Тимски проект

Анализа и предвидување на глобалното загадување на воздухот во различни земји со примена на методи од машинско учење

Линк до кодот:
https://github.com/aleksandrapendovska/Global_Air_Pollution/tree/master

Изработиле:
Марија Јорданоска 203029
Тиана Костова 203061
Александра Пендовска 203089

Ментор:
проф. Д-р Илинка Иваноска

Содржина

ВОВЕД

Во текот на изминатите неколку децении, загадувањето на воздухот стана еден од најзначајните еколошки предизвици со кои се соочува светот. Зголемената урбанизација, индустријализацијата и транспортните активности доведоа до зголемени концентрации на штетни материи во воздухот, што има сериозни негативни влијанија врз јавното здравје и животната средина. Секојдневно, милиони луѓе ширум светот се изложени на високо ниво на загаденост на воздухот, што доведува до респираторни заболувања, кардиоваскуларни проблеми и прерана смрт.

Во овој контекст, од клучно значење е да се развијат ефективни методи за анализа и предвидување на нивоата на загаденост на воздухот. Машинското учење, како дел од областа на вештачката интелигенција, обезбедува моќни алатки и техники за обработка и анализа на големи количини на податоци. Со примена на машинско учење, можеме да развиеме модели кои точно ги предвидуваат идните концентрации на загадувачи, овозможувајќи ни да преземеме соодветни мерки за намалување на негативните влијанија.

Објаснување на проблемот

Еден од најголемите предизвици во решавањето на проблемот со загадувањето на воздухот е неговата сложеност и динамичност. Загадувањето на воздухот е резултат на комбинација од различни фактори кои можат да варираат во времето и просторот. Дополнително, интеракцијата помеѓу локалните и глобалните извори на загадување ја комплицира способноста за точно предвидување и управување со нивото на загадување. Токму тука машинското учење може да одигра клучна улога. Со анализа на големи и комплексни податочни сетови, алгоритмите за машинско учење можат да идентификуваат скриени патерни и врски помеѓу различните фактори на загадување. Преку развој на предиктивни модели, можеме да добиеме појасна слика за тоа како различните извори на загадување влијаат на квалитетот на воздухот и како тие се менуваат со текот на времето.

Цел на проектот

Целта на ова истражување е да ги искористи напредните техники на машинско учење за да се анализираат и предвидат нивоата на загадување на воздухот во различни земји ширум светот. Преку овој пристап, ќе се обидеме да придонесеме кон подобро разбирање на проблемот и да понудиме ефективни решенија за негово намалување и управување.

Пристап на работа

На почетокот на проектот пристапуваме до потребните библиотеки кои ќе ги користиме понатаму во текот на имплементирањето.

Процесот е поделен на неколку фази, секоја со специфични задачи и методи кои обезбедуваат структурирана и систематска реализација на целите на истражувањето.

- **Фаза 1: Прибирање и подготовка на податоци**
  - Искористивме податочно множество од јавна бази на податоци.
  - Справување на вредностите кои недостасуваат
  - Направивме нормализација и стандардизација на податоците за да се обезбеди консистентност.
- **Фаза 2: Енкодирање и трансформација на податоци**
  - Енкодирањето на категориските променливи го направивме со користење на методот Label Encoding за да се претворат во нумерички вредности.
  - Применивме техники за скалирање на податоците, како што е StandardScaler, за да се усогласат различните опсези на вредности.
- **Фаза 3: Развој на модели**
  - Моделот го избравме откако тестиравме различни модели за да утврдиме кој алгоритам најдобро ги исполнува критериумите за точност и перформанси.
  - Ги разделивме податоците на Train и Test множества за валидација. Потоа ги обучивме моделите и направивме оценување на перформансите со помош на тест множеството.
- **Фаза 4: Евалуација и оптимизација на модел**
  - Користевме различни метрики за евалуација на моделот, како што се Accuracy и Classification Report.
- **Фаза 5: Визуелизација**
  - Визуелизацијата на податоците и резултатите ја постигнавме со креирање на графикони кои ги прикажуваат трендовите и предвидувањата на квалитетот на воздухот.
  - Резултатите се презентирани на јасен и разбирлив начин за различни целни групи.

