PROGRESS REPORT

Final Year Project 1

Semester 2, 2018/2019

Bachelor of Computer Science

A. Project Information

Student(s)

Syed Mohammed Khalid (1718487)

Project ID

534R

Project Title

Prediction of the Level of Air Pollution during wildfires using classification methods

Project Category

Research

Supervisor

Dr. Raini Binti Hassan

Table of Contents:

B. Introduction

Project Overview

Air pollution is becoming a major environmental issue in Malaysia. One of the main sources of Air pollution in Malaysia in trans-boundary pollution from neighboring countries.(Syed Abdul Mutalib et al., 2013), "Slash and burn" agricultural activities, deforestation and oil palm plantations on peat areas, particularly in Sumatra and Kalimantan, Indonesia are identified as the contributing factors to high intensity combustions that results in transboundary haze in Malaysia.(Latif et al., 2018). As a consequence of repeated haze episodes, the Malaysian government established the Malaysian Air Quality Guidelines, the Air Pollution Index, and the Haze Action Plan to improve air quality. Air quality monitoring is part of the initial strategy in the pollution prevention program in Malaysia. The Air Pollutant Index is used to identify and classify the ambient air quality in Malaysia based on the possible health implications to the public, if the API value raises then the level of danger or the status of API is raised to "Unhealthy" or "Extremely Unhealthy" , The API values provide a significant means to determine and evaluate the changes in pollution levels. In this project the aim is to be able to predict exposure of a population to pollution or the level of danger with regards to API, through the help of satellite-based data. This will be done with the help of the many different classification algorithms that have come to existence over the years. Generally the API is measured by Air Quality Monitoring Systems which compute the exposure of a population to air pollution, once it's computed it's made public as to whether it's "Good" or "Hazardous", however many regions which do not have any Air Quality Monitoring Systems (AQMS) in their respective locations may not get information about the health hazards in their region. The use of satellite-based data with help of machine learning could potentially be a means for these regions to know the level of air pollution where they live.

Problem Statement

Estimating the exposure of a population to air pollution in certain regions during wildfires can become a difficult task due to lack of ground-based Air Quality Monitoring Systems (AQMS).

Project Objectives

1. To understand the past works related to air pollution level prediction before, during and after wildfires using satellite-based data.
2. To discover techniques used, and to select the most suitable techniques in air pollution level prediction from the past.
3. To implement different selected classification algorithms for air pollution level prediction.
4. To evaluate the performance of the selected algorithms in order to find the optimal algorithm.
5. To perform feature selection in order to ensure the best performance of the optimal algorithm.

Significance of Project

Since the 1997 Indonesian forest fires disaster, forest fires have become recognized as a recurring problem, the most recent episode being the 2019 Kalimantan fires which saw more than 857,756 hectares (2.12 million acres) of land burned (Nangoy, 2019). Haze episodes in Malaysia have contributed to increasing hospital visits for treatments related to respiratory diseases. (Latif et al., 2018). With increase of air pollution, it is important to investigate and predict the air pollution levels or the unhealthiness as a result of it, hence leading to proper actions and controlling strategies so that the adverse effects to human health can be minimized.

Generally, a ground monitoring station equipped with an Air Quality Monitoring System (AQMS) would be responsible for the measurement of the air quality on a given day, however over the years only 65 Continuous Air Quality have come into existence in Malaysia, leaving many regions with close to no way to access it, in most cases a town which was not equipped with a AQMS would have the same API as the nearest town which has the device. An example of this from 2019 is, during the haze episodes residents of Miri, Sarawak raised their concerns concerning the lack of air quality monitoring in the division, as their town must depend on API readings from Samarahan air quality monitoring station (Chua, 2019). Usually the lack of AQMS in regions is because of the cost it takes to implement one and due to poor policy making.

Since, there isn't a cost-efficient alternative to AQMS, the possibility of being able to use satellite-based data comes as a relief. It would be cost efficient in the sense that satellites already exist and are continuously monitoring every region of our planet and sending back data to earth, most of which are available freely in the public domain. The proper use of this data can potentially save the cost of using AQMS and will ensure every region knows the levels and dangers of pollution in the respective region.

Project Schedule

The project schedule is broken down into two sections, one is for FYP 1 and another is for FYP 2. Both are provided in the Gannt chart below (Figure 1 & 2)

