CHAPTER 3
RESEARCH METHODOLOGY

The research is divided into five steps: literature study, data acquisition and preprocessing, data generation, data processing, and analysis.

1. Literature Study

   The research is initialized by collecting journals as a basis to support this research. While the topic resembling the author's research is scarce and far between, there are numerous journals demonstrating the usage prediction algorithms to applicable departments. The author examined and carefully compared available prediction algorithms to find the most ideal to use in this context. After deciding the algorithms and methods to use, the author compiled 10 journals as the initial foundation of this research.

2. Data Acquisition and Preprocessing

   Two datasets are used in this process, the real dataset, and synthetic dataset. This differentiation between synthesized and real is deliberate in order to compare them. Both dataset is used to train prediction algorithm models which in this case are Random Forest regression, and Support Vector Regression.

3. Data Generation

   This step involves using SDV to generate data based the original dataset. The dataset generated by this method depends on the number of sequence parameter within the code. Each sequence generates data equal to the original dataset's amount.

4. Data Processing

   This step is further divided into three steps: data split, model training, and data prediction. Data split is used to allocate some data for testing purposes, while the rest are used to train the algorithm. Both Support Vector Machine and Random Forest Regression need a data model to fit the training data. Aforementioned training data is used for this purpose.

5. Analysis

The prediction between two algorithms is calculated using error evaluation. In this research, the author used Mean Square Error and Mean Absolute Percentage Error to find out which algorithm is more effective than the other, and also to find out whether utilizing synthetic data generation is a viable method to gain higher accuracy or not. The reason for using both is due to the varying scale nature of my dataset. Some parameters in the dataset are greater in number than others. Both evaluation methods handle data scale differently, so they are both used to give varying evaluations from different perspectives.