**COMP4550 HONOURS THESIS DRAFT**

**THESIS OUTLINE**

1. Introduction
   1. Smoking in Health Economics
   2. Legalising Domestic Sale of Vape
   3. Machine Learning
   4. Research Question
   5. Thesis Organisation
2. Literature Review
   1. Machine Learning Methods in Health Economics
   2. Proportional Multistate Life Table Model (PMLTM)
   3. BODE3 Health Intervention League Table
   4. Potential Country-level Health and Cost Impacts of Legalizing Domestic Sale of Vaporized Nicotine Products
   5. Conclusion
3. Materials and Methods
   1. Dataset
      1. Interventions
      2. Modelling Tax Increase
      3. Modelling Outlet Reduction
      4. Modelling Smoking Prevalence
      5. Modelling Decrease in Tobacco Supply
      6. Modelling Smoking Uptake
      7. Full Dataset Overview
   2. Pipeline
   3. Preprocessing
      1. Feature Mapping
      2. Imputing Missing Data
   4. Oversampling
      1. Simple Duplicate
      2. K-Nearest Neighbours
      3. Rebalanced K-Nearest Neighbours
   5. Model Building
      1. Linear Regression
      2. Random Forest without Bootstrapping
      3. Random Forest with Bootstrapping
      4. Extreme Gradient Boosting
      5. AdaBoost
      6. Stacking
      7. Ensemble Learning
   6. Cross Validation
4. Results and Discussion
   1. Model Selection
   2. Selecting Upsampling Method
   3. Effect on Quality Adjusted Life Years (QALYs)
   4. Effect on Health System Costs (HSCs)
   5. Effect on Māori/non-Māori inequities
5. Impact, Strengths and Limitations, and Future Work
   1. Impact
   2. Strengths and Limitations
   3. Future Work
   4. Summary
6. Conclusion

**1. Introduction**

Legalising the domestic sale of vaporised nicotine products, more commonly referred to as vape, is a highly controversial topic in recent years. Proponents argue that vape is a useful cessation tool and legalising it would facilitate a smoker’s journey to quit smoking. On the other hand, detractors argue that vape is equally harmful and may be a gateway for people to eventually pick up smoking. Existing literature does not give a definitive answer as to whether legalising vape has a net positive or negative effect. This study aims to use machine learning methodologies to assess the impact of legalising the domestic sale of vape on smoking-related outcomes. These smoking-related outcomes include (1) Quality Adjusted Life Years (QALYs), (2) Health System Costs (HSCs), and (3) Māori/non-Māori health inequities. Machine learning offers several advantages. Numerous research papers have studied the effect of health interventions on these smoking-related outcomes and machine learning can be used to leverage on these research findings to identify patterns within the data and make predictions based on those patterns. By leveraging past research findings, machine learning models improve predictive accuracy and generalisability, allowing policymakers to assess potential intervention impacts before implementation. Additionally, machine learning enables continuous learning, where new research findings can be easily added and used as data for building the machine learning model. This research adopts New Zealand’s 2025 Smoke-Free Goal as a case study to evaluate these impacts. This chapter provides an overview of the health economics of smoking, and the ambiguity of the effects of legalising the domestic sale of vape. Following this, this chapter introduces the field of machine learning, highlighting its potential in health economics research. Two central research questions are articulated to frame the study, and this chapter concludes with an outline of the remaining structure of this thesis.

* 1. **Health Economics of Smoking**

Health economics is a branch of economics that focuses on the effects of healthcare policies and interventions on a population’s health. Two important metrics of health economics that measure the impact of interventions on population health are Quality Adjusted Life Years (QALYs) and Health System Costs (HSCs).

QALYs measures the increase in the quality and quantity of years lived because of a healthcare intervention. The formula of QALY is:

is a value between 0 and 1 where 0 refers to death and 1 represents perfect health

is the number of years with the health state of

For instance, a QALY value of 1 represents an additional year lived in perfect health because of an intervention.

HSCs measures the costs incurred on the healthcare system, caused by a healthcare intervention, in monetary value, ie. in New Zealand Dollars (NZ$). A negative HSC refers to cost savings in a healthcare system. This is possible when the healthcare intervention improves population health and alleviates the pressure on the healthcare system. This results in the healthcare system incurring lower costs than when in the absence of an intervention. A positive HSC correlates to additional costs being incurred by the healthcare system, with the presence of the healthcare intervention. This happens when the administration costs of the intervention or the decrease in government revenue because of the intervention outweighs the cost savings from the improvement of population health.

