Using classification models to accurately predict disease risk for individuals

GROUP M1: EVANS OCHIENG, JAMES PRESTWICH, MEHMET CAN AYDIN, PAULINE SCHOUTEN, RUDRA AJAY SOMESHWAR, STERYIOS NICOLAIDES

Table of Contents

1. Introduction

The massive growth in recent years of healthcare data and the advancement in machine learning tools and classifiers means there is now a large demand for accurate disease prediction tools. The ability to predict whether an individual is a risk of a serious illness has the potential to not only save lives but also reduce demand and costs for the healthcare system. Columbia Public Health discusses the value of disease risk prediction across sectors ranging from when to prescribe medications such as aspirin in a clinical setting, to Governmental agencies deciding where and how many defibrillators to install in a housing project [1].

The World Health Organization not only states that cardiovascular diseases are the leading cause of death globally but also that most of these could be prevented by addressing behavioural risk factors [2]. Nicholls et al. estimated the mean direct medical care cost of cardiovascular disease to be $18,953 per patient per year in the United States, and significantly higher if the patient had other diseases such as diabetes and kidney disease [3]. These are all diseases we hope to be able to predict using Behavioural Risk Factor Surveillance System (BRFSS) survey data [4].

Insurance companies need to be able to accurately assess the risks when offering a policy to a customer in order to set an appropriate and competitive premium. Our models would allow insurers to gain further insight into their customers' risks of developing certain diseases by asking them a set of survey questions. Healthcare providers both private and public are under increasing demand globally, for them to cope with

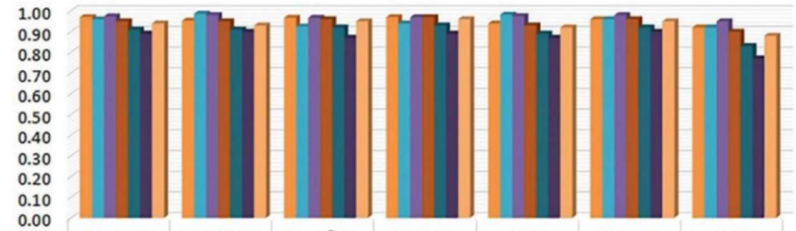
this increasing demand they need to be able to identify which patients are at risks of developing serious illnesses that will require expensive inpatient care. Our models would allow medical practices to target early intervention at high risks patients to reduce the number of patients that will require inpatient care in the long term, reducing both manpower demand and cost.

The aim of this report is to produce a model that will be able to classify an individual's risk of certain diseases based on survey data. This model will allow insurance companies to assess the risk of potential customers in order to adjust insurance premium costs. It will also allow medical practices to identify patients for early intervention. The diseases that will be predicted are:

- Coronary heart disease and myocardial infarction (CHD and MI)
- Kidney Disease
- Skin cancer
- Other cancers
- Diabetes
- Asthma
- Arthritis

1.1 Existing Research

A number of studies have been completed in this field, however the majority of these have focused on predicting heart disease. Soni et al. provided an early overview into how effective different models were at heart disease prediction where they found Decision Trees and Bayesian classification were most effective whereas clustering and Neural Networks did not perform as well. They also found that accuracy improved when genetic algorithms were used to identify the best subsets of attributes for heart disease prediction [5]. In a more recent study, Uddin et al. reviewed 48 studies to compare which algorithms they used to predict disease, focusing on studies that used more than one algorithm [6]. They found that Support Vector Machine was the most commonly used, however, in the studies where it was used Random Forest showed the best comparative accuracy. Ayon, Islam and Hossain compared models directly using the valuated Statlog and Cleveland heart disease dataset and found Deep Neural Network's the most accurate, closely followed by Support Vector Machine [7]. A table showing the accuracy and other performance metrics for the range of models they trialled can be found in Figure 1.1 below.