FYP 1

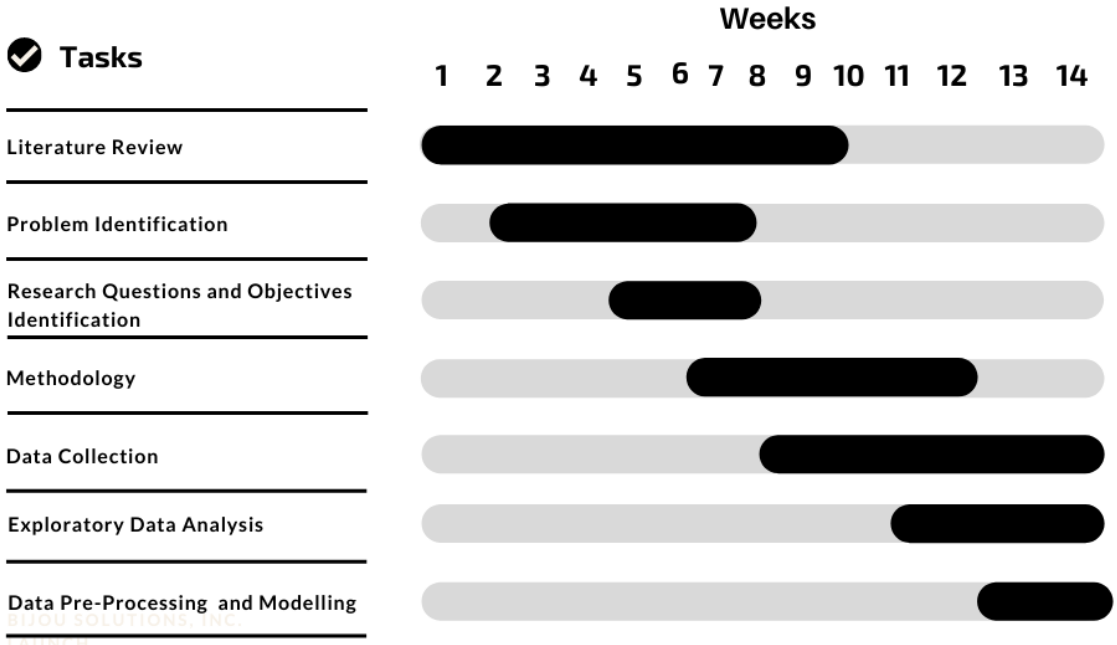

Figure 1: FYP 1 Gannt Chart

FYP 2

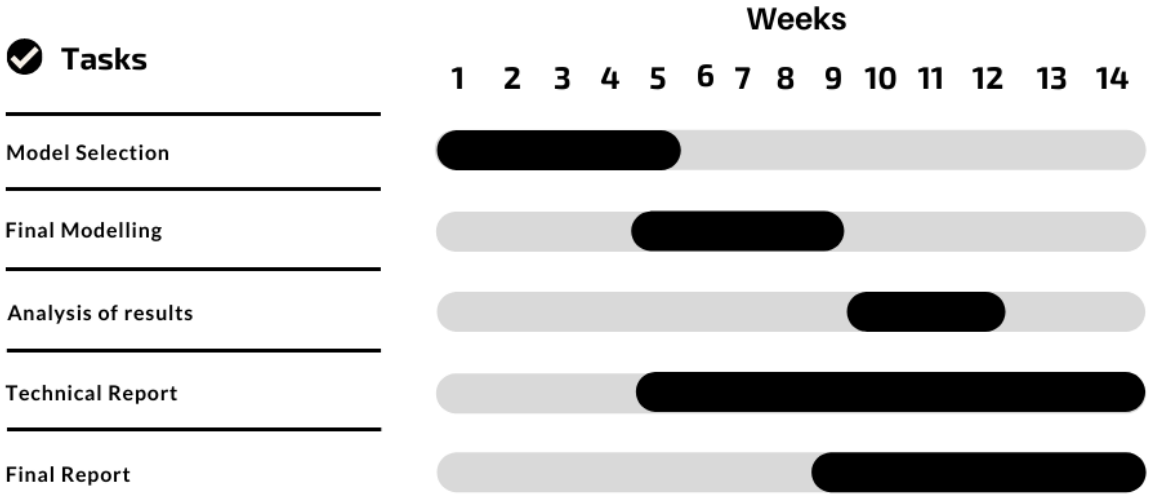

Figure 2: FYP 2 Gannt Chart

C. Review of Previous Works

Table 1: Literature Review Summary

No	Authors	Research problem/ application	Main techniques applied /algorithm	Results	Future Works
1	Akwasi Owusu-Akyaw, Richard Li, Courtney Moran (2019)	Predicting CO concentration of Californian counties near a major wildfire outbreak	-Linear regression -RANSAC Regression.	-Linear Regression MSE = 48.22 -RANSAC MSE = 38.14	- Implement time-series model - Gather more data for less populated counties - Include wind data in the model
2	Jun Wua, Arthur M Winera , Ralph J Delfinob (2006)	Predicting PM concentrations at a zip-code level for southern California before, during and after the 2003 wildfires	-Inverse distance weighting (IDW), Kriging or cokriging methods. -Regression	-	-
3	Colleen E. Reid, Michael Jerret (2015)	Spatiotemporal Prediction of Fine Particulate Matter During the 2008 Northern California Wildfires Using Machine Learning	-Generalized Boosting Model (GBM), Random Forest, Bagged Trees and 8 other algorithms. - 10-fold cross-validation (CV)	- The Generalized boosting model best performed with 29 predictor variables and achieved a CV-R^2 value of 0.803	-

4	Gongbo Chen a, Shanshan Li (2018)	Predicting PM2.5 concentrations across China with remote sensing, meteorological and land use information	- Random forests - Traditional regression models	-10-fold cross-validation: R^2 = 83%	-
5	Thongchai Kanabkaew (2013)	Prediction of Hourly Particulate Matter Concentrations in Chiangmai, Thailand using MODIS Aerosol Optical Depth and Ground-Based Meteorological Data	-Simple linear regression -Multiple linear regression	-Since Multiple Regression worked well, the model was selected. -The model performed reasonably with R^2 of 0.74	-Further study should apply the aerosol vertical profile to improve AOD and PM relationships. -The validity test should be also investigated for other areas to confirm the reliability and applicability of the model.
6	Manlika Sukitpaneenit & Nguyen Thi Kim Oanh (2014)	Analysis of monitoring for carbon monoxide and particulate matter during forest fire episodes in Northern Thailand	- Linear regression analysis	-	-the vertical profiles of satellite data should be incorporated. -longer data series should be used for the regression analysis.
8	Lim Ying Siew, Lim Ying Chin and Pauline Mah Jin We (2008)	ARIMA and integrated ARFIMA models for forecasting Air pollution index in Shah Alam, Selangor	-Time series models - Integrated Autoregressive Moving Average (ARIMA) and the Integrated Long Memory Model (ARFIMA) models.	-The integrated ARFIMA model performed slightly better than the ARIMA model.	-

9	Kar Yong Ng &Norhashidah Awang (2018)	Multiple linear regression and regression with time series error models in forecasting PM10concentrations in Peninsular Malaysia	- Multiple linear regression - Regression with time series error models	- RTSE is chosen as the best performing model	-
10	Azman Azid & Hafizan Juahir (2014)	Prediction of the Level of Air Pollution Using Artificial Neural Network Techniques	-Principal Component Analysis -ANN	-The PCA-ANN R^2 = 0.618 RMSE =10.017	-

Over the years there has been some significant work by researchers in relation to predicting air pollution levels in many regions, below we discuss some of the reviewed papers where some focus on prediction of air pollution levels during wildfires while others focus on air pollution in general circumstances.

If we were to look at the problem in recent times, researchers such as Akwasi, Richard and Courtney conducted their study as recently as last year, in their study they targeted the temporal pollution levels in Californian counties during the 2018 California wildfires, their study was more focused on how wildfires directly contribute to the increase and decrease in pollution levels. Their main predictor variables were distance from the fire, direction from the center of fire and size of fires. In their preliminary models they used two different algorithms, RANSAC and Linear Regression, the RANSAC model performed slightly better with a MSE of 38.14, they hence selected that algorithm. They later achieved better results with the addition of more predictor variables. However, their study only depended on fire and wind pattern data, there was no use of other factors with regards to spatiotemporal environment.

Wu, Winer and Delfino (2006) have looked at predicting PM (particulate matter) concentrations levels at zip code levels during the wildfires in Southern California. The one issue they were trying to address was that the PM concentrations were only recorded on every 3rd and 6th day of a week, hence it became necessary to perform spatial interpolation for the missing data. The methods used for spatial interpolation were inverse distance weighting (IDW), kriging or cokriging methods for the non-fire periods. Kriging was not superior to IDW

in many cases due to the small number of monitoring stations. Since the fire and smoke created highly heterogeneous pollution surfaces, typical IDW and kriging could not work during the fires.

Reid and Jarret (2015) must be one of the closest to what we are trying to achieve, in their work they focused on machine learning techniques to predict particulate matter (PM 2.5) for a California wildfire. Their work focuses on trying to find the optimal algorithm among Generalized Boosting Model (GBM), Random Forest, Bagged Trees, Elastic Net Regression, Multivariate Adaptive Regression Splines, Lasso Regression, Support vector machines, Gaussian processes and Generalized linear model using a 10-fold cross validation. Among the data used for their work, there was fire data, meteorological data and other spatiotemporal variables. In conclusion the best performing model for them was the Generalized boosting model with 29 predictor variables it achieved a CV-R^2 value of 0.803.