To investigate population health, health economics examines behaviours that influence health outcomes, such as smoking, diabetes, substance use, and obesity, among others. This research concentrates on the domain of smoking. Smoking is a leading cause of mortality globally and contributes significantly to a range of severe health conditions, including cardiovascular diseases, lung cancer, and fertility complications, among other adverse outcomes. In 2021, it was estimated that tobacco smoking was the second highest risk factor driving the most deaths and disability combined in New Zealand. Tobacco smoking was responsible for approximately 3,600 deaths, which is about 21.6% of total deaths in New Zealand. In line with New Zealand’s goals of eradicating smoking, various tobacco-related interventions have been modelled, and the effects of these interventions have been investigated. These interventions include: (1) Tobacco Tax Increase, (2) Tobacco Free Generation, (3) Outlet Reduction, (4) Sinking Lid, (5) Combination of multiple interventions, amongst others. This research will focus on the effects of legalising domestic sale of vape.

* 1. **Legalising Domestic Sale of Vape**

Electronic cigarettes or vape were invented in 2003 and first introduced into the New Zealand market in 2006. Vaping is sometimes positioned as a smoking cessation tool that may reduce reliance on tobacco products and potentially facilitate smoking cessation altogether. **Figure 1.1** shows that across the younger age groups (15-17, 18-24, 25-34, and 35-44) from 2011 to 2024, the percentage of daily smokers decreased while the number of daily vapers increased. At first glance, the introduction of vaping appears to correlate with a reduction in the percentage of daily smokers. However, it is important to acknowledge that correlation does not imply causation, and the observed decline in smoking rates across these age groups may be attributable to other factors, such as the implementation of other public health interventions. Additionally, referring to **Figure 1.2**, the introduction of vape has a negligible effect on the percentage of daily smokers across the older age groups (45-54, 55-64, 65-74, and 75+). This indicates that the vaping is not necessarily effective as a smoking cessation tool across all age groups. Finally, vaping is also a potential gateway to smoking, where non-smokers may initiate vaping under the perception that it is less harmful, subsequently progressing to the use of tobacco products. For these reasons, the net impact of legalizing the domestic sale of vape on population health remains inconclusive. This research aims to evaluate whether its effects are ultimately positive or negative on public health outcomes.

Figure 1.1: Percentage of Daily Smokers and Daily Vapers in the New Zealand Population Across Younger Age Groups (15-17, 18-24, 25-34, and 35-44)

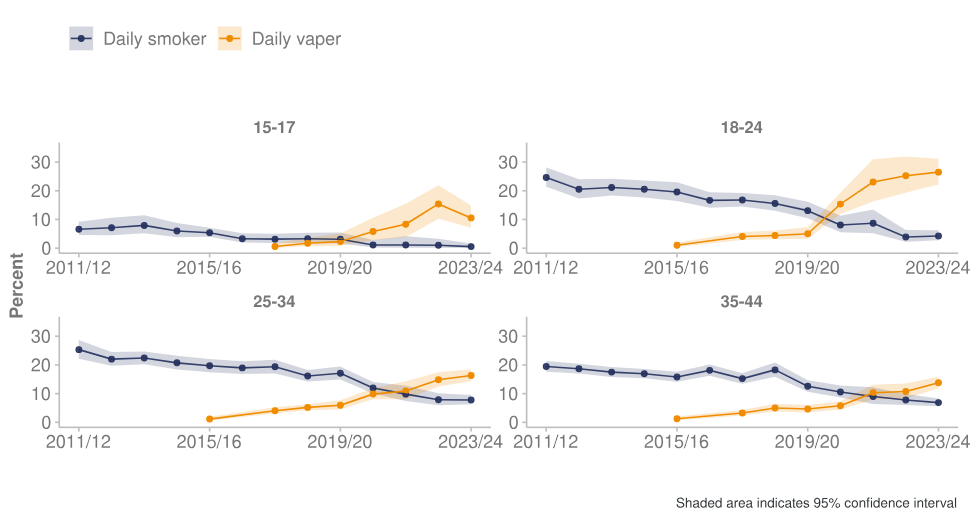
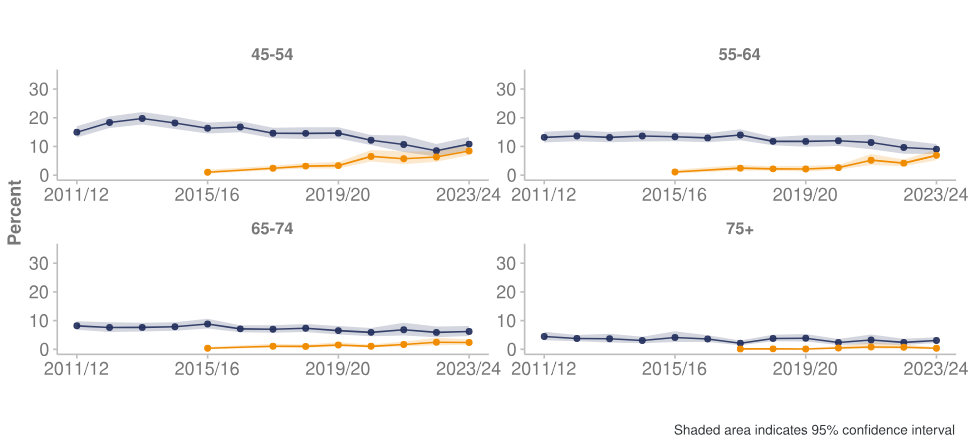


