Air Pollutant Concentration Prediction over Ahmedabad Using Machine Learning

CSE523 - Machine Learning
Winter Semester 2023
Weekly Report - 12/3/2023

Yash Dahima - AU2129002

The tasks of *EDA* and *feature analysis* have been performed this week.

Exploratory Data Analysis (EDA):

The first few rows of the dataset looks like:

datetime	ws	temp	rh	dew_temp	precipitation	pressure	wv	blh	bcaod550	duaod550	omaod550	ssaod550	suaod550	aod469	aod550	aod670	aod865	aod1240	pm2p5
2011-01-01 05:30:00	2.75	11.05	61.08	3.83	0.00	1007.85	7.45	124.60	0.01	0.00	0.09	0.00	0.02	0.15	0.12	0.09	0.06	0.03	98.38
2011-01-01 08:30:00	3.29	12.51	55.55	3.85	0.00	1009.78	7.14	281.45	0.00	0.00	0.07	0.00	0.02	0.12	0.09	0.07	0.05	0.03	90.39
2011-01-01 11:30:00	3.73	22.32	28.96	3.40	0.00	1010.14	7.15	1059.94	0.00	0.00	0.05	0.00	0.01	0.09	0.07	0.05	0.04	0.02	28.07
2011-01-01 14:30:00	3.83	25.07	20.78	1.07	0.00	1006.98	7.51	1449.80	0.00	0.00	0.04	0.00	0.01	0.08	0.07	0.05	0.03	0.02	10.94
2011-01-01 17:30:00	2.46	23.68	23.36	1.53	0.00	1006.20	7.56	124.11	0.01	0.00	0.06	0.00	0.02	0.12	0.10	0.07	0.05	0.02	24.45
2011-01-01 20:30:00	2.46	16.01	45.20	4.14	0.00	1007.56	7.66	94.19	0.01	0.00	0.07	0.00	0.02	0.13	0.10	0.08	0.05	0.03	89.05
2011-01-01 23:30:00	2.83	13.96	51.64	4.15	0.00	1007.94	7.69	129.80	0.01	0.00	0.08	0.00	0.02	0.14	0.11	0.08	0.05	0.03	91.45
2011-01-02 02:30:00	3.02	12.33	55.92	3.78	0.00	1007.27	7.55	126.82	0.01	0.00	0.09	0.00	0.02	0.15	0.12	0.09	0.06	0.03	88.88
2011-01-02 05:30:00	2.74	11.43	57.15	3.24	0.00	1006.95	6.98	118.64	0.00	0.00	0.06	0.00	0.01	0.10	0.08	0.06	0.04	0.02	79.32
2011-01-02 08:30:00	3.06	13.64	50.45	3.53	0.00	1008.81	6.59	223.29	0.00	0.00	0.07	0.00	0.01	0.11	0.09	0.06	0.04	0.02	88.44
2011-01-02 11:30:00	2.88	22.97	30.40	4.65	0.00	1009.43	6.65	894.59	0.01	0.00	0.08	0.00	0.02	0.13	0.10	0.08	0.05	0.03	44.82
2011-01-02 14:30:00	2.44	25.78	21.60	2.19	0.00	1006.37	7.08	1339.41	0.01	0.00	0.09	0.00	0.02	0.15	0.12	0.09	0.06	0.03	20.90
2011-01-02 17:30:00	2.01	24.12	24.57	2.61	0.00	1006.32	7.10	128.25	0.01	0.00	0.10	0.00	0.03	0.18	0.14	0.11	0.07	0.03	35.54
2011-01-02 20:30:00	2.59	16.31	45.43	4.48	0.00	1008.26	7.13	71.80	0.01	0.00	0.10	0.00	0.03	0.18	0.14	0.10	0.07	0.03	91.52
2011-01-02 23:30:00	2.74	14.35	49.77	3.99	0.00	1008.77	7.31	77.45	0.01	0.00	0.10	0.00	0.03	0.17	0.14	0.10	0.07	0.03	96.04
2011-01-03 02:30:00	2.88	12.92	55.18	4.13	0.00	1008.03	7.76	134.39	0.01	0.00	0.12	0.00	0.03	0.20	0.16	0.12	0.08	0.04	93.22
2011-01-03 05:30:00	3.23	12.12	56.25	3.66	0.00	1007.72	8.21	176.72	0.00	0.00	0.08	0.00	0.02	0.13	0.10	0.08	0.05	0.03	71.12
2011-01-03 08:30:00	3.73	14.68	49.39	4.19	0.00	1009.91	9.10	283.47	0.00	0.00	0.08	0.00	0.02	0.13	0.10	0.08	0.05	0.03	62.04
2011-01-03 11:30:00	3.67	22.15	35.37	6.12	0.00	1010.60	10.55	687.30	0.01	0.00	0.09	0.00	0.03	0.17	0.13	0.10	0.06	0.03	31.07
2011-01-03 14:30:00	3.06	26.20	28.38	6.45	0.00	1007.51	12.51	1092.71	0.01	0.00	0.13	0.00	0.06	0.25	0.20	0.15	0.10	0.05	17.43
2011-01-03 17:30:00	2.61	24.56	30.41	6.04	0.00	1007.17	13.26	176.97	0.01	0.00	0.17	0.00	0.08	0.32	0.26	0.19	0.12	0.06	30.10
2011-01-03 20:30:00	2.96	18.06	49.75	7.40	0.00	1009.02	13.42	95.43	0.01	0.00	0.20	0.00	0.09	0.38	0.30	0.22	0.14	0.07	72.12
2011-01-03 23:30:00	2.94	16.67	50.48	6.34	0.00	1009.19	13.81	123.46	0.01	0.00	0.23	0.00	0.10	0.43	0.34	0.25	0.16	0.08	77.77
2011-01-04 02:30:00	2.88	14.75	55.42	5.91	0.00	1008.63	14.89	178.64	0.02	0.00	0.27	0.00	0.11	0.51	0.41	0.30	0.20	0.10	81.97
2011-01-04 05:30:00	3.04	14.06	55.24	5.22	0.00	1008.42	14.95	185.45	0.01	0.00	0.22	0.00	0.10	0.44	0.35	0.26	0.17	0.08	76.23
2011-01-04 08:30:00	3.36	14.67	51.78	4.85	0.00	1009.91	14.58	245.87	0.01	0.00	0.23	0.00	0.10	0.44	0.35	0.26	0.17	0.08	83.86
2011-01-04 11:30:00	2.89	22.36	39.64	7.97	0.00	1010.74	15.32	597.24	0.02	0.00	0.24	0.00	0.10	0.45	0.36	0.26	0.17	0.09	61.23
2011-01-04 14:30:00	2.95	26.76	34.14	9.65	0.00	1007.66	16.76	1013.93	0.02	0.00	0.25	0.00	0.10	0.47	0.38	0.28	0.18	0.09	42.62
2011-01-04 17:30:00	2.37	25.15	38.84	10.15	0.00	1006.51	17.15	174.24	0.01	0.00	0.18	0.00	0.08	0.34	0.27	0.20	0.13	0.07	40.91
2011-01-04 20:30:00	2.69	18.60	57.55	10.06	0.00	1008.42	16.54	79.40	0.01	0.00	0.17	0.00	0.08	0.33	0.27	0.20	0.13	0.07	86.26