	Accuracy	Sensitivity	Specificty	Precision	NPV	F1 score	MCC
SVM	0.97	0.95	0.97	0.97	0.94	0.96	0.92
LR	0.96	0.99	0.93	0.94	0.98	0.96	0.92
DNN	0.97	0.98	0.97	0.97	0.97	0.98	0.95
DT	0.95	0.95	0.96	0.97	0.93	0.96	0.90
NB	0.91	0.91	0.92	0.93	0.89	0.92	0.83
RF	0.89	0.90	0.87	0.89	0.87	0.90	0.77
K-NN	0.94	0.93	0.95	0.96	0.92	0.95	0.88

Figure 1.1: Evaluation metrics of the heart disease prediction (ten-fold) system using Statlog dataset [7].

In a comprehensive literature review of the use of health data for disease prediction Hossein et al. compared the advantages and disadvantages of different models and found that there were benefits to each but the biggest limitations to the studies they examined was data quality [8].

Advances in machine learning, in particular, Convolutional Neural Networks (CNNs) and their ability to analyse images and unstructured text has meant that even more sources of data are available for disease prediction which may help counteract the problems Hossein et al. identified. Chen et al. achieved a 94.8% accuracy with a CNN when predicting cerebral infarction using a combination of both structured and unstructured data from hospitals in China [9].

It is clear from the studies discussed above and multitude of others they reference, that machine learning has the potential to accurately predict disease and advance the quality of healthcare provision around the world. However, the studies often disagree on the best algorithm to be used as their performances vary on both the data used and the type of disease being predicted. Our study differs from the studies before as it aims to predict several diseases from the same dataset rather than just heart disease like most before. It also aims to use telephone survey data rather than hospital data which is easier and cheaper to gather.

1.2 Data used in this report

As discussed earlier, this report looks at telephone survey data, specifically the Behavioural Risk Factor Surveillance System (BRFFS) survey data collected in 2020 by the Centers for Disease Control and Prevention (CDC) in the United States of America [4]. The data was gathered using landline and cellular telephone surveys across all 50 states, collecting information on both health risk behaviours and chronic conditions.

The data itself consists of 279 columns and 401958 rows. The column features range from demographic questions such as race and income to health specific questions such as tobacco use and whether the interviewee has ever been diagnosed with diabetes. Due to the large number of features contained in those dataset care was needed to select the most appropriate to be used in this study. All columns are populated with numerical data and there is a codebook [10] which comes with the dataset to explain in detail the questions asked and what the numerical values correspond to. Blank responses are separated from refused to respond which are normally assigned values of 9, 99, or 999, although this varies from column to column.

2. Methodology

2.1 Data Preparation and Cleaning

The BRFSS dataset has some responses might be considered as noise during the data preparation. Due to the data collection method, via telephone survey, some records that have no relation with the project aim were also stored in the dataset as columns. Even though some columns might have a correlation with the project aim, they might have many null values as some questions were not asked or there was not a response from the respondent.

It was also crucial to reduce the complexity of the models by decreasing the number of columns used as input columns. Therefore, the first step was removing the columns that had no relation with the project case. Also, in the dataset there were columns that had to be removed since they stored some calculated values. These columns were generated with the combination of other columns which resulted in duplication of attribute values. The further step was detecting columns with high proportion of null values.

For all columns, modes were checked, and the columns that had the mode as null value were removed. The column removal operation was operated with different approaches for discrete and continuous columns.

Although most of the null values were removed from the dataset during column removal operation, there were still null values spread throughout the rest of the columns. The remaining null values needed to be swapped to logical measures. In this operation, depending on the column type, the null values were swapped with mean or mode value of the column.

Once ensured there were no left null values in the dataset, the correlation matrix was created for detecting the duplicate columns. There were also duplicate columns that were not detected in correlation matrix but known from the domain knowledge. These were also removed.

Some columns have values with low frequency that represent missing and rejected to answer responses in the survey. These values can be treated as noise since they could affect the scale of normalization. Depending on properties of the columns, the noises were eliminated by swapping them mean or mode of the columns. Before normalization, exploratory data analysis was practiced in the dataset as discussed below. Finally, the dataset was normalized but in the output columns, the label of having disease was 1 and 0 for healthy. The labels were swapped, and the dataset was ready for dealing with class imbalance and ML applications. See Appendix A.1 to see columns names left after data processing and detailed explanations of what they represent.

