CHAPTER 3

RESEARCH METHODOLOGY

3.1 Data Science Project Life Cycle

The data science project life cycle for this study involves several key stages: include problem definition, data collection, data pre-processing, data modelling, and evaluation. The first one is to ensure that the problem is clearly identified and understood. In this study, the main research issue is analysing the effect of air pollution on the population's overall health in Indonesian large cities. This entails determining the nature of pollutants of interest, the disease endpoints to investigate and regions of interest. It assists in setting specific goals of the research and in defining the parameters for collection and analysis of data.

Following that, the process of data collection takes place once the problem to be solved has been identified. This entails the collection of data from the government website, related website, and hospital records. All the sources of data pose important information essential in explaining the link between air pollution and health consequences. It is important here to emphasise that data should be collected in a standardized and accurate manner in order to provide efficient analysis.

The next step in the process is data pre-processing, which aims to clean the data before analysis and transform it into a suitable form. This includes managing cases where the data is missing, eradicating unnecessary data, and standardizing the data which might be in different formats. Data pre-processing serves to make sure the data is clean, complete, and free of inconsistencies and errors before feeding it to the model. This is important in order to eliminate any mistakes or prejudice in the next stages of analysis.

The process of data modelling can be described as the use of mathematical and statistical methods of identifying the patterns of given data. This covers the use of descriptive models for creating insights into existing patterns and a predictive model to make future forecasts. In the modelling phase, features are recognized that can be useful in the pattern analysis and, to a certain extent, causality between air pollution and health effects. The last process is the evaluation, where the models' results are analysed in terms of accuracy and relevancy. It entails testing the models with test data and looking at the results obtained from the models and the actual results. Evaluation brings out the credibility of the models, and the conclusion that is arrived at is valid and relevant.

3.2 Data Source and Collection Methods

3.2.1 Air Quality Data

Gathering data on air quality is an essential prerequisite in order to determine the rates of pollution in the cities and possible effects on human health. Therefore, data from this study will be obtained from different authentic sources to enhance the coverage of the topic and validity of the results. One of the primary sources is IQ Air that provides a real time air quality index through the network of its stations. IQAir highlights the index of main pollutants and PM2. 5, PM10, NO2, SO2, CO, and VOC which are the key sources of air pollution. More specifically, the collected data for the analysis of Jakarta and Surabaya include the following sources.

1. Jakarta Air Quality Data: The real-time data from this source will enable the assessment of the pollution level in Jakarta for instance, one of the Indonesian provinces with high population density. The source link is https://www.iqair.com/air-quality-map/indonesia/jakarta
2. Surabaya Air Quality Data: This data will give information regarding the air quality in Surabaya, another large populated area that has many industries. The source link is https://www.iqair.com/air-quality-map/indonesia/east-java/surabaya

Moreover, the *Badan Pusat Statistik* (BPS) gives long-term air quality data and regional data which helps to understand more details about changes and fluctuations in environmental pollution in Indonesia. For that BPS data plays a crucial role here and the patterns can be easily determined regarding CO level and regional characteristics. The source link is https://www.bps.go.id/

3.2.2 Health Data

Health information is crucial for determining the interaction between air pollution and the population's health, with special reference to respiratory and cardiovascular diseases. Data collection in the study will primarily draw data from the Indonesian Central Statistics Agency (BPS) and the Ministry of Health's which the *"Satu Sehat"* program. BPS offers detailed health statistics from the nationwide surveys and papers; it involves data of diseases, admissions, and fatalities. This data will assist in establishing a relation between the level of pollution and the certain health impacts. The source link is https://www.bps.go.id/

Satu Sehat is another information system which is provided by the Ministry of Health of Indonesia and it consists of various figures of health over different indicators. Patient information database including the disease history and treatment details is provided by this platform, which is crucial for the examination of the connection between pollution and health on the individual level. The source link is https://satusehat.kemkes.go.id/

Furthermore, to secure additional and more comprehensive health information from hospitals in big cities where the study partners with hospitals to gain these records. Reports from specific hospitals such as admissions of the patients admitted, and the diagnoses of the treatments received which may relate to respiratory or cardiovascular illnesses will hence give the exact correlation of the effects of pollution and health. This localized data will improve the test and the specificity of the health impact assessment. Therefore, collecting the national and local health data, this study will aim to construct the understanding of the impact of air pollution on public health with reference to the urban settings in Indonesia.

3.3 Data Pre-Processing

Data pre-processing is an important and preliminary step in the process of preparing the dataset to make sure the data is clean and appropriate for further analysis. This process entails cleaning, transformation and integrating data by using both Microsoft Power BI and python with the advantage of the two platforms.

3.3.1 Data Cleaning

The first step in data pre-processing is data cleansing to remove noise from the data. Transferring the data into the Power BI, the sources of the air quality and health data include CSV, Excel, and online databases. In Power BI, features for data transformation will be used to filter out the existing duplicated records. To deal with missing values, this study will apply interpolation methods or in cases where there is a lot of missing data in a specific row, discard such a row. At the same time, libraries of Python, namely Pandas and Numpy, will also be used for reading and preparing data. The descriptions of some functions including fillna will be able to work on missing values through the imputation of different types while the drop_duplicates will get rid of any entries that may be duplicates. Anomalies in the data will be identified with the help of statistical tools and represented with the help of Matplotlibs and Seaborn.

