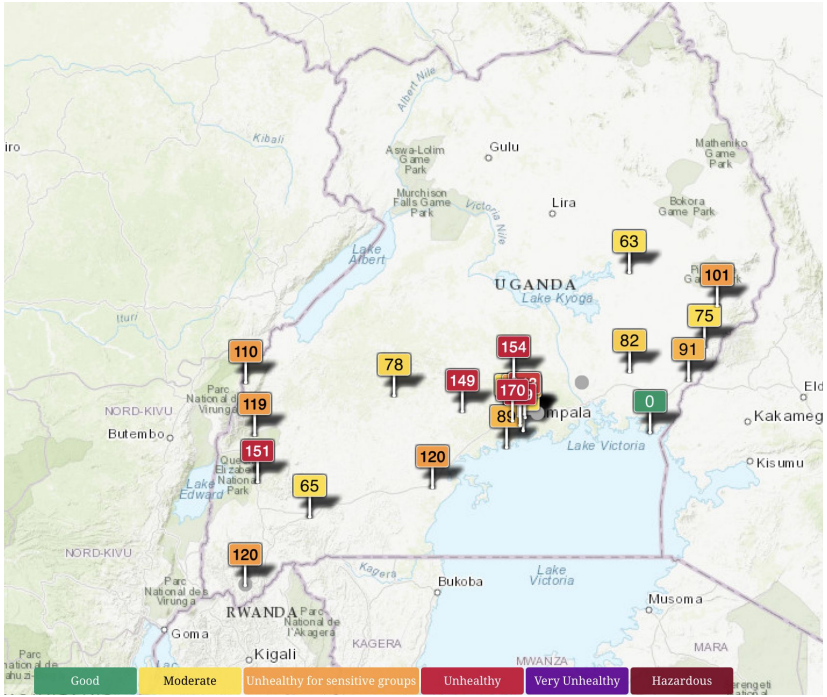
AIR QUALITY - ZINDI

Timorsha Rafiq-Dost, Aljoscha Wilhelm and Lina Willing

Background

PM2.5 Air Quality Index



Data

Features

Statistics	Weather Indicators
maximum	*Temperature*
minimum	*Precipitation*
mean	*Relative Humidity*
standard deviation	*Wind Direction*
variance	*Wind Speed*
median	*Atmospheric Pressure*
ptp (max-min)	
percentile	

- Data from 5 sensors
- 15539 time series over 5 days in train set
- hourly weather readings

Target PM2.5

Health Concern	$PM_{2.5}$ (μgm^{-3})	Precautions
Good	0 - 12	None
Moderate	13 - 35	Unusually sensitive people should consider reducing prolonged or heavy exertion
Unhealthy for Sensitive Groups	36 - 55	Sensitive groups should reduce prolonged or heavy exertion
Unhealthy	56 - 150	Everyone should reduce prolonged or heavy exertion, take more breaks during outdoor activities
Very Unhealthy	151 - 250	Everyone should avoid prolonged or heavy exertion, move activities indoors or reschedule
Hazardous	250 +	Everyone should avoid all physical activities outdoors.

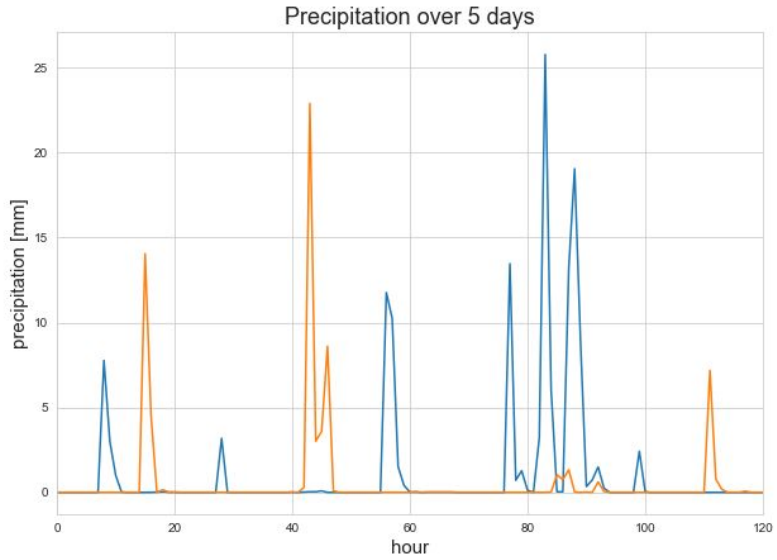
Imputing

- To each case the location is known
- At what time the data was taken is not known
- No relations between the cases can be made