Figure 1.2: Percentage of Daily Smokers and Daily Vapers in the New Zealand Population Across Older Age Groups (45-54, 55-64, 65-74, and 75+)



* 1. **Machine Learning**

Machine learning spans across various domains due to its ability to learn patterns from data and make intelligent decisions or predictions. In the field of health economics, machine learning has been primarily utilised for predicting future events, such as disease prevalence or patient outcomes. Increasingly, machine learning is also being applied to predict health economic outcomes, including healthcare resource utilization and associated costs. This study is one such example and leverages existing data on various healthcare interventions and their impacts on Quality-Adjusted Life Years (QALYs) and Health System Costs (HSCs) to develop and optimize a predictive model. The model is designed to forecast the effects of other interventions to evaluate the implications of legalising the domestic sale of vape. This research uses methods such as random forest with bootstrapping, random forest without bootstrapping, extreme gradient boosting (XGBoost), AdaBoost, Stacking and a combination of these methods.

* 1. **Research Aim**

This research aims to leverage the findings of previous research papers to build a ML model that predicts the effects of legalising domestic sale of vaporised nicotine products, or vape for short, on (1) Quality Adjusted Life Years (QALYs), (2) Health System Costs (HSCs), and (3) Māori/non-Māori health inequities.

* 1. **Thesis Organisation**

Following this chapter, Chapter 2 will include a Literature Review of the topic. Details on the materials and methods, which include the building the dataset, data preprocessing, and building the model, are included in Chapter 3. Chapter 4 will present the results of the study in relation to the research aims outlined earlier and provide a detailed discussion of the findings. Chapter 5 will discuss the research impact, limitations and future work. Finally, Chapter 6 will provide the concluding remarks for this study.

**2. Literature Review**

Research papers that model the effect of health interventions on QALYs and HSCs use a Proportional Multistate Life Table Model. The findings from these studies are integrated into the Burden of Disease Epidemiology, Equity and Cost-Effectiveness (BODE3) Health Intervention League Table. In the BODE3 Table, there are a total of 9 research papers that specifically study the effect of smoking-related health interventions on health outcomes in New Zealand.