Sukitpaneenit and Oanh (2014) proposed in their work to study on how satellite data can be used to monitor carbon monoxide (CO) and particulate matter (PM) in Northern Thailand when forest fires occur, the authors acknowledge that forest fires known to be an important cause of air pollution. The target variables being the CO and PM concentrations were obtained from monitoring stations across the northern region of Thailand, and the main predictor variables were data obtained from the Moderate Resolution Imaging Spectroradiometer Satellite (MODIS) , which provided the Measurement of Pollution in the Troposphere (MOPPIT), Aerosol Optical Depth (AOD) and MODIS fire hotspots data. Their results showed that correlations between the ground-monitored CO and PM, respectively, with satellite monitoring data were in reasonable ranges in comparison with earlier studies conducted for other regions around the world. AOD and PM10 were generally better correlated (R=0.50–0.73) than MOPITT CO and ground monitored CO which correlated at (R=0.36–0.71).

Like Sukitpaneenit and Oanh (2014), Kanabkaew (2013) focused on the Thailand region. He indicates that the constant reoccurrence of forest fires Chiangmai and northern Thailand is a matter of concern, he acknowledges that lack of ground monitoring systems may cause unreliability for warning information, hence he suggests satellite remote sensing is now a good way of predicting air quality at the ground level. His study is focused on coming up with a satellite model to predict PM concentrations using satellite data. AOD data was collected from MODIS- Terra platform and ground level air quality were retrieved from ground stations. Two models were implemented using the data, the first model being single linear regression

and the second one being multiple linear regression. The second model gave a slightly better performance with $R^2$of 0.77 and 0.71, respectively for PM2.5 and PM10. In order to investigate the validity of the model, the regression equation obtained from the second model was then applied with smog data over Chiangmai in March 2007. The model performed decently with an R^2 of 0.74.

The next few papers reviewed were studies focused on the Malaysian air pollution levels. Yong and Awang (2017) in their paper used Multiple linear regression (MLR) and regression with time series error (RSTE) models in order to PM10 concentrations in Peninsular Malaysia. The predictor variables used by them were hourly temperature, humidity, wind speed and direction. If any missing values occurred, they used linear interpolation to perform the imputation. The performance of models at each station was evaluated by six indices, namely the root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), maximum absolute percentage error (maxAPE), fractional bias (FB) and percent bias (PBIAS). The overall evaluation statistics suggest that MLR and RTSE are comparable in their predictive ability. The lowest values of RMSE, MAPE and maxAPE for the MLR model showed that it is the best model. On the other side, RTSE is chosen as the best model in preference to the two MLR models by the MAE, FB and PBIAS. In addition, the positive FB and negative PBIAS indicate that all three models underestimated the actual concentrations in general. The RTSE model does not show any superiority over the simple MLR model.

Siew and Chin (2008), they start off by highlighting the danger of haze and how much it has affected Malaysia over the years. Hence, the objective of their project was to fit and illustrate the use of time series models in forecasting the API in Shah Alam, Selangor. The data used them in this study consisted of 70 monthly observations of API (from March 1998 to December 2003) published in the Annual Reports of the Department of Environment, Selangor. The time series models that were in consideration for selection were the Integrated Autoregressive Moving Average (ARIMA) and the Integrated Long Memory Model (ARFIMA) models. Through model evaluation it came to their attention that the integrated ARFIMA model is a better model as it has the lowest MAPE value. However, they noticed that the actual value of May 2003 falls outside the 95% forecast interval, this could be possibly due to emissions from mobile sources. industrial emissions, burning of solid wastes and forest fires.

Azid and Juhair (2014) in their study focused on the pattern recognition of Malaysian air quality based on the data obtained from the Malaysian Department of Environment (DOE). Eight air quality parameters in ten monitoring stations in Malaysia for 7 years (2005–2011)

were collected. They made use of Principal component analysis (PCA) was used to help in identifying the sources of pollution in the study locations. The combination of PCA and artificial neural networks (ANN) was modelled in order to determine its predictive ability for the air pollutant index (API). The PCA-ANN models showed a slightly better predictive ability in the determination of API with less variables, with R^2 and root mean square error (RMSE) values of 0.618 and 10.017, respectively.

From the literature review conducted above, it can be concluded that most of the research has been done in other regions, very little research has been done on the Malaysian region. Even the few papers which worked on the Malaysian region did not use the combination of fire data, meteorological data and AOD data. In this project the aim is to entirely use satellite-based data to evaluate the predictive ability of machine learning models in predicting the exposure of pollution to different regions before, during and after wildfires. Another thing one might notice is that all the previous studies have relied on regression algorithms for similar objectives. In this project the aim is to try to convert this into a classification problem in order to achieve better precision and accuracy, this will be done with help through data binning. Often the question arises as to how to pick the classes while performing the binning? This is made easier with the classes already being specified and standardized (Table 2) by the Malaysian Ambient Air Quality Standard (MAAQS).

Table 2: Health classifications used by the MAAQS

API	Air Pollution Level
0 -50	Good
51 - 100	Moderate
101 - 200	Unhealthy
201 - 300	Very unhealthy
301 - 500	Hazardous
500+	Emergency

D. Methodology

The workflow can be best understood with the help of the diagram below (Figure 3).