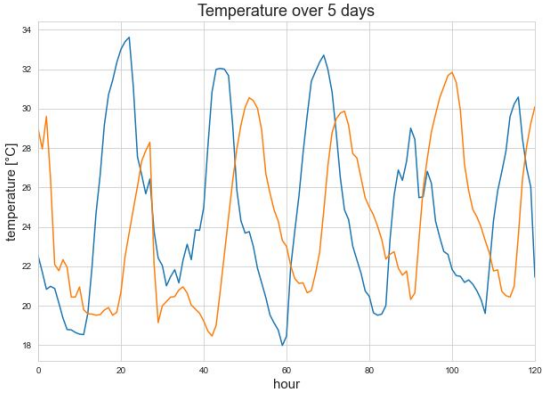
[nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,nan,26.90909091,27.20833333,26.18333333,24.7,23.65833333,22.74166667,22.15833333,21.55,21.16666667,21.0,20.925,20.35833333,19.84166667,19.275,19.53333333,19.58333333]

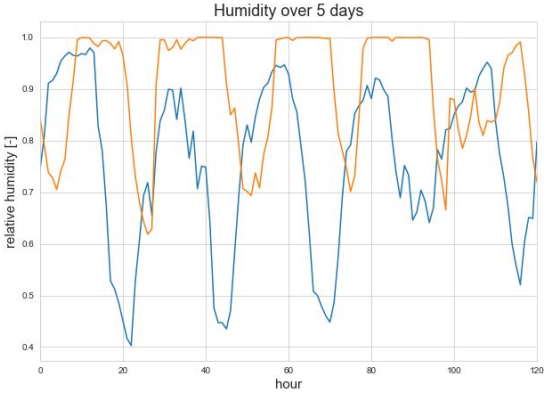
Imputing: Precipitation

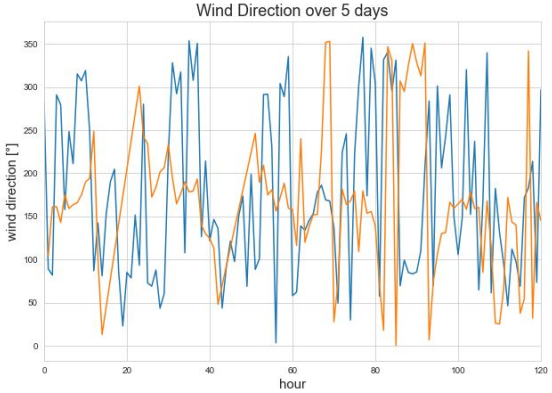
- Most values are 0
- If there is no data it is assumed that no rain has fallen
- NaN's will be set to 0

