**3. Materials and Methods**

This chapter examines the materials utilized in this research, with a primary focus on the curation of the dataset employed for training and testing the machine learning model. Additionally, it outlines the method adopted in this study. This chapter begins by outlining the research pipeline, which details the process through which the results of this study will be obtained. The chapter will then discuss how the dataset is curated and the preprocessing steps undertaken before its use in model training. Subsequently, it will explore the oversampling techniques employed, the various models considered, and the process used to select the final model.

**3.1 Dataset**

The dataset used to train the model is derived from the Burden of Disease Epidemiology, Equity and Cost-Effectiveness Programme Health Intervention League Table (BODE3 HILT). Further elaboration about this dataset can be found in Chapter 2, Subsection 2.3. Since the dataset provides a list of interventions along with their effects on Quality Adjusted Life Years (QALYs) and Health System Costs (HSCs), the objective is to leverage this data to train a machine learning model capable of identifying patterns between health interventions and their corresponding health outcomes. However, the BODE3 HILT in its raw form requires further curation before it can be utilised for model development, and it must be transformed into a structured format suitable for machine learning. The interventions need to be parameterised into a set of input features to facilitate model training. These input features were selected because most interventions, along with their corresponding research studies, provide the necessary data to parameterise these interventions within the model. The input features are shown in **Figure 3.1**:

|  |  |
| --- | --- |
| Input Features | Explanation |
| tax\_increase | **Percentage of Tax Increase**  Certain interventions within the BODE3 HILT dataset involve the implementation of taxation on tobacco products. An increase in tobacco taxation leads to a corresponding rise in retail prices, which is expected to deter consumption and reduce smoking prevalence. A decline in smoking behaviour is associated with improvements in quality of life, a reduced risk of smoking-related illnesses—such as lung disease and other chronic conditions—and, consequently, an increase in both life expectancy and QALYs. Furthermore, as smoking prevalence declines, the incidence of smoking-related diseases is expected to decrease, thereby alleviating pressure on the healthcare system and leading to a reduction in HSCs. |
| outlet\_reduction | **Percentage of Outlet Reduction**  Other interventions involve decreasing the number of outlets that can legally retail tobacco products. By limiting the number of retail outlets, the intent is to inconvenience smokers, which is expected to deter consumption, therefore improving QALYs and decreasing HSCs. |
| dec\_smoking\_prevalence | **Decrease in Smoking Prevalence** |
| dec\_tobacco\_supply |  |
| dec\_smoking\_uptake |  |
| age |  |
| gender |  |
| ethnicity |  |

**3.2 Pipeline**

The pipeline starts off with the curation and processing of the dataset, coloured blue in the pipeline image shown in **Figure X**. A total of two datasets will be generated. The first dataset comprises the parameterization of various interventions outlined in the BODE3 Health Intervention League Table (HILT) and their respective effects on QALYs and HSCs. This dataset will be used to train, optimise, and test the built models. The dataset is called “Full Dataset” in the pipeline image. The second dataset is called “Vaping Intervention Dataset”. This dataset encapsulates the parameterisation of the intervention involving the legalisation of vaping, mapped onto the input features outlined in **Figure 3.1** above.

The “Full Dataset” will then be split into the training set and testing set. Given the limited number of samples in the dataset, the training set will not be further divided into separate training and validation sets. Instead, five-fold cross-validation will be employed to optimise the model's hyperparameters. Oversampling is also applied to expand the training set. To identify the optimal combination of model and oversampling method, each model-oversampling pair is trained and tested systematically. To determine the best models, or combination of models to use to predict the results,

A diagram of a program

AI-generated content may be incorrect.