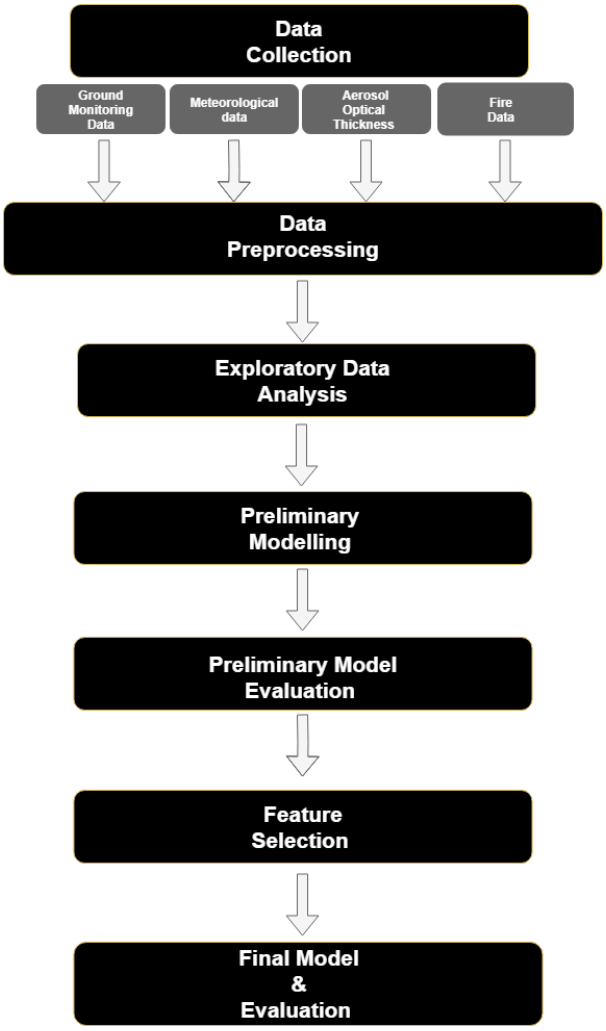

Figure 3: Project Workflow

As it has been illustrated in Figure 3 for the prediction of Air Pollution Level, different datasets are to be collected from multiple sources which contain air quality data, fire data, meteorological data and aerosol optical depth data for the date range of August 1st 2019 to October 31st 2019. The aim is to analyze the collected data in order to predict the level of air pollution. For this to become a possibility first, data preprocessing is done, and an exploratory data analysis is done to better understand the data which has been collected. Next the dataset is divided into training and testing sets, 80 percent and 20 percent respectively. For FYP 1 two algorithms are selected, one algorithm which has been successful and recommended by the reviewed literature, the other algorithm will be one which works best with architecture and the nature of the dataset. Both selected algorithms will be evaluated and compared accordingly and based on the result it will be determined if they are the best options, if not other alternative algorithms will be explored in FYP 2. After the selection of the optimal algorithm, feature

selection techniques will be implemented in order to identify the best set of predictors. Once the feature selection process is completed, the optimal algorithm will be used as the final model and it will be accordingly evaluated. The tool used for this project will mainly be Python.

E. Progress

In this section we will be looking into the progress made until now, and the challenges faced throughout the process in every step.

Literature Review

The literature review was conducted in order to discover different methods that have been used to achieve similar goals and to identify the problems they faced and their respective existing solutions. The literature which was studied helped to identify the gaps in the existing research. 10 papers were reviewed.

Data Collection

The process of data collection was initiated by the collection of daily ground-based monitoring data for Air Quality Index (AQI). In Malaysia the AQI is referred to as the Air Pollution Index (API), it is a simple and generalized way to describe the air quality, it is calculated from several sets of air pollution data. To determine the API for a given day, the sub-index values of 6 air pollutants which are particulate matter with the size of less than 10 micron (PM10), sulfur dioxide (SO2), carbon monoxide (CO), nitrogen dioxide (NO2), and ground level ozone (O3) as well as 1 additional parameter which is particulate matter with the size of less than 2.5 micron (PM2.5) are calculated based on the average concentration calculated. The maximum sub-index of all six pollutants is then selected as the API for the day. The API data this research was retrieved from the Air Quality Historical Data Platform (https://aqicn.org/data-platform/) which obtained its data from the Department of Environment, Ministry of Environment And Water. Since the research's focus is on the east Malaysia region data was collected from stations located in this region. 7 different stations were selected in an arbitrary manner.

List of selected stations

1. Kapit, Sarawak
2. Kota Kinabalu, Sabah
3. Kuching, Sarawak
4. Samarahan, Sarawak
5. Sibu, Sarawak
6. Sri Aman, Sarawak
7. Tawau, Sabah

The Aerosol optical depth (AOD) measurements were next retrieved from the Suomi National Polar-orbiting Partnership (SNPP) Visible Infrared Imaging Radiometer Suite (VIIRS) satellite, this satellite provides daily deep blue aerosol product provides satellite-derived measurements of Aerosol Optical Depth (AOD) and their properties over land and ocean as gridded aggregates, on a daily basis, globally. is provided in a 1° x 1° horizontal

resolution grid. Each data field, in most cases, represents the arithmetic mean of all the cells whose latitude and longitude coordinates positions them within each grid element's bounding limits. This data was in netCDF format and contained 44 Science Data Set (SDS) layers, hence the dataset required for the research needed to be extracted and converted into a csv format file.

The most important set of data needed for the research being the fire data, was retrieved with the help of the Visible Infrared Imaging Radiometer Suite (VIIRS) satellite where the fire layer shows active fire detections and thermal anomalies, such as volcanoes, and gas flares. The fire layer is useful for studying the spatial and temporal distribution of fire, to locate persistent hot spots such as volcanoes and gas flares, to locate the source of air pollution from smoke that may have adverse human health impacts.

Lastly the meteorological data for the selected region was retrieved from satellite data belonging to ERA5-Land, which is a reanalysis dataset providing a consistent view of the evolution of land variables over several decades at an enhanced resolution. All the above-mentioned datasets were downloaded for a specific time period of three months from August 2019 to October. The selection of the time period was on the basis that the haze episode in Malaysia occurred around the selected time period. They were also specific to the selected ground stations as in that case it would correspond with the air quality. Table 3 displays all the collected datasets and the variables.

Table 3: Collected Data

Dataset	Feature	Description
Air Quality Data	Date	The dates from 1st August 2019 to October 31st 2019
	Air Pollution Index	API readings from ground-based monitoring stations
Aerosol Optical Depth Data	Average Aerosol Optical Depth	A quantitative estimate of the amount of aerosol present in the atmosphere.
	Angstrom Exponent	A parameter which describes how the optical depth of an aerosol depends on the wavelength of the light.
	Spectral AOD Land	A quantitative estimate of the amount of aerosol present in the atmosphere.
	Spectral AOD Ocean	A quantitative estimate of the amount of aerosol present in the atmosphere.
Fire Data	Distance from the center of fires	The distance from the center of the all fires surrounding the stations.
	Distance from the nearest fires	The from the nearest fire from the stations
	Fire count	The count of fires surrounding the stations at a given time.

Meteorological Data	10m u-component of wind	Eastward component of the 10m wind.
	10m v-component of wind	Northward component of the 10m wind.
	2m dewpoint temperature	Temperature at 2 meters above the surface of the Earth.
	2m temperature	Temperature of air at 2m above the surface of land or sea.
	Forecast albedo	A measure of the reflectivity of the Earth's surface.
	Skin temperature	Temperature of the surface of the Earth.
	Surface latent heat flux	Exchange of latent heat with the surface through turbulent diffusion.
	Surface net solar radiation	Amount of solar radiation reaching the surface of the Earth.
	Surface net thermal radiation	Net thermal radiation at the surface.
	Surface pressure	Pressure of the atmosphere on the surface of land.
	Surface sensible heat flux	Transfer of heat between the Earth's surface and the atmosphere.
	Surface solar radiation downwards	Amount of solar radiation reaching the surface of the Earth.
	Surface thermal radiation downwards	Amount of thermal radiation emitted by the atmosphere and clouds that reaches the Earth's surface.
	Total precipitation	Accumulated liquid and frozen water, including rain and snow, that falls to the Earth's surface.