Анализа на глобалното загадување на воздухот во различни земји

Во делот за анализа главна цел е да се разбере влијанието на различните фактори врз квалитетот на воздухот и да се спореди квалитетот на воздухот во различни земји во светот.

Податоците вклучуваат непрекинати атрибути, како што се концентрациите на различни загадувачи во воздухот, како и категориски атрибути кои ги одредуваат нивните категории.

Различните фактори кои влијаат на квалитетот на воздухот, кои ќе ги разгледаме се:

Азот диоксид [NO2]: NO2 се формира од емисиите на автомобили, камиони и автобуси, електрани и теренска опрема. Изложеноста во кратки периоди може да ги влоши респираторните заболувања. Луѓето со астма, децата и постарите лица се изложени на поголем ризик за здравствените ефекти на NO2.

Озон [O3]: На ниво на површината, озонот се создава со хемиски реакции помеѓу оксидите на азот и испарливите органски соединенија (VOC). Озонот на ниво на земја може да предизвика неколку здравствени проблеми како болка во градите, кашлање, иритација на грлото и воспаление на дишните патишта. Озонот влијае и на вегетацијата и на екосистемите.

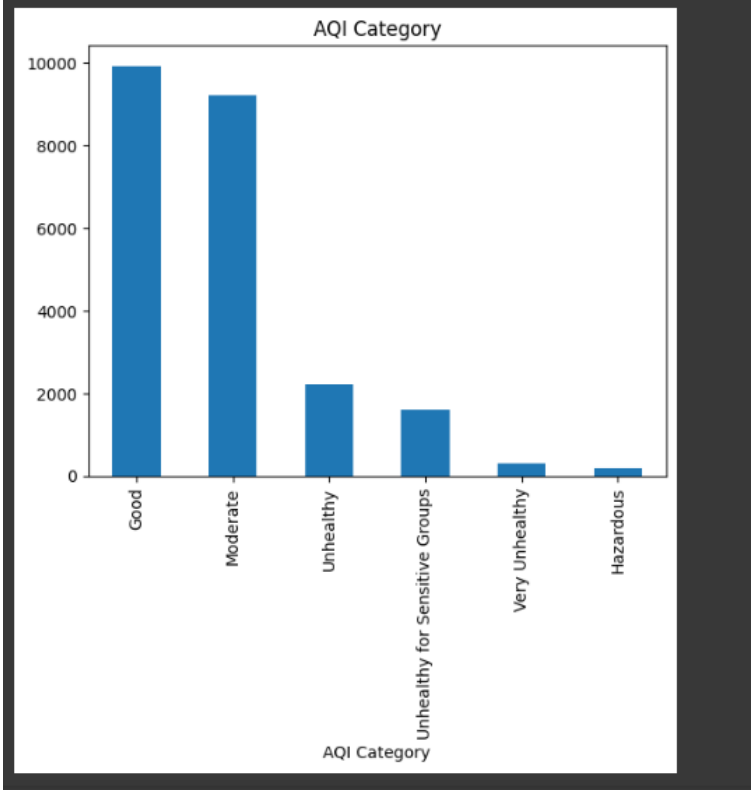
Јаглерод моноксид [CO]: Се емитува во воздухот пред сè од возила или машини кои согоруваат фосилни горива. Артиклите како што се керозин и грејачи на гас, шпорети на гас исто така ослободуваат CO, што влијае на квалитетот на воздухот во затворените простории. Дишењето воздух со висока концентрација на CO ја намалува количината на кислород што може да се транспортира во крвотокот до критичните органи како срцето и мозокот.

Честички [PM2.5]: Атмосферските честички, исто така познати како атмосферски аеросоли честички, се сложени мешавини од мала цврста и течна материја што влегуваат во воздухот. Ако се вдишат, може да предизвикаат сериозни проблеми со срцето и белите дробови. Тие се класифицирани како канцерогени од групата 1 од страна на Меѓународната агенција за истражување на ракот (IARC).