2.2 Data Exploration

In this step, valuable information can be extracted from our cleaned dataset. The BRFSS data can allow one to validate certain relationships presented in medical research as well as explore unexpected correlations. Having a good idea of the strength of any trends in the data can also provide an insight into how well classification models may perform.

2.2.1 Categorical Target Variable Analysis

First, suspected class imbalances for our target diseases were identified, these are all rare diseases so some imbalance towards the 'does not have disease' class was expected.

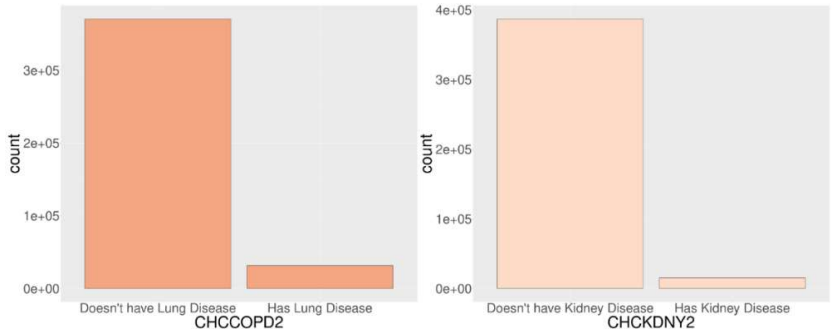


Figure 2.2.1: Bar charts plotted with ggplot2 demonstrating class representation for lung disease and kidney disease

Clearly, this imbalance was to be adjusted, any classifier model will be biased towards the majority class: 'Does not have disease x'. It is particularly strong for kidney disease, a very rare disease. To fix this problem, class balancing techniques were explored.

2.2.2 Categorical Predictor Variable Analysis

Next, the categorical attributes were investigated. Medical research shows strong relationships between exercise, smoking and heart disease, the BRFSS clearly supports this:

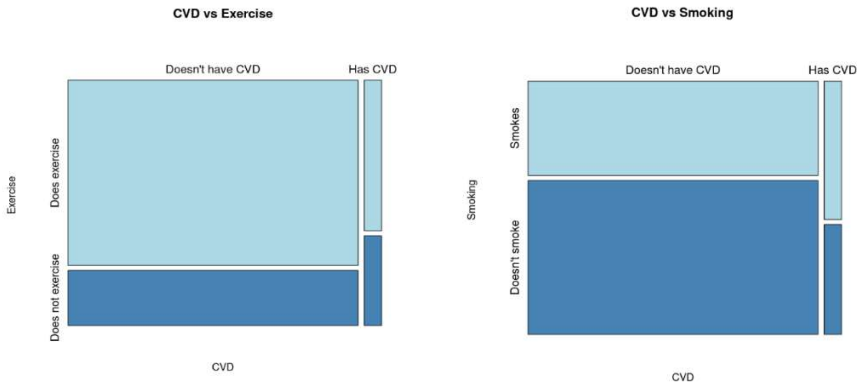


Figure 2.2.2.1 Mosaic plots plotted showing the proportion of who exercise/smoke and have CVD

This suggests the effects of exercise and smoking on CVD are opposite. There is clearly a smaller proportion of people who exercise and have CVD compared to those who don't have CVD. Whereas there is larger proportion of people who smoke and have CVD compared to those who don't smoke.

Another interesting correlation to check was between smoking and pulmonary disease, a clear link has been made in many research papers [11].