3.3.2 Data Transformation

Data transformation entails making the data suitable for analysis by eliminating or reducing differences in format. Hence, in MS-Power BI, normalization of data will be done through the transformations where data should be brought to a similar scale. Additional columns produced through calculations will include the trend analysis of specific air quality indexes or health related calculated criteria, e.g., moving average. All date and time fields will be properly written, and such values as day, month, and year will be derived. At the same time, using the Python's Scikit-learn library, it will be possible to normalize the data with the help of such tools as MinMaxScaler or StandardScaler. The extracted date and time data will be formatted consistently from different sources by pandas datetime functions.

3.3.3 Data Integration

Data integration is the last step in the pre-processing phase where air quality and health data are combined into a single database. Microsoft Power BI will be used to combine the datasets as per the matching key like date and location, set up the relationship between the different tables, and develop a complete data model which holds all applicable tables and relationships. On the other hand, using the Python library known as Pandas, the merge function will connect air quality and health data based on specific keys and form the data on the right format for analysis. The dataset that will be merged will be checked on its compatibility on integration so that inconsistencies will be undone before integration is made.

To ensure that the integrated dataset is credible and of high quality, validation checks will be conducted. In the MS Power BI, there are tools that will help to validate the data set for errors such as missing values and outliers. Final examinations will involve line charts, histograms, and scatter plots to determine any other problems. Descriptive statistics will be performed in other to ensure Theological School data distribution is as expected. Cohort-similarly, Python's Pandas will include validation checks, while Matplotlib and Seaborn will aid the performance of visual inspection. The calculation of the summary statistics with the help of Pandas will also help in checking whether the data distributions match expectations.

3.4 Analytical Methods

The analytical techniques used in this research are intended to examine the link between air pollution and health concerns in urban regions of Indonesia comprehensively. This section involves using techniques such as descriptive analysis, predictive analysis, statistical analysis, and machine learning to understand the data. The analytical methods would allow the findings to be valid, consistent, and useful in informing policy and intervention efforts aimed at addressing the negative impacts of air pollution.

3.4.1 Descriptive Analysis

Descriptive Analysis means using basic statistical methods to describe the data, in other words, using tables and charts showing the main data properties. Microsoft Power BI utilised to analyse data and present it in forms of charts, graphs and dashboards that display changes over time, between/among categories, and between different variables. Measurements, including arithmetic mean, mode, median, standard deviation and range will be calculated in order to get the measures of central tendency and dispersion of the data collected. There would be heat maps and geographical plots used in the analysis of spatial occurrences of the air pollution and the health implications of the different cities. At the same time, the graphical means offered by

Python's native libraries Matplotlib and Seaborn shall be used to build histograms, box-and-whisker and scatter plots which will help to reveal distributions of variables and their relationships, respectively. Another type of graphic to be produced with the help of Matplotlib will be a time line where trends and seasonal variations in the levels of air quality as well as health consequences of pollution will be depicted.

3.4.2 Predictive Analysis

In Predictive Analysis, future state of air pollution and the likely health risks, are estimated with the help of past data. This step remains crucial in public health practice and policy in order to prevent different illnesses and diseases. In Microsoft Power BI, built-in forecasting algorithms containing exponential smoothing will be applied to anticipate future air quality levels and observe trends and seasonal variations. In the high-level programming language, namely, Python, powerful time series analysis models like the ARIMA (AutoRegressive Integrated Moving Average) and more specifically, the Prophet model will be used to forecast future emissions of air pollution along with their effect on human health. Scenario analysis will also be carried out in Python as a way of illustrating the possible future conditions as influenced by pollution and other factors.

3.4.3 Statistical Analysis

Statistical analysis covers the use of statistics procedures in Data Analysis to get the association between the variables and the hypotheses testing. This step is crucial in establishing the generality of the findings and proving the statistical tests used. The strength and direction of the relationship between the air quality parameters and the health consequences will be determined via correlation coefficients, in Microsoft Power BI. To assess the effects of air pollution on public health simple & multiple regression analysis shall be applied with the help of analytical features in power BI. In Python, SciPy statistical packages will be applied to conduct hypothesis (for example, t-tests, chi-square tests) to investigate the correlation of variables. In order to compare the effects of different pollutants with health outcomes, multiple linear regression and logistic regression models will be applied with the introduction of confounding factors with the help of Python's two libraries namely statsmodels and scikit-learn.

3.4.4 Machine Learning

Based on the collected datasets, Machine Learning methods will be used for constructing predictive models, which will help predict future tendencies in air pollution and potential adverse health impacts. These models are capable of a nonlinear relationship between variables, such as the existence of an interaction effect. Different classification and regression algorithms will be studied in detail using Python such as decision tree, random forests, support vector machines (SVM) and neural networks, mainly from the Scikit-learn package.