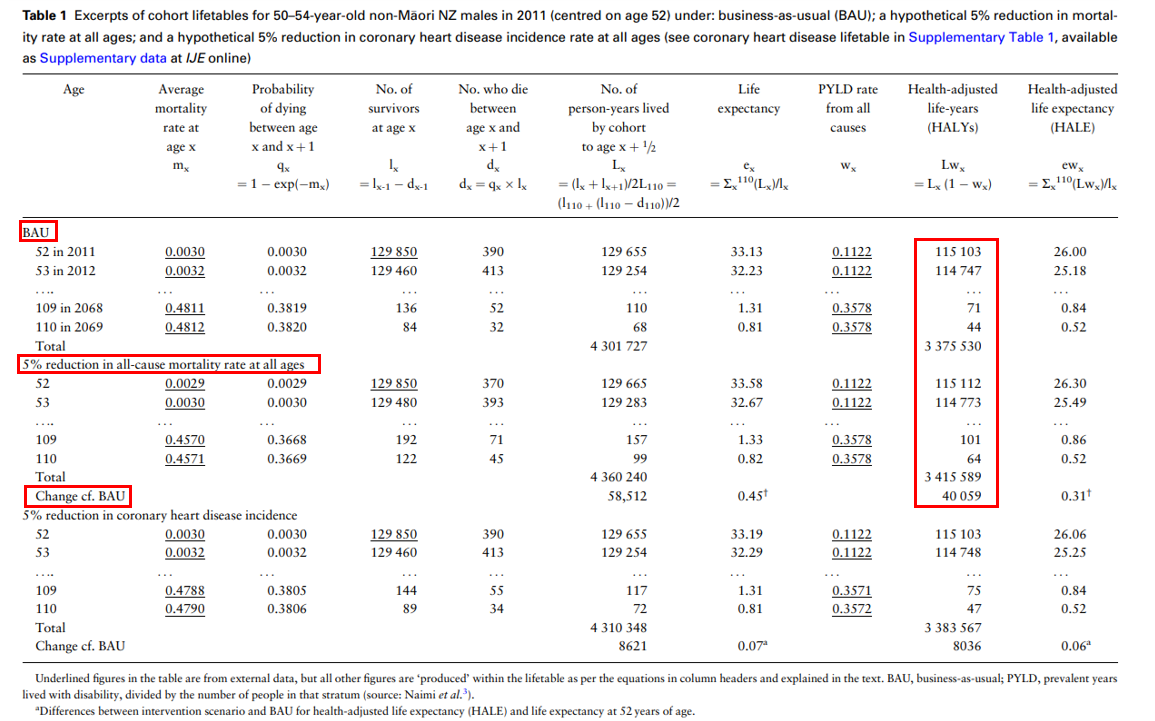
**2.1 Machine Learning Methods in Health Economics**

Recent advancements in technology have ushered in a widespread use of machine learning methods in health economics. One reason is the introduction of wearable technology such as fitness trackers, that capture personal data at any given moment. Such technologies, in addition to more traditional sources such as electronic medical records (EMRs) create large datasets that can be used for machine learning. In health economics, machine learning is primarily used to predict the probability of a clinical event, disease incidence, or treatment outcomes. For example, researchers can estimate the probability of disease incidence by developing predictive models using large datasets containing personal information. In these models, the input variables represent individuals' personal characteristics, while the output variable indicates the presence or absence of the disease. Machine learning was less used to predict economic outcomes such as HSCs. Tree-based methods such as random forests and boosting were the most used machine learning methods due to their ease of application and intuitiveness. However, this research deviates from the conventional machine learning methodologies, further explained in **Chapter 3**. Instead of using a large dataset containing personal data, this research uses the results of existing literature to curate a dataset, which will be used to build a machine learning model. The results are retrieved from the BODE3 Health Intervention League Table, which will be discussed further in this chapter.

**2.2 Proportional Multistate Life Table Model**

Proportional Multistate Lifetable (PMSLT) modelling quantifies intervention impacts in terms of QALYs and HSCs. A key concept in this modelling method is the life table, which summarises the mortality rate, life expectancy, and QALY (denoted as Health-adjusted life-years in **Figure 2.1**) and other health outcomes of a demographic at a specific age. Referring to the highlighted areas in **Figure 2.1**, we can compare the QALYs between Business as Usual (BAU) and the QALYs after an intervention. BAU refers to the scenario where no policy changes occur. The increase in QALYs indicates the effectiveness of an intervention.

Figure 2.1: Example Life Table in Proportional Multistate Lifetable Modelling



In a PMSLT, QALYs are calculated by simulating the transitions of a population through different health states over time. Each transition has a probability calculated using past data, and each health state has a corresponding utility value that reflects the quality of life. For each state, the person’s years lived in that state are multiplied by the state's utility value (ranging from 0 for death to 1 for perfect health), producing QALYs for that state. The total QALYs for the population are obtained by summing these values across all states and time periods. Similarly, HSCs are calculated by assigning a per-person healthcare cost to each health state and multiplying it by the number of individuals in that state at each time step. These costs, which reflect medical expenses such as hospitalizations and treatments, are derived from historical data. The total HSC is obtained by summing the costs across all health states and time periods.