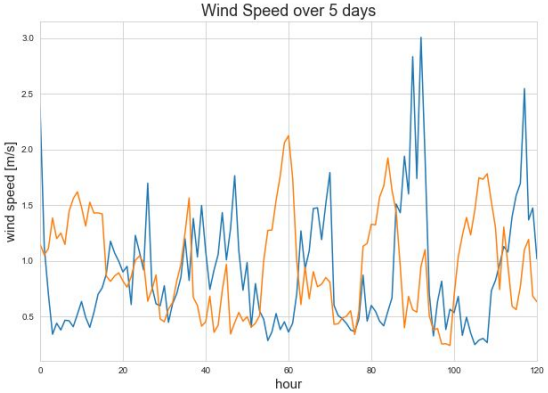


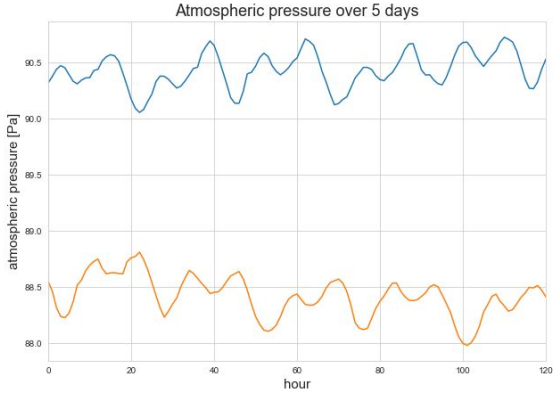
Imputing: Other Features











Imputing: Other Features

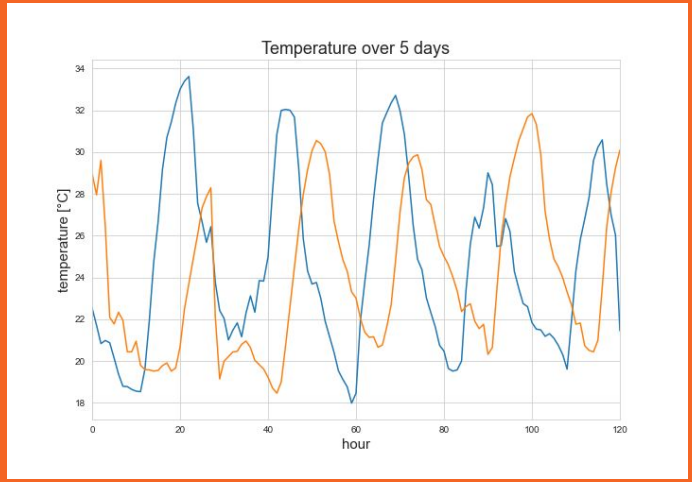
- All features show a periodic progression or at least a signal-shaped progression

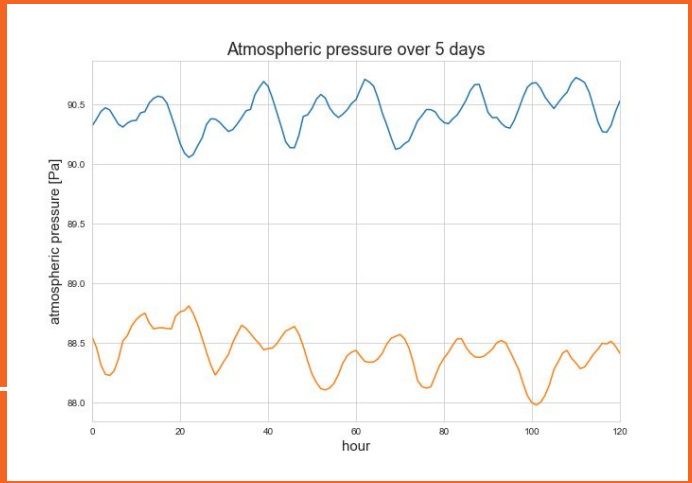
Approaches:

1. NaN's filled with mean's
2. NaN's calculated by Fourrier Transformation

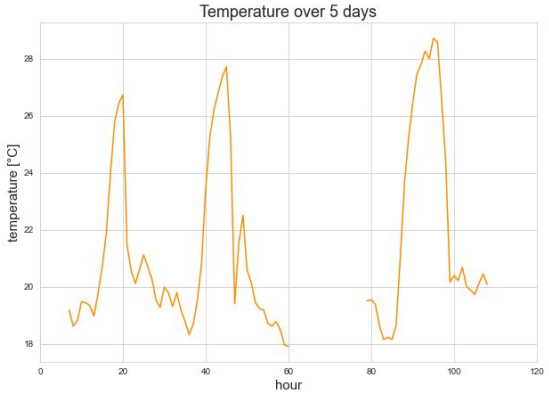
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi}{N}kn}$$

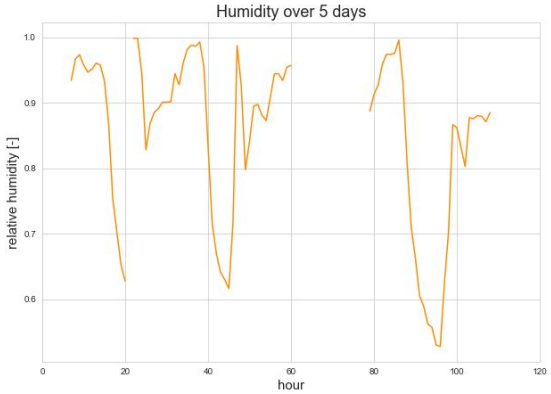
$$Y_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}kn}$$