The description of columns is as follow:

column name	full name	unit
datetime	Date & Time	yyyy-mm-dd hh:mm:ss
ws	Wind Speed at a height of ten metres above the surface of the Earth	meter/second
wd	Wind Direction at a height of ten metres above the surface of the Earth	0 (northward), 90 (eastward)
temp	Temperature of air at 2m above the surface of land, sea or in-land waters	degree Celcius
dew_temp	dew point temperature - Temperature to which the air, at 2 metres above the surface of the Earth, would have to be cooled for saturation to occur. It is a measure of the humidity of the air	degree Celcius
precipitation	Total Precipitation - Accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface	milimeter
pressure	Surface Pressure - Pressure (force per unit area) of the atmosphere on the surface of land, sea and in-land water. It is a measure of the weight of all the air in a column vertically above the area of the Earth's surface represented at a fixed point.	hPa
wv	Total Column Water Vapour	kg/m2
bcaod550	Black carbon aerosol optical depth at 550 nm	-
duaod550	Dust aerosol optical depth at 550 nm	-
omaod550	Organic matter aerosol optical depth at 550 nm	-
ssaod550	Sea salt aerosol optical depth at 550 nm	-
suaod550	Sulphate aerosol optical depth at 550 nm	-
aod469	Total aerosol optical depth at 469 nm	-
aod550	Total aerosol optical depth at 550 nm	-
aod670	Total aerosol optical depth at 670 nm	-
aod865	Total aerosol optical depth at 865 nm	-
aod1240	Total aerosol optical depth at 1240 nm	-
pm2p5	Particulate matter d < 2.5 µm (PM2.5)	ug/m3
blh	Boundary layer height - This parameter is the depth of air next to the Earth's surface which is most affected by the resistance to the transfer of momentum, heat or moisture across the surface.	meter
rh	Relative Humidity (derived)	%