The key advantage of PMSLT is that it allows the inclusion of multiple diseases without state explosion. For instance, in a standard Markov model, a model with five diseases would have 32 possible combination of disease states. Given the number of diseases in the real world, using a Markov model would result in an exponential number of states. On the other hand, the PMSLT allows the population to exist simultaneously in many states, significantly reducing the number of states needed for the model. This significantly computational and parameterisation costs. In addition, social group heterogeneity such as sex, age, and ethnicity can be incorporated in PMSLT models. This means that the effect of interventions on various demographics can be analysed. For these reasons, PMSLT models remain the primary model used to analyse the effect of smoking interventions.

However, PMSLT models can be data-intensive, requiring detailed information on transition probabilities and utility weights. These models utilise vitals and registry data, census data, GBD epidemiological data, survey data, data forecasts for trends in BAU, among others. Data also needs to be processed to ensure coherence. Grasping the intricacies of the PMSLT model necessitates a deep understanding of health economics, making it challenging to comprehend without prior knowledge in the field. This research seeks to build upon the findings of studies utilizing the PMSLT by developing a simplified machine learning model. In this process, the dataset—composed of results from multiple research papers—will be considerably smaller while still preserving the outcomes derived from the PMSLT model.

**2.3 BODE3 Health Intervention League Table**

The BODE3 [Health Intervention League Table](https://league-table.shinyapps.io/bode3/) (HILT) comprises the health gain and cost-effectiveness of health interventions evaluated by the BODE3 program. Each data sample is the result of a research paper on the topic. This enables researchers to get an informed 'first impression' on the efficacy of health interventions. **Figure 2.2** describes the columns in the HILT.

Figure 2.2: Health Intervention League Table for Tobacco Domain

|  |  |
| --- | --- |
| **Column** | **Description** |
| Country | Indicates the country the research was conducted in. The research is centred on New Zealand's 2025 Smoke-Free Goal. Consequently, only records with the country designation "NZ" will be included in the dataset for model building. |
| Base Year | Indicates the start year from which the effects of an intervention are modelled. Research papers that use New Zealand’s 2025 Smoke-Free Goal as a case study unanimously have 2011 as its base year. |
| Domain | Indicates the health domain of the study, such as alcohol, cancer, diabetes, among others. This research will focus on records with the *Tobacco* domain. |
| Intervention | Indicates the name of the intervention associated with the study. Records that indicate *Het* under the “Heterogeneity” column (explained below) also specifies the age, gender, and ethnicity. |
| Heterogeneity | Indicates if the record is disaggregated by age, gender, or ethnicity. *Main* indicates that the QALY and HSC of the record is aggregated across all age groups, gender, and ethnicity. *Het* indicates the QALY and HSC value represents a specific age group, gender, or ethnicity. |
| Comparator | Indicates the scenario in which QALYs and HSCs are compared against. For research papers that use New Zealand’s 2025 Smoke-Free Goal as a case study, the comparator is *Business-as-usual.* |
| Discount Rate | The discount rate is used to adjust the value of future costs and health benefits to their present value. All records in the *Tobacco* domain use a discount rate of 3%, which means that future costs and health benefits are reduced by 3% per year. This reflects the principle that people generally prefer to receive benefits sooner rather than later and incur costs later rather than sooner. |
| QALYs/DALYs | Indicates the change in QALYs due to the intervention. QALYs per 1,000 population are used in the dataset for model building to account for the specific demographic's population size, ensuring the QALY value is sufficiently large for effective machine learning application. |
| Health System Costs | Indicates the change in HSCs due to the intervention. A negative HSC refers to cost savings in a healthcare system. A positive HSC refers to additional costs being incurred by the healthcare system, with the presence of the healthcare intervention. |
| Intervention Costs | Indicates the cost in New Zealand Dollars (NZ$) of implementing the intervention |
| Time Horizon | Indicates the time frame in which the effect of the intervention is measured. In this study, the effect of all the interventions mentioned is 14 years (from 2011 to 2025). In the dataset used, the QALY and HSC values represent the lifetime effect of the intervention. |
| Duration and Frequency | Indicates the duration of the intervention implementation. |
| Perspective | Indicates the viewpoint from which costs and benefits of health interventions are assessed. All data records are viewed from the perspective of the *health system*. |
| Link | Link to the research paper |
| Publication Year | Year in which the research paper is published |

The HILT offers a concise summary of results from existing literature. However, it has certain limitations that restrict its overall utility. Some of these limitations can be addressed through data processing techniques, while others represent more significant challenges that require specific methods to overcome. These methods will be explored in detail in **Chapter 3**.