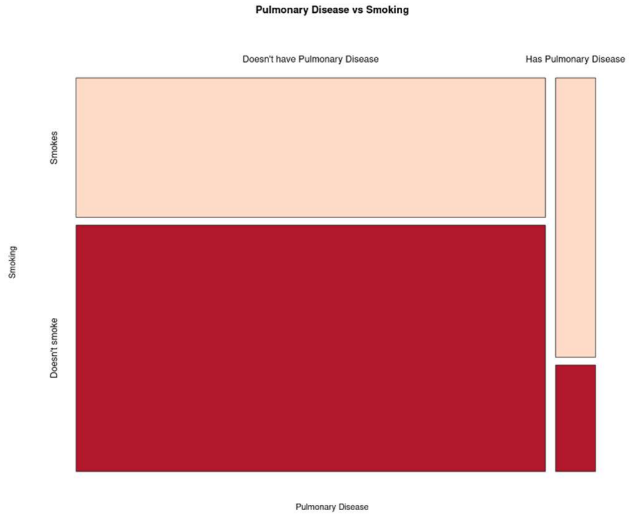


Figure 2.2.2.2: Mosaic plot demonstrating proportion of people who smoke and have pulmonary disease

The BRFSS data seems to validate this link. We can retrieve the exact proportions in this mosaic plot:

	Doesn't_have_Pulmonary_Disease	Has_Pulmonary_Disease
Smokes	0.36	0.72
Doesn't smoke	0.64	0.28

In the group of people who don't have lung disease, there is a higher proportion of those who don't smoke than those who do (64% do not smoke). Then for the group of those with lung disease, there is a much higher proportion of those who do smoke than those who do not (72% smoke).

After various visualisations of the data and reading available research [12], the following relationships between input and output fields were identified.

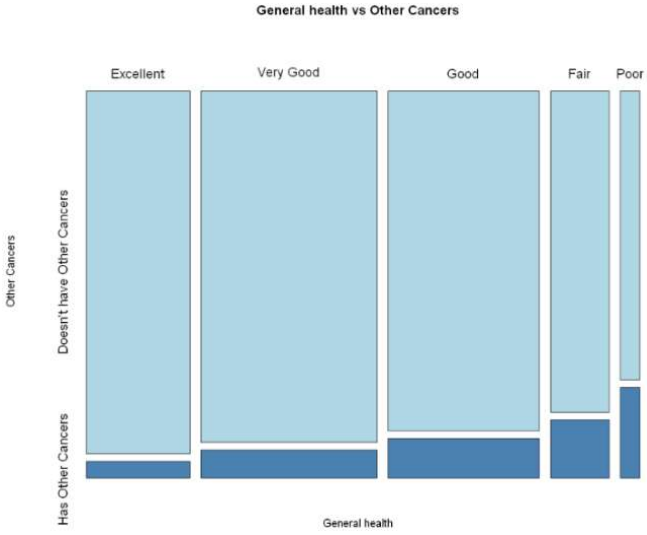


Figure 2.2.2.3: Mosaic plot demonstrating relationship between general health and whether a person has cancer

The poorer the general health, the higher the proportion of having any type of cancer. This also gives insight to the distribution of perceived health in the survey participants, most people seem to report themselves as in 'Very Good' or 'Good' health.