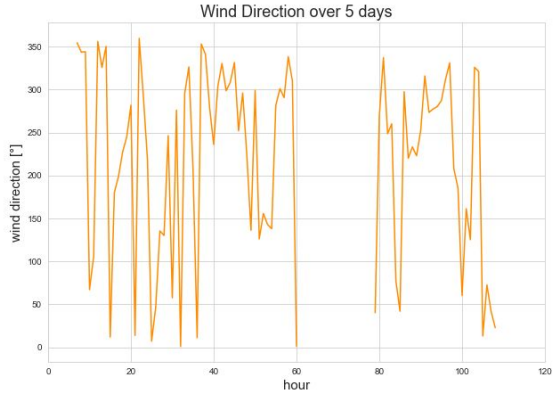


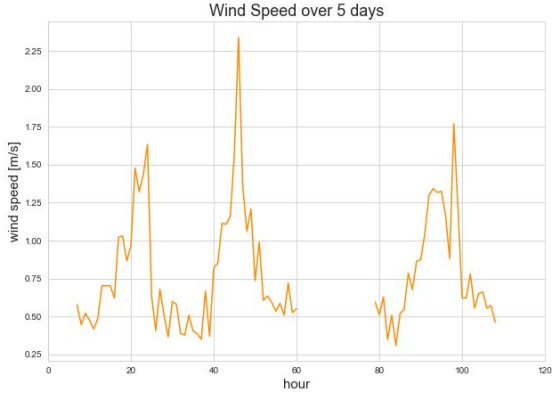


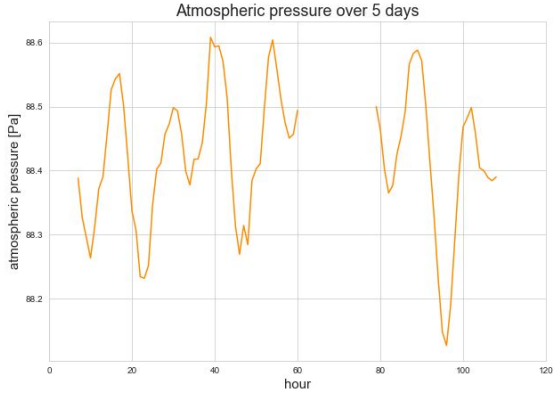
Imputing: Other Features



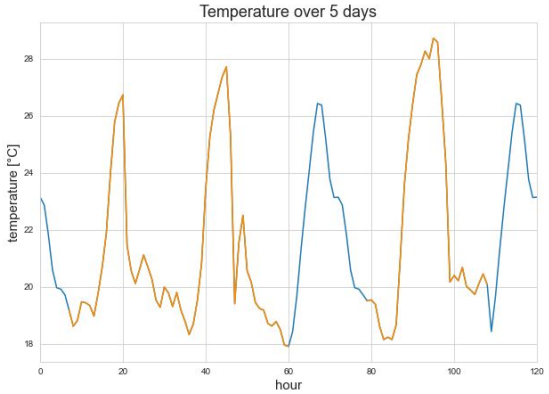


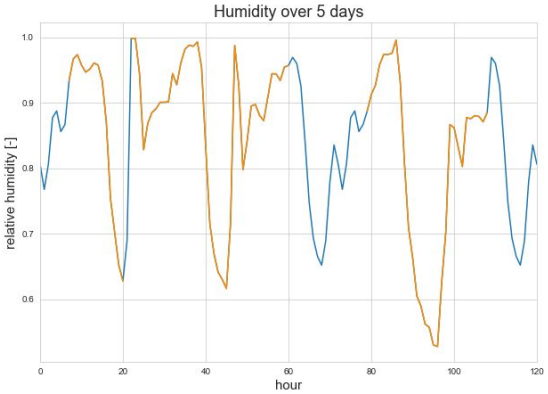


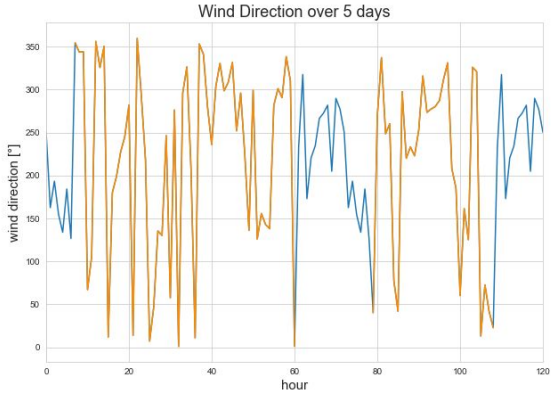


