овие гасови** во воздухот, што значи дека условите на воздухот можат да го зголемат или намалат ефектот на загадувањето.

Сето ова помага при креирање на **ефективни мерки за заштита на животната средина и јавното здравје**.

Литература / Извори

- **AirQuality Dataset:** https://archive.ics.uci.edu/dataset/360/air+quality
- **Документација за користени пакети и функции:**
  - **Python** *(основни функции и библиотеки)*
    https://docs.python.org/3/
  - **NumPy** *(нумерички пресметки, матрици)*
    https://numpy.org/doc/stable/
  - **Pandas** *(табеларни податоци)*
    https://pandas.pydata.org/docs/
  - **Matplotlib** *(визуализации)*
    https://matplotlib.org/stable/contents.html
  - **Seaborn** *(статистички визуелизации)*
    https://seaborn.pydata.org/
  - **Scipy** *(статистика, научни функции)*
    https://docs.scipy.org/doc/scipy/reference/
  - **Scikit-learn** *(машинско учење, кластеризација, предобработка)*
    https://scikit-learn.org/stable/modules/classes.html
  - **Statsmodels** *(статистички модели, ACF/PACF, seasonal_decompose)*
    https://www.statsmodels.org/stable/index.html
  - **XGBoost**
    https://xgboost.readthedocs.io/en/stable/
  - **TensorFlow / Keras** *(Neural Networks: LSTM, GRU, Dense, Sequential)*
    https://www.tensorflow.org/api_docs/python/tf/keras
  - **TSLearn** *(TimeSeriesKMeans, TimeSeriesScalerMeanVariance, cdist_dtw, TimeSeriesResampler)*
    https://tslearn.readthedocs.io/en/stable/
  - **Scipy.cluster.hierarchy** *(linkage, dendrogram, fcluster)*
    https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html
  - **Spectral Analysis** *(пресметка на спектрални карактеристики за Time Series Spectral Clustering)*
    https://numpy.org/doc/stable/reference/routines.fft.html