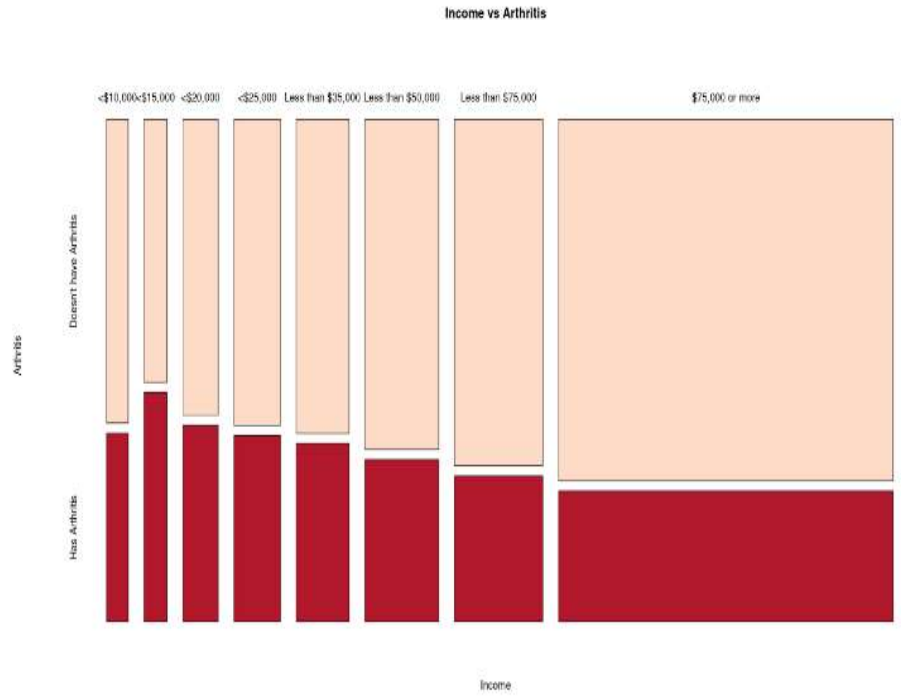


Figure 2.2.2.4 Mosaic plot demonstrating proportion of people who have arthritis in different income groups

The higher the income, the lower the proportion of people having arthritis. How the exact proportion varies for each income group can be obtained by extracting the exact values.

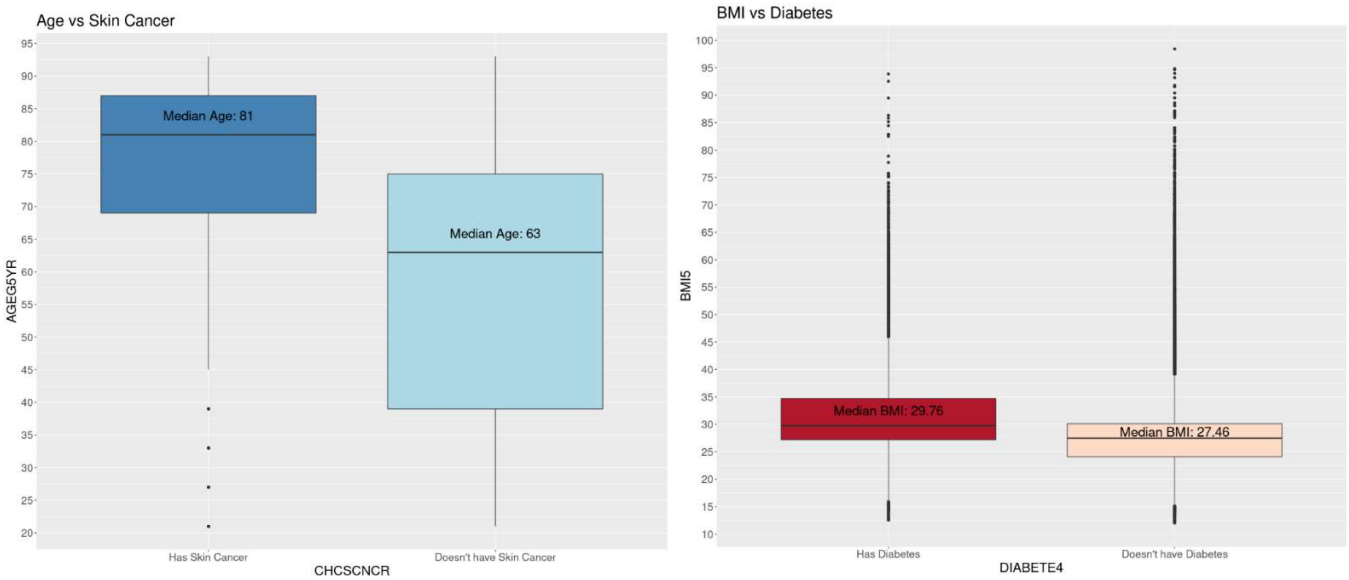
doesn't_have_Arthritis		Has_Arthritis
<$10,000	0.62	0.38
<$15,000	0.53	0.47
<$20,000	0.60	0.40
<$25,000	0.62	0.38
Less than $35,000	0.64	0.36
Less than $50,000	0.67	0.33
Less than $75,000	0.70	0.30
$75,000 or more	0.73	0.27

Excluding the first income group which does not fit the trend, when the income group increases, the proportion of people who do not have arthritis increases. 27% of people who earn above 75k have arthritis compared to 47% for those who earn less than 15k per year.