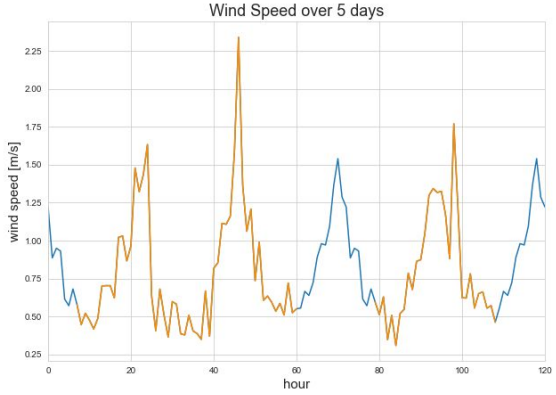


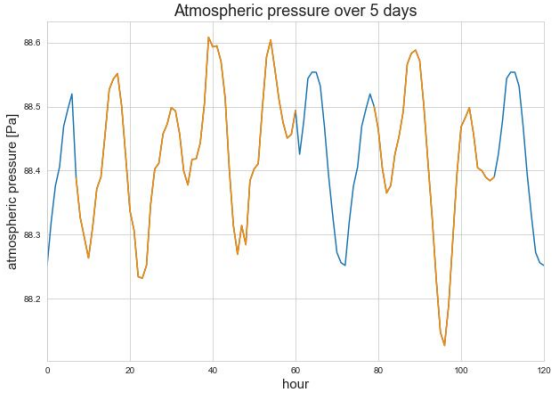
Imputing: Other Features





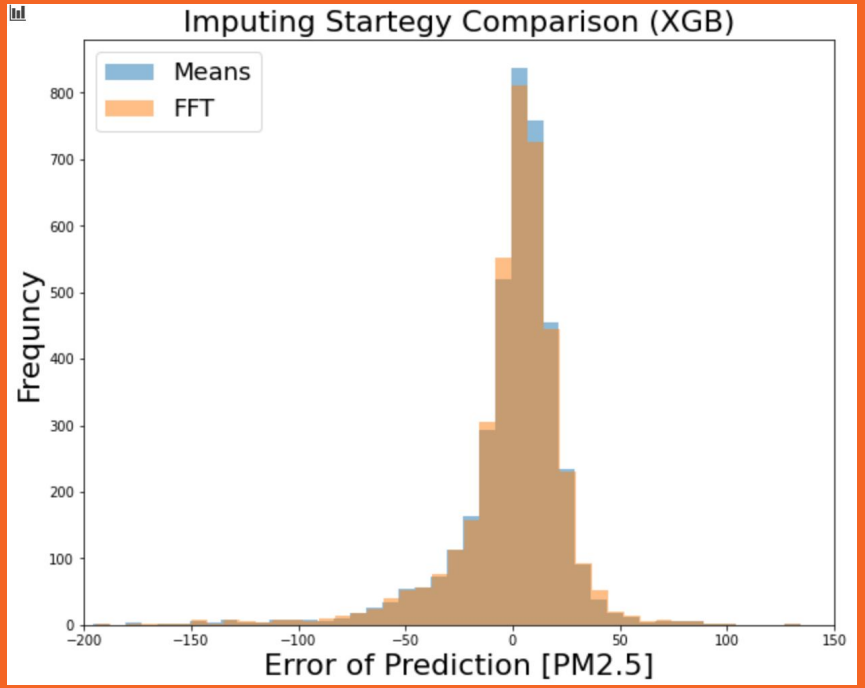






Applying Models

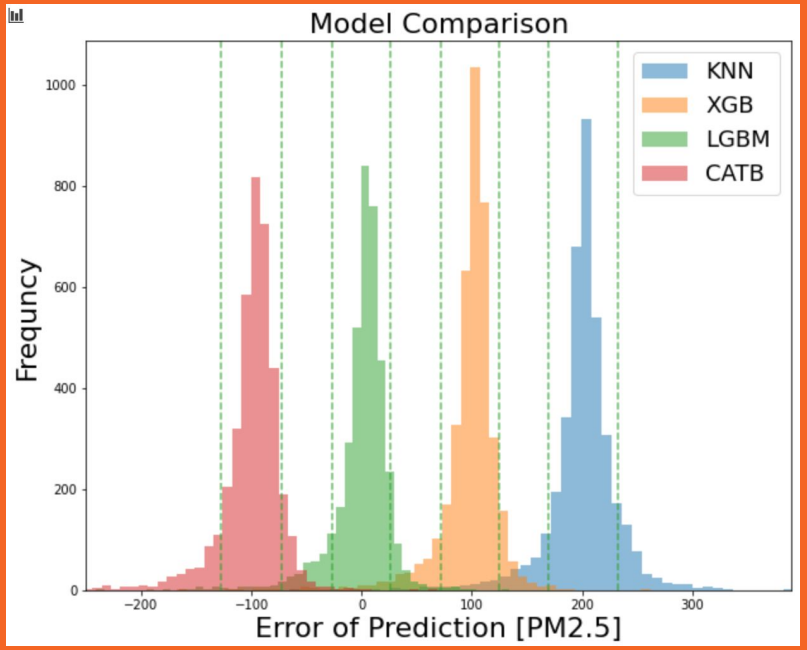
Both imputing strategies give mostly the same Results!