Data Preprocessing

The process of data preprocessing began with API dataset, however the API dataset only contained two columns, one being date and other the AQI and most of data was clean and did not contain any missing values, hence very limited treatment was performed to this dataset. Since our problem is a classification problem the target cannot be a continuous variable hence the API went through the data binning process which is a way to group a number of more or less continuous values into a smaller number of "bins" resulting in a categorical variable. The data binning process was done in accordance with the Malaysian Ambient Air Quality Standard (MAAQS). Table 2 shows the different classes to which the continuous can be binned into, the

newly created variable will be called “Air Quality Level” and it will act as the target variable for the future selected models.

The next dataset, which was the Aerosal Optical Thickness, when it was collected the data was in NetCDF format. Generally, the NetCDF format is popular for storing multi-dimensional data. The task was to see what types of variables are inside the dataset and accordingly extract the data. When looking at the literature that has been reviewed, there was no mention of how they dealt with this kind of file, however after tirelessly searching online for some type of tool which will help in identifying the attributes inside this dataset, a tool named ‘Panoply’ developed by the National Aeronautics and Space Administration (NASA) was discovered. Using this tool, it confirmed the existence of more than 40 different attributes (Figure 4).

Name	Long Name	Type
AERDB_D3_VIIRS_SNPP.A2019213.001.2...	SNPP VIIRS Deep Blue Level 3 daily aerosol data, 1...	Local File
Aerosol_Optical_Thickness_550_Land_...	number of retrievals used for aerosol optical thickn...	Geo2D
Aerosol_Optical_Thickness_550_Land_...	maximum aerosol optical thickness estimated at 55...	Geo2D
Aerosol_Optical_Thickness_550_Land_...	aerosol optical thickness estimated at 550 nm over ...	Geo2D
Aerosol_Optical_Thickness_550_Land_...	minimum aerosol optical thickness estimated at 550...	Geo2D
Aerosol_Optical_Thickness_550_Land_...	number of retrievals used for aerosol optical thickn...	Geo2D
Aerosol_Optical_Thickness_550_Land_...	maximum aerosol optical thickness estimated at 55...	Geo2D
Aerosol_Optical_Thickness_550_Land_...	aerosol optical thickness estimated at 550 nm over ...	Geo2D
Aerosol_Optical_Thickness_550_Land_...	minimum aerosol optical thickness estimated at 550...	Geo2D
Aerosol_Optical_Thickness_550_Land_...	standard deviation of aerosol optical thickness esti...	Geo2D
Aerosol_Optical_Thickness_550_Land_...	standard deviation of aerosol optical thickness esti...	Geo2D

Figure 4: Panoply

Despite being able to view the attributes, the tool couldn’t be used to extract the dataset into a csv format. Hence a python script was written in order to extract the data from 92 different files each corresponding to a different day of the selected time period. This process was performed for all the selected 7 stations. The next immediate task was to select the most relevant attributes from the given 40, the selection was made, and 4 columns were selected, the basis of the selection was relevancy and amount of missing data. Although the selected columns had the least amount of missing data among all the columns, they still had a considerable amount of missing data and hence required interpolation. When the API was plotted against different selected attributes of the AOD (Figure 5) it was noticed that the relationship was linear and that the increase and decrease in AOD took place in a linear manner, hence it was established that linear interpolation was the best way to treat the missing data.

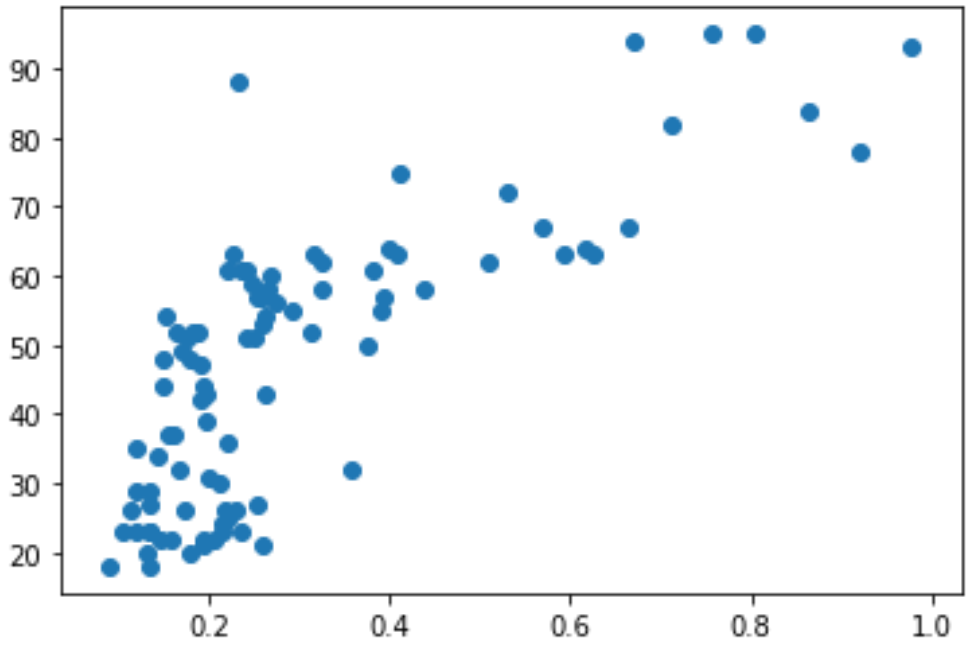

Figure 5: Scatter plot (AOD vs AQI)

The next important dataset was the fire dataset, generally if we study the literature we can notice that fire data is available with name, date, size and location of the fire but this was the case for data in regions in the US or Europe. However, for Malaysia and Indonesia there has been no convention of naming fires despite there being many instances of wildfire incidents in these regions. Hence for this research it was incumbent to depend on satellite detected hotspots or thermal anomalies, as mentioned earlier hotspots do not necessarily mean there is fire in a certain location, it could also mean there's a volcanoes or gas flares, this isn't a problem as both volcanos and gas flares in some way contribute to pollution in a similar manner to fires.