In [3]:
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns, numpy as np, os

# Load the dataset
df = pd.read_excel('C:/Users/Yash Dahima/PhD/Course Work/ML/Project/AQI/Datasets/data4.xlsx', parse_dates= True)
```

In [4]:
```python
# Check the dimension of the dataset
print(df.shape)

# Check the first few rows of the dataset
print(df.head())
```

```
(26295, 20)
             datetime        ws       temp         rh  dew_temp  \
0 2011-01-01 05:30:00  2.748318  11.052734  61.081372  3.833771
1 2011-01-01 08:30:00  3.285797  12.509644  55.546595  3.849792
2 2011-01-01 11:30:00  3.731982  22.320221  28.955012  3.398590
3 2011-01-01 14:30:00  3.831188  25.073883  20.784734  1.071045
4 2011-01-01 17:30:00  2.461802  23.678711  23.357070  1.532501

   precipitation     pressure        wv          blh  bcaod550  duaod550  \
0       0.000148  1007.853027  7.451631   124.595856  0.005653  0.003228
1       0.000000  1009.779663  7.143795   281.450348  0.003787  0.002302
2       0.000000  1010.135498  7.146959  1059.936401  0.003310  0.001736
3       0.000000  1006.981384  7.514349  1449.800903  0.003369  0.001533
4       0.000000  1006.195496  7.559806   124.113068  0.005465  0.001438

   omaod550  ssaod550  suaod550    aod469    aod550    aod670    aod865  \
0  0.086111  0.001924  0.024665  0.151883  0.121593  0.090554  0.059739
1  0.066265  0.002047  0.019914  0.117729  0.094319  0.070313  0.046448
2  0.053539  0.001933  0.011374  0.089529  0.071884  0.053829  0.035817
3  0.044161  0.001899  0.014208  0.081334  0.065161  0.048617  0.032210
4  0.063244  0.001875  0.024782  0.121511  0.096802  0.071530  0.046574