За таа цел, искористивме различни визуелизации:

- Загаденост на воздухот според AQI:

```
df['AQI Category'].value_counts().plot.bar()
plt.title('AQI Category')
plt.show()
```

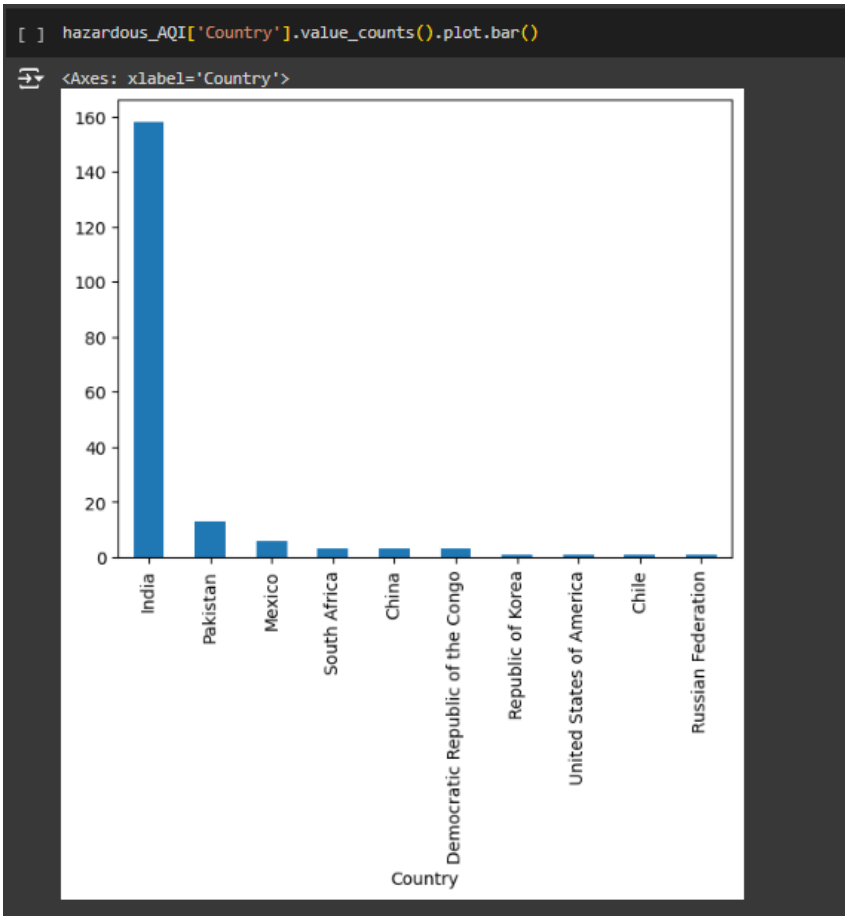


Legend:

- Good: < 50
- Moderate: 51-100
- Unhealthy: 101-150
- Unhealthy for sensitive groups: 101-150
- Very unhealthy: 151-200
- Hazardous: >200

- Резултат:
  9936 градови имаат добар квалитет на воздух
  2227 имаат лош квалитет на воздух
  286аат многу лош, додека пак 191 имаат опасен квалитет на воздух