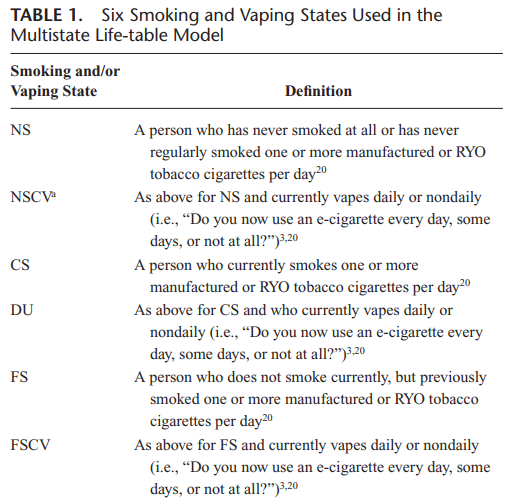
Firstly, the HILT has 243 entries, which without further processing, is too small a dataset for machine learning techniques to be used. Amongst the 243 entries, not all can be used for model building. Some entries study the effect of health interventions in the Australian population. However, this research focuses on the effects on the New Zealand population. Comparing the effect of interventions using different demographics will result in an inaccurate comparison. In addition, this research aims to explore the health inequities between the Māori and non-Māori people. Even if entries that use Australia as a case study granularise their findings by ethnicity, we cannot assume that the indigenous population of Australia and New Zealand have the same smoking and vaping behaviour.

Secondly, research papers often lack consistency in the granularity with which they analyse the effects of an intervention on specific subgroups. For example, some studies disaggregate data by gender and ethnicity but omit age, while others provide a more comprehensive breakdown that includes all three variables. Data imputation is necessary for entries that omit certain variables, requiring specific assumptions to fill in the missing data. This process can introduce inaccuracies into the dataset, potentially affecting the reliability of the results.

**2.4 Potential Country-level Health and Cost Impacts of Legalizing Domestic Sale of Vaporized Nicotine Products**

This research is not the first to predict the effect of legalising vape on QALYs and HSCs. Van der Deen et al. used a multistate life-table model of 16 tobacco-related diseases to simulate lifetime QALYs and HSCs. Conventional multistate lifetables that are used to study the health impacts of smoking incorporate transitions from never smoker, former smoker, and current smoker. To introduce the element of vaping, Van der Deen et al.’s research also introduces more states, shown in **Figure 2.3**.

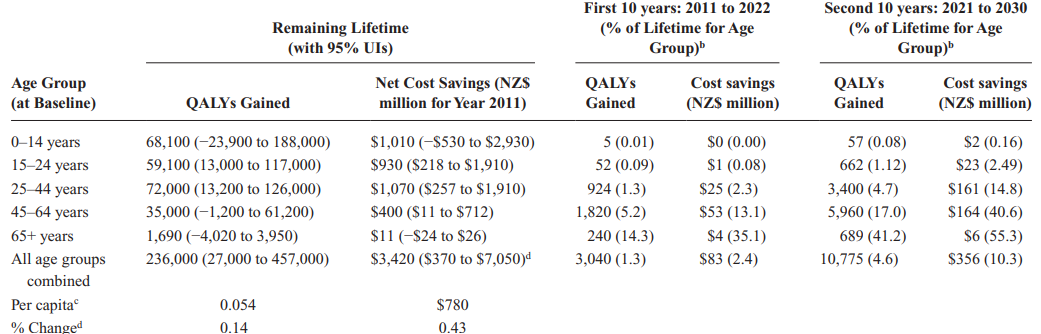
Figure 2.3: Smoking and Vaping States Used in Multistate Life Table Model by van der Deen et al.



The effect of legalizing vaping on future QALYs and HSCs is assessed by analysing shifts in the population distribution across the six health states in **Figure 2.3**, compared to business as usual (BAU). This analysis involves integrating these population shifts with the relative risks associated with vaping and smoking for 16 specific diseases. By doing so, population impact fractions (PIFs) are calculated. PIFs are then used to set intervention disease incidence rates in the 16 parallel disease life tables. These incidence rates are then used to calculate the changes in morbidity and mortality rates of a population because of a disease. The total QALYs, which indicate life-years lived adjusted for morbidity, and HSCs can then be calculated, and aggregated by age.