Applying Models

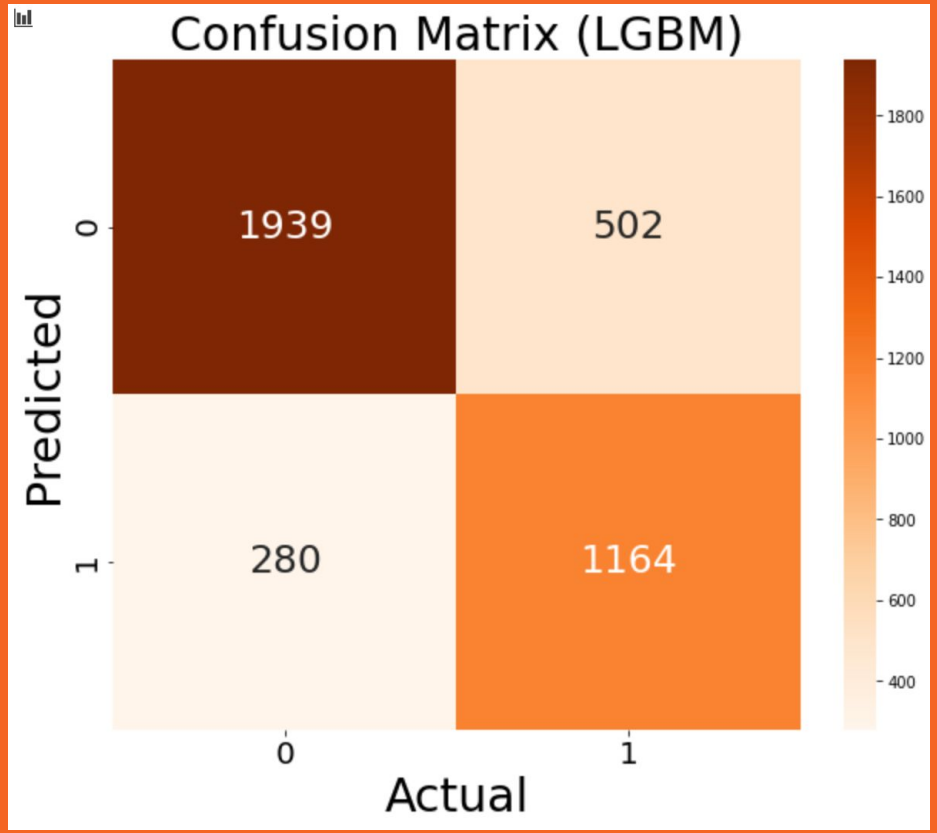
Error Ranges:

LGBM = 26.1
XGB = 26.2
CATB = 27.5
KNN = 32.0



Applying Models

Health Concern	$PM_{2.5}$ (μgm^{-3})	Precautions
Good	0 - 12	None
Moderate	13 - 35	Unusually sensitive people should consider reducing prolonged or heavy exertion
Unhealthy for Sensitive Groups	36 - 55	Sensitive groups should reduce prolonged or heavy exertion
Unhealthy	56 - 150	Everyone should reduce prolonged or heavy exertion, take more breaks during outdoor activities
Very Unhealthy	151 - 250	Everyone should avoid prolonged or heavy exertion, move activities indoors or reschedule
Hazardous	250 +	Everyone should avoid all physical activities outdoors.



Accuracy : 0.8
Recall: 0.81

Conclusion

- **Results enable prediction of Healthy vs Unhealthy**
- Imputing with FFT did not improve the results
- All models produce similar results
- Gridsearch and feature importance improved results only marginally

Outlook

- Data scaling and the use of other models
- Classification instead of regression and then handling of imbalance
- More complex models e.g. Arima Model for time series

Questions?