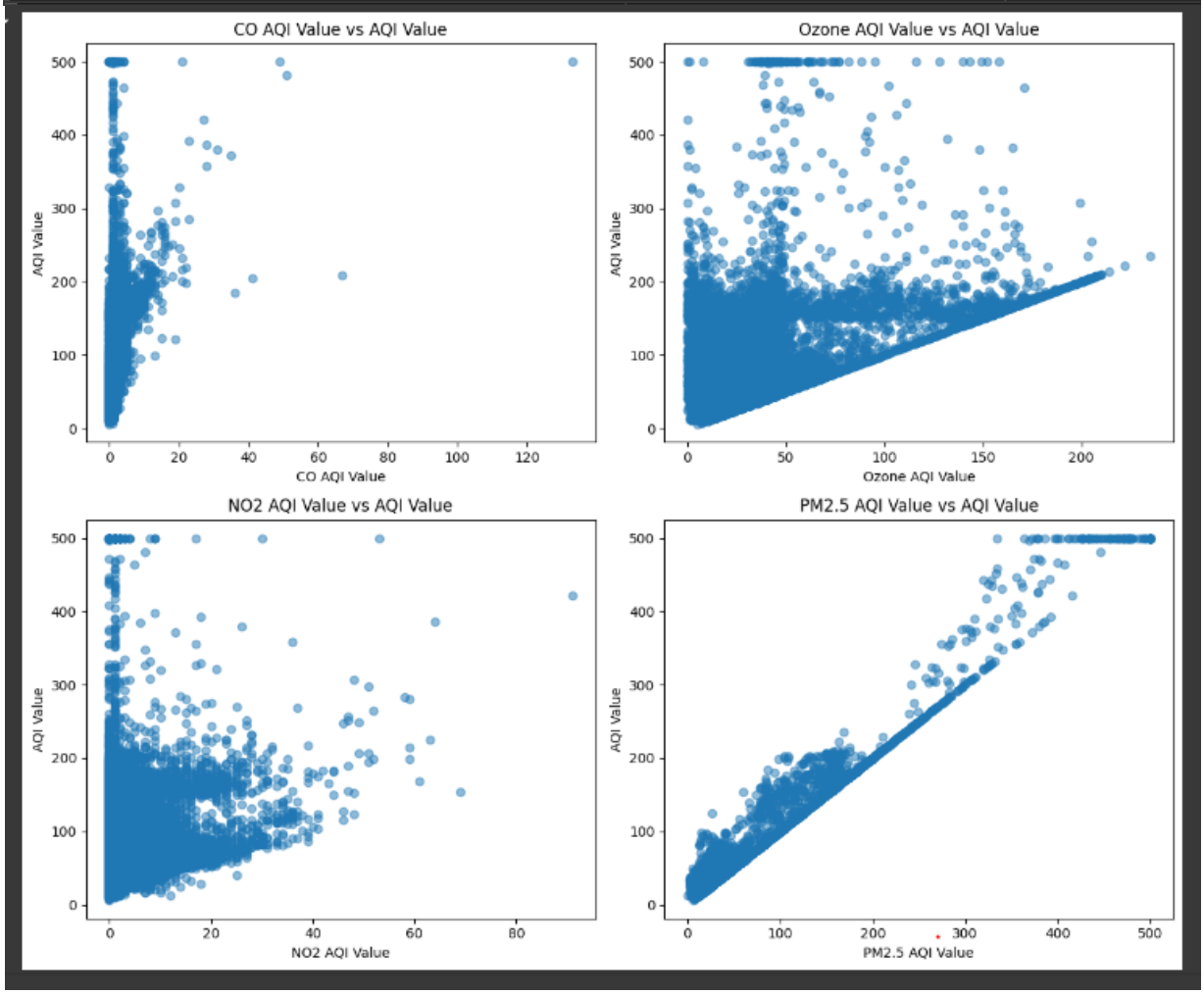
- Во која држава има најмногу градови со Hazardous AQI:



- Резултат:
  Од 190 градови, 158 се во Индија, 13 во Пакистан, 6 во Мексико и останатите се во Јужна Африка, Конго, Кореја, Кина, Чиле, Америка и Русија.

- Односите помеѓу Индексот за квалитет на воздухот (AQI) и поединечните вредности на AQI специфични за загадувачите

```
pollutants = ['CO AQI Value', 'Ozone AQI Value', 'NO2 AQI Value', 'PM2.5 AQI Value']
aqi_values = df['AQI Value']

fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(12, 10))

for i, ax in enumerate(axes.flat):
    ax.scatter(df[pollutants[i]], aqi_values, alpha=0.5)
    ax.set_xlabel(pollutants[i])
    ax.set_ylabel('AQI Value')
    ax.set_title(f'{pollutants[i]} vs AQI Value')

plt.tight_layout()
plt.show()
```



- Резултат:
  PM2.5 има подиректна и посилна корелација со целокупниот AQI, додека CO, Озонот и NO2 покажуваат помалку директни врски со повеќе расфрлани придонеси.