While the result of this paper is also integrated into the HILT, it is not disaggregated by age, gender, and ethnicity. Van der Deen et al. disaggregated the results of the paper by age, as shown in **Figure 2.4**. To maximise the usefulness of the HILT, research papers should ideally be consistent in disaggregating results by age, gender, and ethnicity. In addition, vapers who vape every day, or only occasionally are all grouped under the same smoking or vaping state. Intuitively, the frequency of vaping should impact future QALYs and HSCs but there is insufficient information to accurately parameterize a model for vaping frequency. Finally, the potential impact of legalising vape with other tobacco control policies was not considered when building the model. The interactions between different policies could lead to even greater reduction in smoking rate, which van der Deen et al. acknowledges to be a potential future work.

Figure 2.4: Result Table by van der Deen et al.



**2.5 Conclusion**

To conclude, while there are existing works that predict the effects of legalising vape, this research deviates from the methodologies of those works, by curating a dataset using the results of relevant papers and using that data to build a machine learning model. The next chapter will discuss how the dataset is curated and how the machine learning model is built and optimised.

**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) across a 14 year time period, 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**:

Figure 3.1: Model Input Features

|  |  |
| --- | --- |
| 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.  All interventions that have a non-zero value for this feature involve some form of annual tax increase. The value in this input feature indicates the percentage increase in tax every year, with the assumption that the increase in tax is constant in the 14-year period from 2011 – 2025, which is the time period for all relevant interventions in the BODE3 HILT. |
| 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**  Health interventions are aimed at decreasing the smoking prevalence of the population. A trickle-down effect of that would be the increase in QALYs and health system cost savings on the government. As such, research papers in the BODE3 HILT all contain the estimated effect on smoking prevalence because of a health intervention. |
| dec\_tobacco\_supply | **Decrease in Tobacco Supply**  Decrease in tobacco supply here refers to a decrease of tobacco imports into New Zealand on a national level, as opposed to the “outlet\_reduction” input parameter which focuses on the number of tobacco retail outlets. A decrease in tobacco supply reduces the smoking rate primarily by making tobacco products less accessible and often more expensive. When supply is restricted, such as by reducing import quotas, or tightening distribution channels, it becomes more difficult for consumers to purchase tobacco. This can lead to higher prices and reduced convenience, both of which discourage smoking initiation and promote cessation. In interventions such as the sinking lid strategy, the supply of tobacco into New Zealand steadily decreases to 0, making it ultimately illegal to purchase any tobacco in New Zealand. |
| smoking\_uptake | **Smoking Uptake**  Some interventions aim to prevent people—especially youth—from starting to smoke by reducing tobacco’s appeal and accessibility. Measures like raising the legal age, banning advertising, and public education discourage smoking initiation. This leads to fewer tobacco-related illnesses, improving QALYs. Over time, it also reduces demand on healthcare services, lowering HSCs. |
| age | **Age**  To facilitate the identification of patterns between age and health outcomes, the samples may be divided into five distinct age groups:   * 0 – 14 years old * 15 – 24 years old * 25 – 44 years old * 45 – 64 years old * 65 years and older |
| gender | **Gender**  The samples also indicate the effect of the intervention on Male and Female. |
| ethnicity | **Ethnicity**  In line with the research aims of finding the effect of legalising vape on Māori and non-Māori health inequities, the samples also indicate if the effect of the intervention on Māori and non-Māori populations. |

In total, there are 10 interventions from the BODE3 HILT that can be parameterised into the input features shown in **Figure 3.1**. **Figure 3.2** below shows the interventions and how they are parameterised into the input features.