    aod1240      pm2p5
0  0.032268  98.381378
1  0.025200  90.393173
2  0.019805  28.066624
3  0.017714  10.941221
4  0.024734  24.454000
```

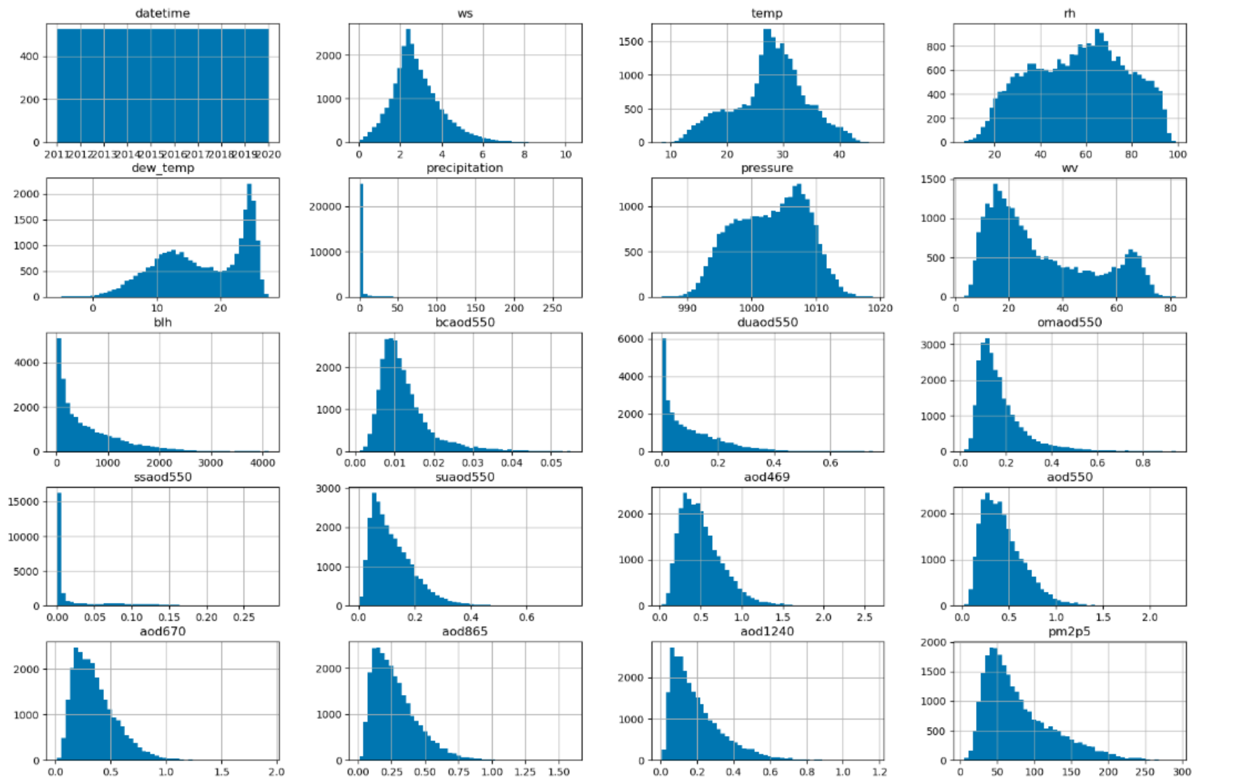
In [5]:
```python
# Check the basic information about the dataset
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26295 entries, 0 to 26294
Data columns (total 20 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   datetime       26295 non-null  datetime64[ns]
 1   ws             26295 non-null  float64
 2   temp           26295 non-null  float64
 3   rh             26295 non-null  float64
 4   dew_temp       26295 non-null  float64
 5   precipitation  26295 non-null  float64
 6   pressure       26295 non-null  float64
 7   wv             26295 non-null  float64
 8   blh            26295 non-null  float64
 9   bcaod550       26295 non-null  float64
 10  duaod550       26295 non-null  float64
 11  omaod550       26295 non-null  float64
 12  ssaod550       26295 non-null  float64
 13  suaod550       26295 non-null  float64
 14  aod469         26295 non-null  float64
 15  aod550         26295 non-null  float64
 16  aod670         26295 non-null  float64
 17  aod865         26295 non-null  float64
 18  aod1240        26295 non-null  float64
 19  pm2p5          26295 non-null  float64
dtypes: datetime64[ns](1), float64(19)
memory usage: 4.0 MB
```

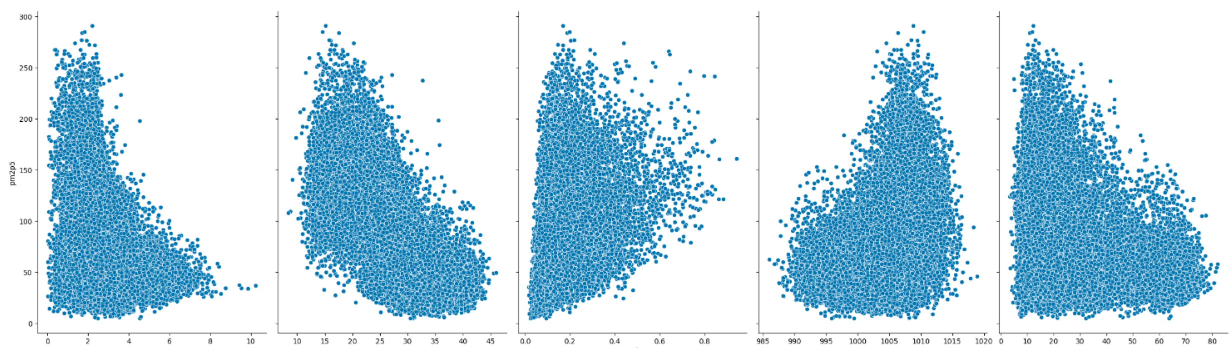
In [7]:
```python
# Check for missing values in the dataset
print(df.isnull().sum())
```