Предвидување на квалитетот на воздухот во земјите

Користејќи податочно множество од јавно достапна база, се фокусиравме на истражување и анализа на квалитетот на воздухот во различни земји. За да ги подготвиме податоците за анализа, искористивме методи за справување со Missing Values, како и нормализација и стандардизација на податоците, за целосно обезбедување консистентност и точност на нашата анализа.

Енкодирањето на категориските променливи го направивме со користење на методот Label Encoding за да се претворат во нумерички вредности.Исто така ги скалиравме податоците со StandardScaler, за да се усогласат различните опсези на вредности.

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
df['AQI Category'] = le.fit_transform(df['AQI Category'])

df['AQI Category'].value_counts()
```

```
AQI Category
0    9936
2    9231
3    2227
4    1591
5     287
1     191
Name: count, dtype: int64
```

```
from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

x = scaler.fit_transform(x)

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)
```

Нашиот избор за модел беше DecisionTreeClassifier (одлучувачко дрво за класификација), затоа што ги поддржуваше нашите потреби за класификација и даваше најдобри резултати според критериумите за точност и перформанси. Во истражувањето пробавме да користиме и LSTM Neural Network, меѓутоа, откако ја анализиравме потребата и природата на нашите податоци, утврдивме дека тој модел не беше погоден за нашиот случај, бидејќи LSTM е подобар за time series податоци.

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

from sklearn.tree import DecisionTreeClassifier

classifier = DecisionTreeClassifier(max_depth=3)
classifier = classifier.fit(x_train,y_train)

y_pred = classifier.predict(x_test)
```

Евалуацијата на моделот ја направивме со користење на метриките од библиотеката sklearn: Accuracy и Classification Report.

```
from sklearn import metrics

print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.977859778597786
```