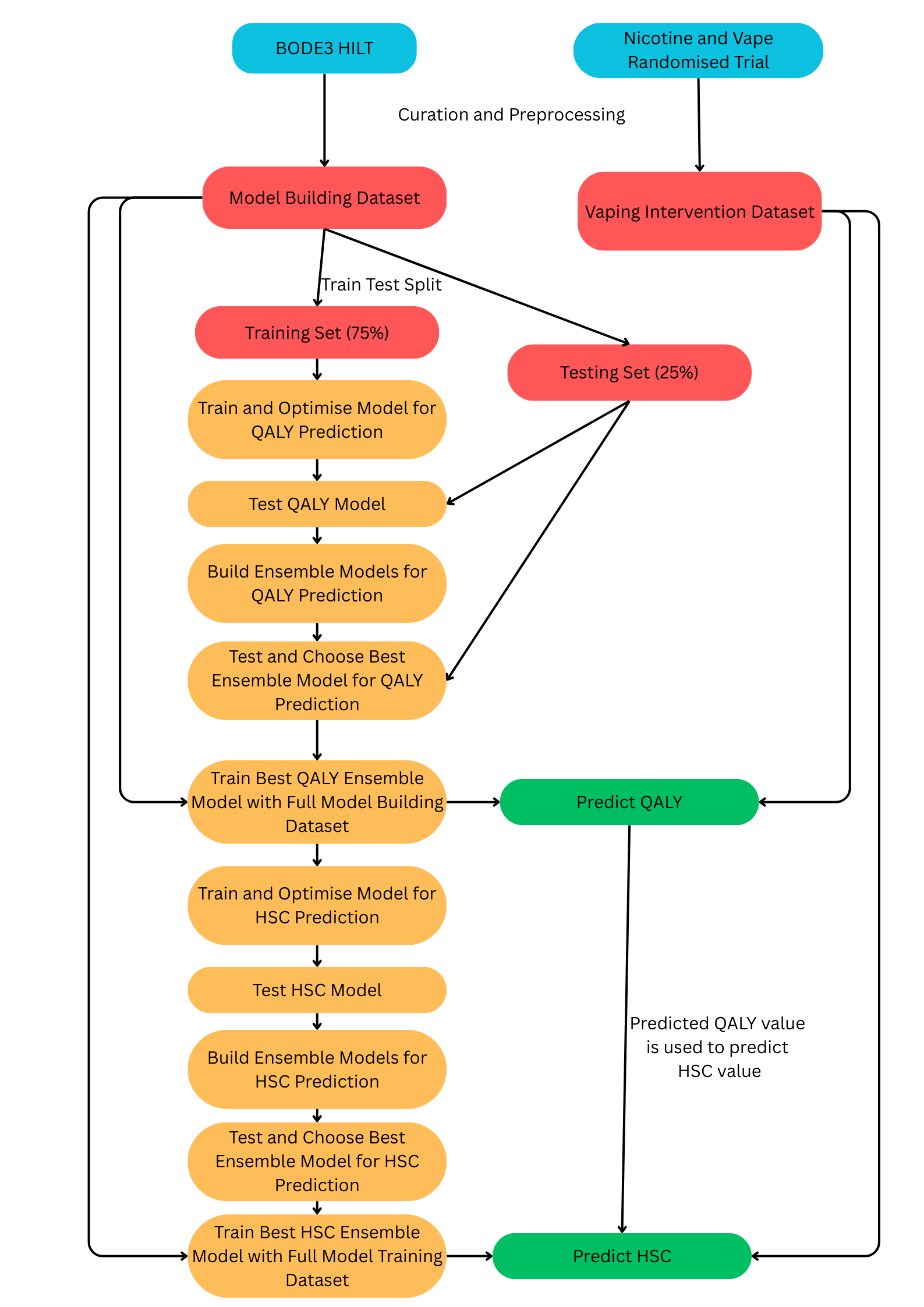
Figure 3.2: Intervention Parameterisation

|  |  |  |  |
| --- | --- | --- | --- |
| **S/N** | **Intervention** | **Explanation** | **Parameterisation** |
| 1 | Ongoing Tobacco Tax Increase | This intervention refers to a 10% annual tobacco tax increase in a 14-year period (from 2011 – 2025). | tax\_increase |
| 2 |  |  |  |
| 3 |  |  |  |
| 4 |  |  |  |
| 5 |  |  |  |
| 6 |  |  |  |
| 7 |  |  |  |
| 8 |  |  |  |
| 9 |  |  |  |
| 10 |  |  |  |

**3.2 Pipeline**

The pipeline starts off with the curation and processing of the dataset, coloured blue in the pipeline image shown in **Figure 3.3**. 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 “Model Building 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 “Model Building 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 optimize the model's hyperparameters.



**Hyperparameters optimized for each model:**

Random Forest without Bootstrapping

* n\_estimators
* max\_depth
* min\_samples\_leaf

Random Forest with Bootstrapping

* n\_estimators
* max\_depth
* min\_samples\_leaf
* max\_samples

XGBoost

* n\_estimators
* max\_depth
* min\_child\_weight
* reg\_lambda
* reg\_alpha

AdaBoost

* n\_estimators
* learning\_rate
* loss

Stacking Ensemble

* See Table x

**Oversampling Techniques Attempted:**

* Simple Duplicate with Gaussian Noise
* Duplicate with K-Nearest Neighbours
* Synthetic Minority Oversampling

\*Each oversampling technique was attempted with n=50, 100, 200, 300, and 500