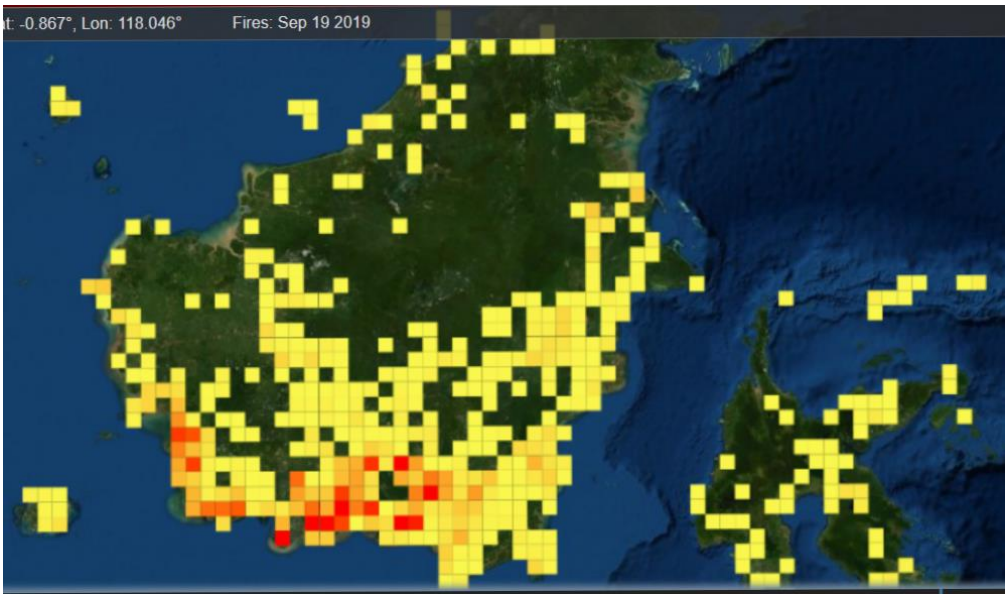

Figure 6: Fires at a given time

The problem that was encountered when working with fire data is that at a given day there could be multiple fires occurring (Figure 6). However, in our dataset we can only include information of one fire at a given day, but all the fires on the given day play a part in the contribution of pollution. After many discussions and consultations with lecturers, it was decided that the center of all the fires would be taken as the location of the fire, this was done with assumption that the effect of the fires will be the highest at the center. So, it was decided that the distance from the center of all the fires, the count of the fires in that region and the distance from the nearest fire. In the dataset fires had two main attributes, one being the Confidence which is intended to help users gauge the quality of individual hotspot/fire pixels. Confidence values are set to low, nominal and high each indicating the quality of the fires, low confidence fires were removed from the dataset, the other attribute being Fire Radiative Power (FRP), the FRP depicts the intensity of the fire any fire with an FRP less than average FRP was removed in the final dataset. There are close to no missing data in the dataset hence no treatment was done.

The last dataset would be the meteorological data, this dataset like the AOD dataset was in the NetCDF format, hence it had to be converted, however the earlier script would not work for this dataset due to some differences, hence a brand new python script was used to extract the data. All the datasets were next combined into one dataset, and a correlation test was performed in order to study the linear dependence between the predictors and the target. The results suggested which features had low correlation with the target and these features were removed accordingly. A total of 12 features were removed from the dataset. The final dataset had a total of 10 predictors and one target feature.

Exploratory Data Analysis

Once completed with data preprocessing, the next stage would be to perform exploratory data analysis which is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. It is important to understand the types of data we are working with in order to understand the architecture of the data at hand, this will help in selecting the best algorithm. We first take a look at the behavior of API over the three selected months, plotted below (Figure 7) is the changes in API from August 2019 to October 2019 in all the selected stations, as we can see between day 30 to day 60 the API seems to be at the peak for all of the stations, however the extent to which the API increased differs from station to station, one example can be Kota Kinabalu where the change in API was very minimal, while Sri Aman had a drastic change. The plot indicates that the geographic location of the stations really matters, each of them might be cities which are slightly near to each other, but the pollution affects them differently.

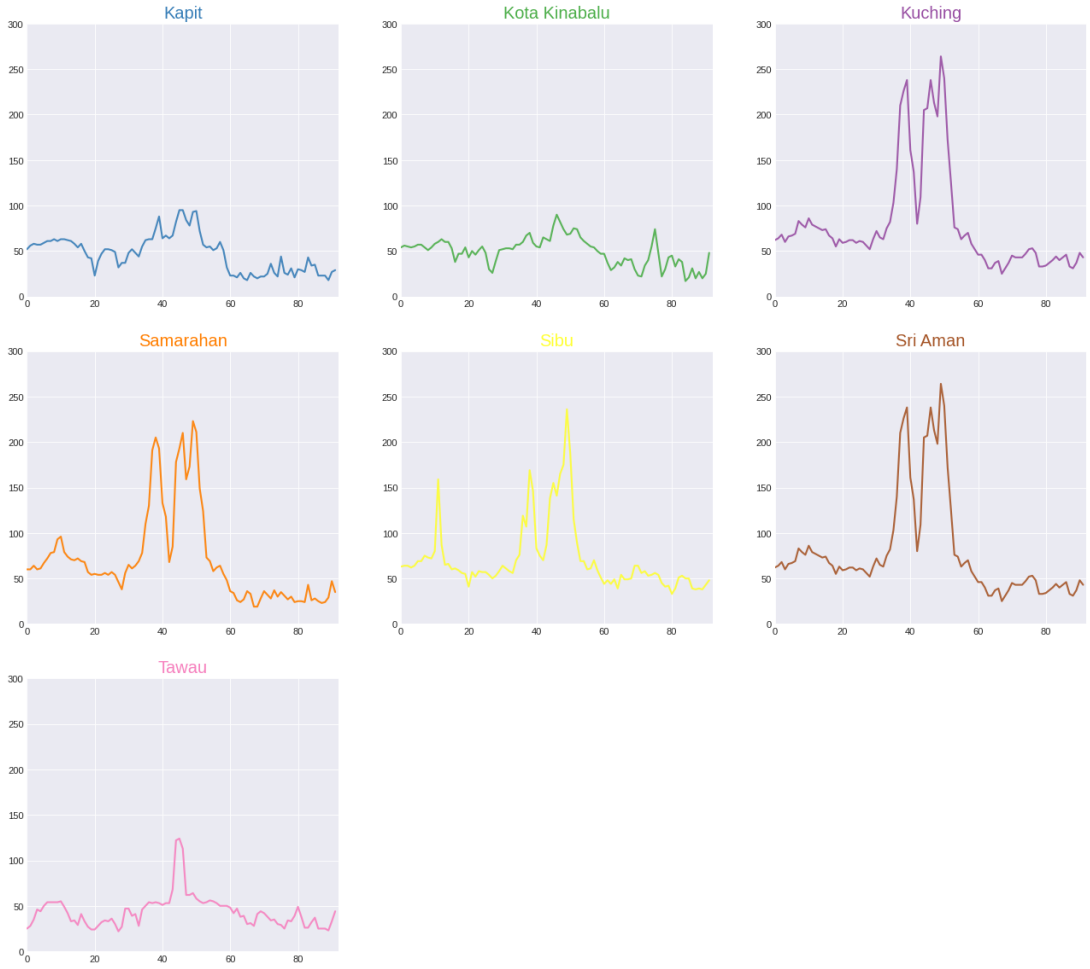

Figure 7: Line graph (API vs Date)

Next came the visualization of the newly created variable “Air Quality Level”, this plotted pie chart (Figure 8) indicates that the class “Very Unhealthy” occurs very few times, which points out that the dataset being worked on is an imbalanced dataset. An imbalanced dataset is a special case for classification problem where the class distribution is not uniform among the classes. Specific algorithms can handle imbalanced datasets well, however there are very few of them, there are other techniques to deal with imbalanced datasets, two of the most popular methods include Random Under/Over-sampling and using class weights during the modelling stage.