```
datetime         0
ws               0
temp             0
rh               0
dew_temp         0
precipitation    0
pressure         0
wv               0
blh              0
bcaod550         0
duaod550         0
omaod550         0
ssaod550         0
suaod550         0
aod469           0
aod550           0
aod670           0
aod865           0
aod1240          0
pm2p5            0
dtype: int64
```

In [11]:
```python
# Visualize the distribution of each variable using histograms
df.hist(bins=50, figsize=(20,13))
plt.show()
```



In [17]:
```python
# Visualize the relationship between the target variable and the other variables using scatterplots
sns.pairplot(df, x_vars=['ws', 'temp', 'omaod550', 'pressure', 'wv'], y_vars=['pm2p5'], height=7, aspect=0.7)
plt.show()
```



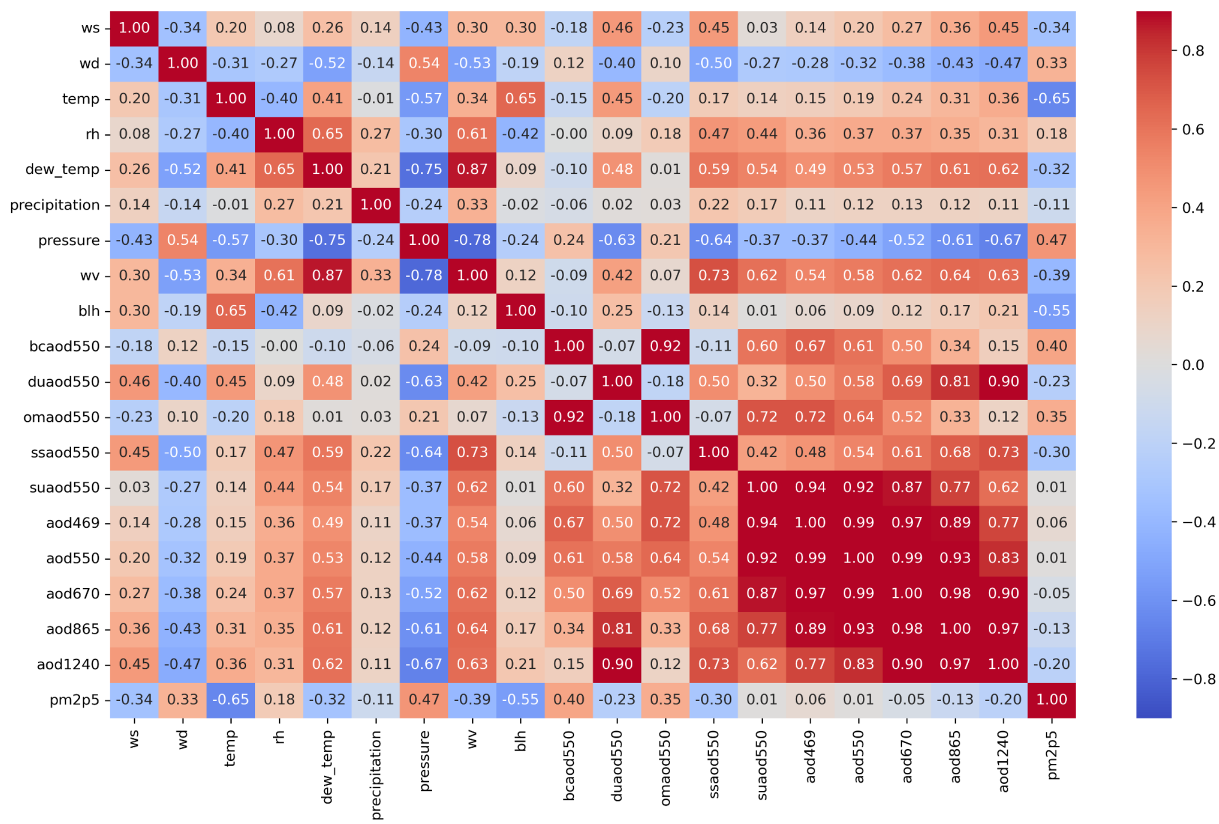
<u>**Feature Analysis:**</u>

Feature analysis and selection is performed primarily based on correlation matrix and principal component analysis (PCA).

Correlation Matrix:

In [14]:

```
# Visualize the correlations between variables using a heatmap
sns.heatmap(df.corr(), cmap='coolwarm', vmin=-0.9, vmax=0.9, annot=True, fmt='.2f')
plt.show(figsize=(20,13))
```



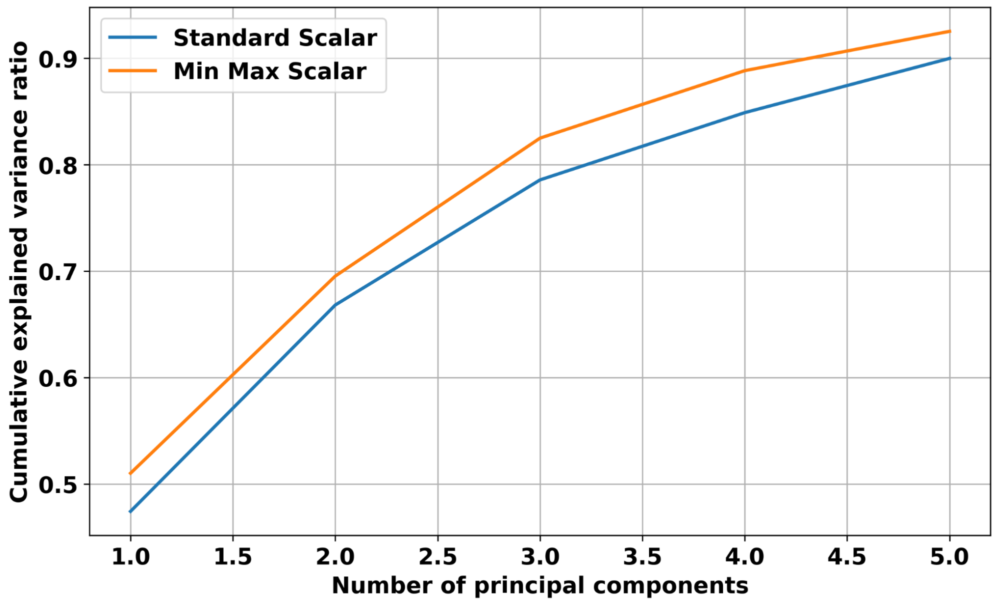
Target variable for prediction is PM2.5. As we can see, wind speed, temperature, pressure, blh, omaod are strongly correlated with it. We can also see that different AOD values are also correlated with each other. So we need only 1 or 2 types of AOD values. We can perform further feature selection by reducing dimensionality with PCA.

Principal Component Analysis (PCA):

```
import matplotlib as mpl
import pandas as pd, numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScaler
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_excel('C:/Users/Yash Dahima/PhD/Course Work/ML/Project/AQI/Datasets/data4.xlsx')
df['datetime'] = pd.to_datetime(df['datetime'])
df = df.set_index('datetime')

# Separate the features and the target variable
X = df.drop('pm2p5', axis=1)
y = df['pm2p5']

# Scale the data
std_scaler = StandardScaler()
min_max_scaler = MinMaxScaler()
X_scaled_std = std_scaler.fit_transform(X)
X_scaled_min_max = min_max_scaler.fit_transform(X)

# Perform PCA
pca_std = PCA(n_components=5)
pca_min_max = PCA(n_components=5)
X_pca_std = pca_std.fit_transform(X_scaled_std)
X_pca_min_max = pca_min_max.fit_transform(X_scaled_min_max)

# Calculate the explained variance ratio
explained_variance_ratio_std = pca_std.explained_variance_ratio_
explained_variance_ratio_min_max = pca_min_max.explained_variance_ratio_
```

```
# Plot the cumulative explained variance ratio
cumulative_explained_variance_ratio_min_max = np.cumsum(explained_variance_ratio_min_max)
cumulative_explained_variance_ratio_std = np.cumsum(explained_variance_ratio_std)

mpl.rcParams.update({'font.size': 14, 'font.weight':'bold', 'lines.linewidth':2})

plt.plot(np.array([1,2,3,4,5]), cumulative_explained_variance_ratio_std, label='Standard Scalar')
plt.plot(np.array([1,2,3,4,5]), cumulative_explained_variance_ratio_min_max, label='Min Max Scalar')

plt.xlabel('Number of principal components', fontweight ="bold")
plt.ylabel('Cumulative explained variance ratio', fontweight ="bold")
plt.legend()
plt.grid()
plt.show()



