ML VAPING EFFECT PROJECT WEEK 7 REPORT

Prepared By: Ng Jun Kiat (u7338876)

INTRODUCTION

1. This report describes the steps taken to build the Random Forest model to predict Quality Adjusted Life Years (QALYs) and Health System Cost Savings using the BODE3 dataset.

DATA PREPERATION

2. The following columns were added to the original dataset:

   a. age_group.

      The five age groups from the BODE3 dataset, (1) 0 – 14 years old, (2) 15 – 24 years old, (3) 25 – 44 years old, (4) 45 – 64 years old, (5) and 65 years old and above, were mapped to integers from 0 to 4 respectively. This is to preserve the ordering of the different age groups.

   b. gender_idx.

      Males were mapped to the integer 0, and females to the integer 1.

   c. ethnicity_idx.

      Māori were mapped to the integer 0 and non-Māori to the integer 1.

3. The dataset was duplicated five times to increase the number of samples. When the dataset was not duplicated, the mean squared error (MSE) for both QALYs and Health System Costs were observed to be high, which indicates that the model did not converge.

4. The following columns were specified as the independent variables (*X*):
   - 'tax_increase'
   - 'outlet_reduction'
   - 'dec_smoking_prevalence'
   - 'dec_tobacco_supply'
   - 'dec_smoking_uptake'
   - 'age_group'
   - 'gender_idx'
   - 'ethnicity_idx'

5. The following columns were specified as the dependent variables (*y*):
   - 'qalys_pc'
   - 'hs_costs_pc'

6. 80% of the dataset will be used for training to maximise the size of the training set. 10% will be used for validation, and 10% will be used for testing.

VALIDATION

7. Two model hyperparameters were optimised, (1) the number of decision tree estimators, and (2) the maximum depth of the trees. The optimal value for *n_estimators* was determined to be 50. Using *n_estimators = 50*, the optimal value for *max_depth* was determined to be 15. **Figure 1** and **Figure 2** show the MSE for QALYs and Health System Costs against *n_estimators* and *max_depth* respectively.

Figure 1: MSE for QALYs and Health System Cost Savings against *n_estimators*.

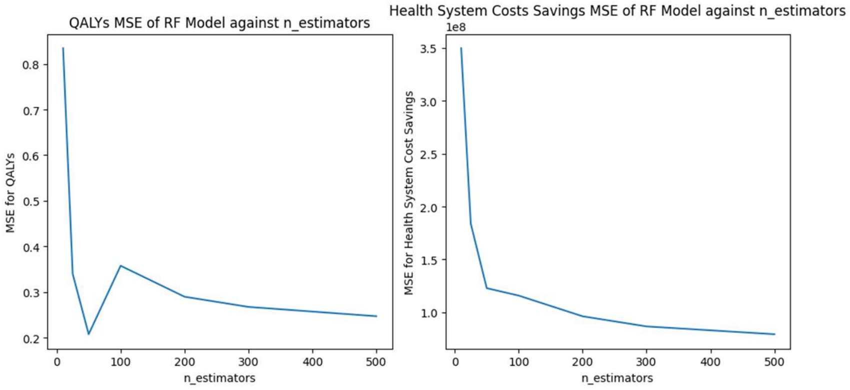
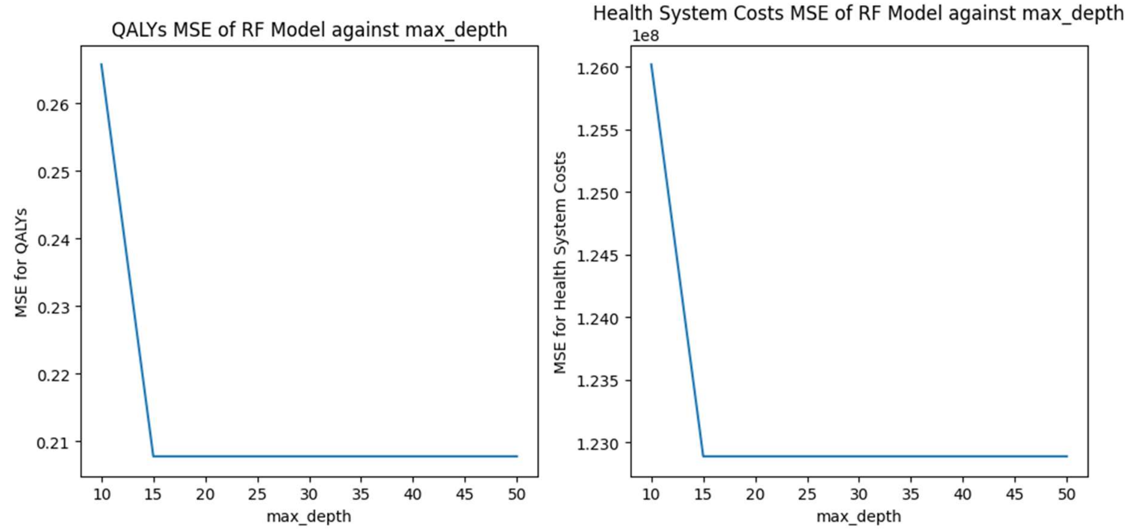


Figure 2: MSE for QALYs and Health System Cost Savings against *max_depth*.



TESTING

8. Using *n_estimators = 50* and *max_depth = 15*, **Figure 3** shows the MSE for QALYs and Health System Cost Savings as an absolute value as and as a percentage of the range of the respective indicators.

Figure 3: MSE for QALYs and Health System Cost Savings using *n_estimators = 50* and *max_depth = 15*

Indicator	MSE	MSE as Percentage of Range/%
QALYs	0.64	0.23%
Health System Cost Savings	279077008.57	4288.07%

9. **Figure 3** shows that while the model performs well when predicting QALYs, its MSE is extremely high when predicting Health System Cost Savings.

FOLLOW-UP ACTIONS

10. Follow-up actions include improving the model's performance for predicting health system cost savings. One option would be to duplicate the dataset more times, or to try a different model altogether. Another improvement would be to model the uncertainty of different independent variables, such as 'tax_increase', 'outlet_reduction', 'dec_smoking_prevalence', 'dec_tobacco_supply', and 'dec_smoking_uptake'.