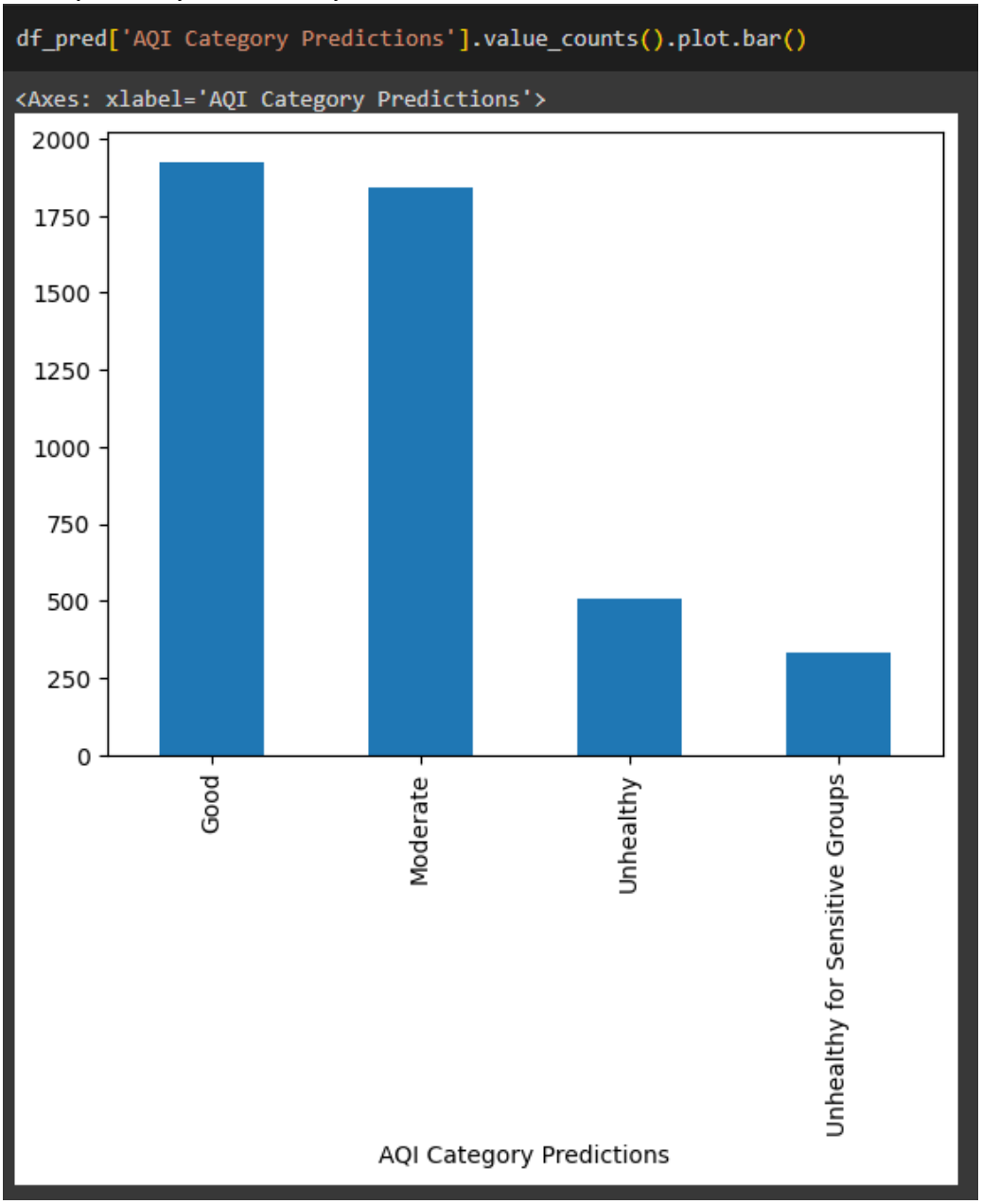
```
from sklearn.metrics import classification_report

print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1926
           1       0.00      0.00      0.00        45
           2       1.00      1.00      1.00      1841
           3       0.80      1.00      0.89       405
           4       1.00      1.00      1.00       333
           5       0.00      0.00      0.00        57

    accuracy                           0.98      4607
   macro avg       0.63      0.67      0.65      4607
weighted avg       0.96      0.98      0.97      4607
```

Графикон кој кажува колкав број од градовите во одредените држави ќе припаѓаат на категориите Good, Moderate, Unhealthy, Unhealthy for Sensitive Groups, Very Unhealthy во иднина:

```
df_pred['AQI Category Predictions'].value_counts().plot.bar()
```

```
<Axes: xlabel='AQI Category Predictions'>
```



Заклучок

Ова истражување за анализа и предвидување на глобалното загадување на воздухот во различни земји обезбеди значајни увиди во факторите што влијаат на квалитетот на воздухот и распределбата на нивоата на загадување низ целиот свет. Користејќи јавно достапно податочно множество, направивме детално претпроцесирање на податоците, вклучувајќи справување со недостасувачки вредности и нормализација, за да обезбедиме точност и конзистентност во нашата анализа.

Преку експериментирање со различни модели, утврдивме дека DecisionTreeClassifier беше најсоодветен за нашите потреби, нудејќи најголема точност и перформанси. Нашата првична анализа го покажа следното:

- Голем број градови, најмногу во Индија и Пакистан, се соочуваат со сериозни проблеми со загадувањето на воздухот.
- Земји како САД и Русија исто така имаат значајни градови со слаб квалитет на воздухот.
- Од друга страна, Бразил, Русија, САД и Германија имаат многу градови со добар квалитет на воздухот.

Идентификуваните трендови и корелации помеѓу различните загадувачи и целокупниот AQI можат да служат како основа за понатамошни истражувања и развој на поефективни мерки за контрола на загадувањето.

Овој проект ја нагласува важноста од континуирано следење и анализа за намалување на загадувањето на воздухот и неговите штетни ефекти на глобално ниво.

Користени алатки

Програмски јазик: Python

Библиотеки:
Pandas
matplotlib.pyplot
seaborn
sklearn