# Get the contribution of each original feature to each principal component
feature_contributions_std = pd.DataFrame(pca_std.components_, columns=X.columns)
feature_contributions_min_max = pd.DataFrame(pca_min_max.components_, columns=X.columns)
```

The following graph shows how much variance in the data is explained by how many principal components. It is clear that the first 5 principal components are able to capture almost 90% of the variance in the data.



The following table shows the contribution of each feature to the first 5 principal components using min max and standard scalars.

feature_contributions_min_max - DataFrame

Index	ws	temp	rh	dew_temp	precipitation	pressure	wv	blh	bcaod550	duaod550	omaod550	ssaod550
0	0.10	0.13	0.32	0.41	0.01	-0.31	0.52	0.04	0.01	0.20	0.03	0.28
1	0.13	0.53	-0.61	-0.02	-0.01	-0.21	-0.03	0.41	-0.12	0.21	-0.17	0.01
2	-0.06	-0.00	-0.23	-0.19	-0.01	0.17	-0.19	0.05	0.48	0.10	0.41	-0.06
3	0.36	-0.36	0.04	-0.29	-0.01	-0.02	-0.30	-0.14	-0.15	0.43	-0.25	0.35
4	0.30	-0.19	0.00	-0.28	0.03	0.17	0.26	0.54	0.03	-0.40	0.12	0.46

feature_contributions_std - DataFrame

Index	ws	temp	rh	dew_temp	precipitation	pressure	wv	blh	bcaod550	duaod550	omaod550	ssaod550
0	0.13	0.12	0.16	0.26	0.07	-0.24	0.27	0.06	0.11	0.24	0.12	0.25
1	-0.25	-0.27	0.05	-0.16	-0.05	0.30	-0.15	-0.21	0.46	-0.21	0.47	-0.17
2	-0.06	-0.40	0.56	0.22	0.26	-0.06	0.24	-0.45	-0.19	-0.18	-0.07	0.17
3	-0.31	0.41	-0.04	0.25	0.38	-0.15	0.28	0.31	0.04	-0.36	0.18	-0.14
4	0.52	-0.18	-0.10	-0.28	0.72	0.10	-0.10	0.19	0.10	-0.06	0.08	0.10

feature_contributions_min_max - DataFrame

Index	re	wv	blh	bcaod550	duaod550	omaod550	ssaod550	suaod550	aod469	aod550	aod670	aod865	aod1240
0		0.52	0.04	0.01	0.20	0.03	0.28	0.17	0.16	0.17	0.18	0.20	0.23
1		-0.03	0.41	-0.12	0.21	-0.17	0.01	-0.08	-0.05	-0.03	-0.00	0.04	0.10
2		-0.19	0.05	0.48	0.10	0.41	-0.06	0.28	0.32	0.30	0.27	0.22	0.17
3		-0.30	-0.14	-0.15	0.43	-0.25	0.35	-0.15	-0.01	0.03	0.09	0.18	0.30
4		0.26	0.54	0.03	-0.40	0.12	0.46	0.03	-0.00	-0.01	-0.04	-0.08	-0.12

feature_contributions_std - DataFrame

Index	re	wv	blh	bcaod550	duaod550	omaod550	ssaod550	suaod550	aod469	aod550	aod670	aod865	aod1240
0		0.27	0.06	0.11	0.24	0.12	0.25	0.29	0.31	0.32	0.33	0.33	0.32
1		-0.15	-0.21	0.46	-0.21	0.47	-0.17	0.24	0.23	0.19	0.12	0.01	-0.09
2		0.24	-0.45	-0.19	-0.18	-0.07	0.17	0.01	-0.08	-0.08	-0.08	-0.09	-0.09
3		0.28	0.31	0.04	-0.36	0.18	-0.14	0.19	-0.00	-0.04	-0.10	-0.17	-0.26
4		-0.10	0.19	0.10	-0.06	0.08	0.10	-0.05	0.01	0.01	0.01	0.00	0.00

The features will be finalized and the model will be attempted to be developed in the next week.