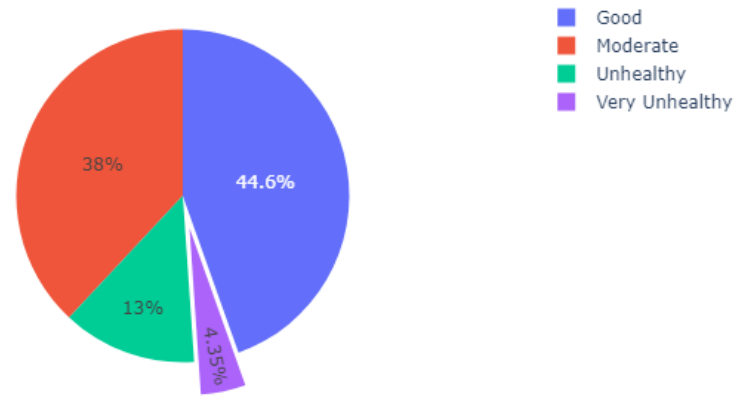

Figure 8: Air Quality Level class distribution

The next and final visualization which was done was in relation to the fire data, in order to investigate the relationship between API and the count of fires, we notice from the plot (Figure 9) that it's not necessary that when the number of fires surrounding increases the API should increase too, the reason for this relationship not being linear in nature is because the distance from the fires matters. The count of fires won't affect the selected region to great extent if the fires are located at a great distance from it. This highlights the importance of the distance from fires as a feature.

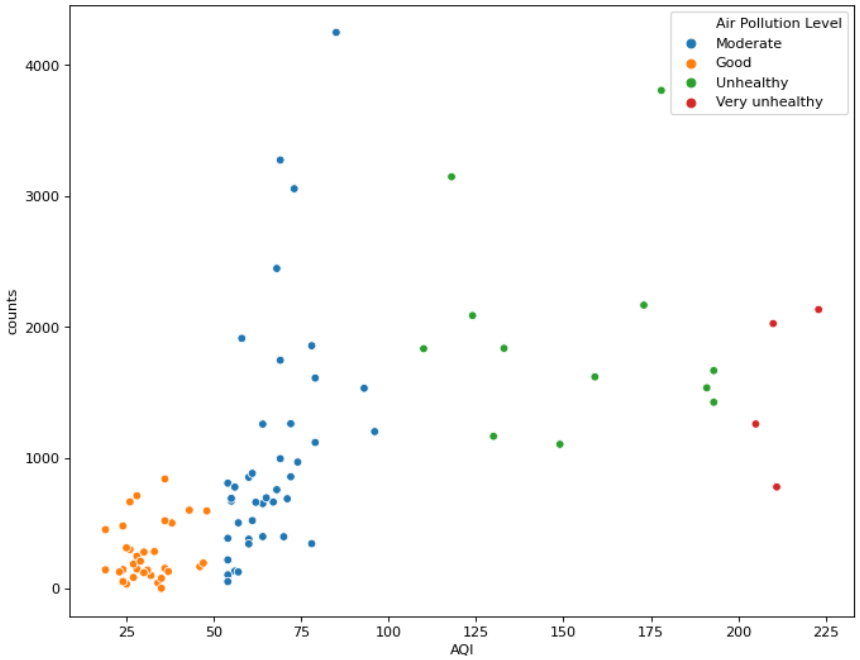

Figure 9: Scatter Plot (Counts vs API)

Preliminary Modelling and Preliminary Results

After the exploratory data analysis stage, the immediate next step would be to proceed to the preliminary modelling stage, which involves selecting two different classification algorithms and evaluating their performance for each of the selected stations. As a result of the data preprocessing performed earlier the remaining were 10 predictor features and one target (Table 4) were used to perform the modeling. The earlier used "API" feature was removed from the dataset as it will be no longer be used following the introduction of "Air Quality Level".

Table 4: Selected Predictors and Target Variable

Predictors	Average Aerosol Optical Depth
	Angstrom Exponent
	Spectral AOD Land
	Spectral AOD Ocean
	Fire count
	Skin temperature
	Surface net solar radiation
	Surface solar radiation downwards
	2m temperature
	Forecast albedo
Target	Air Quality Level

Two different algorithms were selected and trained accordingly is similar circumstances. The algorithms selected were Random Forest and Gradient Boosting. In order to be able to evaluate the models, a train-test split procedure was performed, this acts an estimate of the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow one to compare the performance of machine learning algorithms for their predictive modeling problem. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset. For this project the common split percentage of Train dataset: 80%, Test dataset: 20% was performed. The estimated accuracy of each of the models is highlighted in Table 5.

Table 5: Evaluation results (Accuracy)

Algorithms	Kapit	Kota Kinabalu	Kuching	Samarahan	Sibu	Sri Aman	Tawau
Random Forest	0.929	0.750	0.893	0.750	0.821	0.857	0.928
Gradient Boosting	0.964	0.821	0.857	0.785	0.714	0.857	0.892

Random Forest

Random forests is a supervised learning algorithm, which is an ensemble learning method for classification or regression. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It has also become one of the most popular algorithms, because of its simplicity and its flexibility in terms or working on both regression and classification problems. Based on literature review conducted, random forest was one of the best performing algorithms when it comes to dealing with air quality data, hence this algorithm was selected, Figure 10 highlights the performance of random forest using the confusion matrix. As we can see there are 3 misclassifications which occurred (Figure 10). Figure 11 & 12 are the comparisons of how the algorithm fared with data from a different station.

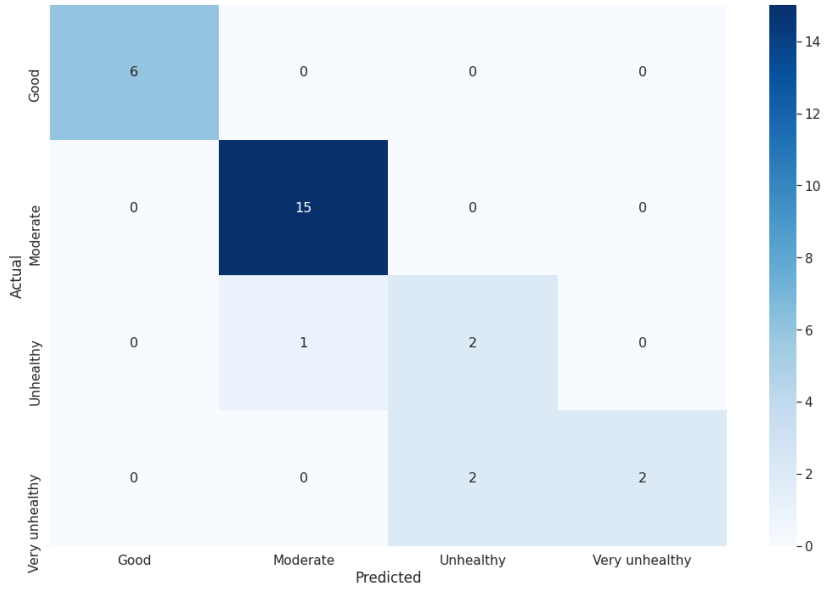

Figure 10: Confusion Matrix (Random Forest)

Figure 11: Actual Air Quality Level

Figure 12:: Predicted Air Quality Level (Random Forest)

Gradient Boosting

Like random forest, gradient boosting is a supervised learning algorithm which is also an ensemble learning method for classification and regression. Boosting is a method of converting weak learners into strong learners. In boosting, each new decision tree is a fit on a modified version of the original data set. Gradient Boosting trains many models in a gradual, additive and sequential manner. The motivation for selecting gradient boosting is that it is classification model that has built-in approaches in order to combat class imbalance, as noticed earlier through the visualizations the dataset is quite imbalanced hence gradient boost will help with this constructing successive training sets based on incorrectly classified examples. Figure 13 is the resulting confusion matrix when gradient boosting was performed for a given location. The confusion matrix indicates that there were 4 misclassifications.

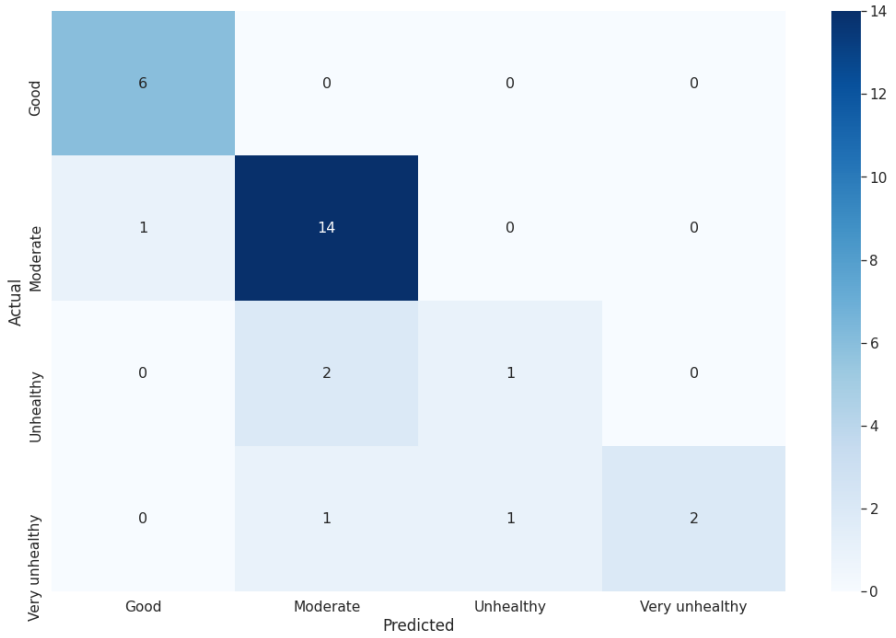

Figure 13: Confusion Matrix (Gradient Boosting)

Future work

For FYP 2, the work will continue from where it has been left off. Model selection, feature selection and model evaluation will be the main focuses of FYP 2. Further evaluation techniques will be used to estimate the performance of the current models, if the current models fail to impress, other classification algorithms may be explored. Other tasks which must be performed. are that a proper measure for distance from fires to the stations must be discovered, as the method used in the initial preprocessing stage resulted in an extremely low correlation with API and more predictor variables may be added in the future if required.

Acknowledgement

I would like to express my deepest appreciation to my supervisor Dr. Raini Binti Hassan for her constant guidance throughout this project. I would also like to extend my gratitude to Dr. Rawad Abdulghafor and the FYP Coordinator Dr. Norzariyah Binti Yahya for helping me with the project.

References

Mutalib, S. N. S. A., Juahir, H., Azid, A., Sharif, S. M., Latif, M. T., Aris, A. Z., ... & Dominick, D. (2013). Spatial and temporal air quality pattern recognition using environmetric techniques: a case study in Malaysia. *Environmental Science: Processes & Impacts*, *15*(9), 1717-1728.

Latif, M. T., Othman, M., Idris, N., Juneng, L., Abdullah, A. M., Hamzah, W. P., ... & Sahani, M. (2018). Impact of regional haze towards air quality in Malaysia: a review. *Atmospheric Environment*, *177*, 28-44.

Nangoy, F. (2019, October 21). Area burned in 2019 forest fires in Indonesia exceeds 2018 - official. Retrieved August 15, 2020, from https://www.reuters.com/article/us-southeast-asia-haze/area-burned-in-2019-forest-fires-in-indonesia-exceeds-2018-official-idUSKBN1X00VU

CHUA, S. (2019, September 24). MCAQM station arrives in Serian to provide accurate data on air quality in the division. Retrieved August 15, 2020, from https://www.theborneopost.com/2019/09/24/mcaqm-station-arrives-in-serian-to-provide-accurate-data-on-air-quality-in-the-division/

Owusu-Akyaw, A., Li, R., & Moran, C. (2019). Spread of Wildfire Pollutants in California.

Wu, J., Winer, A. M., & Delfino, R. J. (2006). Exposure assessment of particulate matter air pollution before, during, and after the 2003 Southern California wildfires. Atmospheric Environment, 40(18), 3333-3348.

Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., ... & Balmes, J. R. (2015). Spatiotemporal prediction of fine particulate matter during the 2008 Northern California wildfires using machine learning. Environmental science & technology, 49(6), 3887-3896.

Chen, G., Li, S., Knibbs, L. D., Hamm, N. A., Cao, W., Li, T., ... & Guo, Y. (2018). A machine learning method to estimate PM2. 5 concentrations across China with remote sensing, meteorological and land use information. Science of the Total Environment, 636, 52-60.

Kanabkaew, T. (2013). Prediction of Hourly Particulate Matter Concentrations in Chiangmai, Thailand Using MODIS Aerosol Optical Depth and Ground-Based Meteorological Data. *EnvironmentAsia*, *6*(2).

Sukitpaneenit, M., & Oanh, N. T. K. (2014). Satellite monitoring for carbon monoxide and particulate matter during forest fire episodes in Northern Thailand. *Environmental monitoring and assessment*, *186*(4), 2495-2504.

Siew, L. Y., Chin, L. Y., & Wee, P. M. J. (2008). ARIMA and integrated ARFIMA models for forecasting air pollution index in Shah Alam, Selangor. *Malaysian Journal of Analytical Sciences*, *12*(1), 257-263.

Ng, K. Y., & Awang, N. (2018). Multiple linear regression and regression with time series error models in forecasting PM 10 concentrations in Peninsular Malaysia. *Environmental monitoring and assessment*, *190*(2), 63.

Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Saudi, A. S. M., Hasnam, C. N. C., ... & Osman, M. R. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. *Water, Air, & Soil Pollution*, *225*(8), 2063.

D. (n.d.). Air Pollutant Index (API). Retrieved August 15, 2020, from https://www.doe.gov.my/portalv1/en/info-umum/english-air-pollutant-index-api/100