This provides a good insight into relationships between our categorical variables. The numerical features of our dataset were also investigated.

2.2.2 Numerical variable analysis

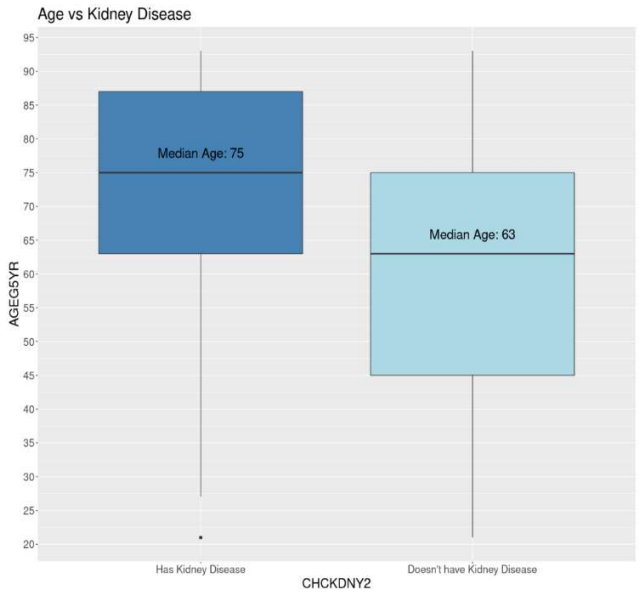
The effect of continuous features such as age and BMI on our target diseases was observed using boxplots.



Figures 2.2.2.1 and 2.2.2.2 Age/Skin Cancer and BMI/Diabetes relation depends on having disease or not

The group that has skin cancer has a much higher median age which follows current medical information around skin cancer risks [13]. The boxplots also give an idea of the spread of ages in each group.

The group having diabetes has a higher median BMI than those without diabetes which again follows medical research.



Once again the median age is a lot higher for those with the target disease, Kidney disease. Further plots of trends in the data can be found in the appendix A.3

2.2.3 Summary

The BRFSS data has severely unbalanced target classes and many of the predictor fields are subject to recall and social bias.

However, several strong relationships between the input and output variables have been identified. Both discrete and continuous features seem to impact the risk of having one of the target diseases. This gives good reason to believe classification models will be able to identify the trends in the BRFSS data and perform well. .

2.3 Data Splitting

Firstly, before addressing the class imbalance problem, 10% of the original dataset was taken for faster results and split it randomly into two subsets used for training and testing. It is common practice to use a bigger portion of the data to train the classification models, so a 70-30 split is implemented. This leaves enough data to perform evaluation and further analysis and ensures sufficient data left for the training. By setting the same 'seed' value while splitting the dataset, it is possible to generate identical data subsets for experiments whenever necessary. There are now 28,136 rows in the 'train' subset and 12,059 rows in the 'test' subset.

2.4 Class Balancing

Class imbalance can negatively affect the performance of the models and is a common issue faced while dealing with medical datasets as the vast majority of patients do not have the target disease being predicted. Overall accuracy as a performance evaluation criteria is not useful because if most of the samples used for training belong to the '0' class, it is likely to classify all samples as '0' and still attain a relatively high accuracy [14]. However, for this business case, it is more beneficial to correctly predicting the '1' class – where a target disease exists. Furthermore, since most algorithms work towards optimizing metrics such as error rates and disregard data distribution [14]; before implementing algorithm-level fixes like optimizing the decision threshold, it is recommended to use data-level resampling techniques to fix the imbalance for the training dataset [15]. The test subset does not need to be changed as it resembles a more realistic scenario and can be used to check the robustness of the models.

Common methods of resampling data shown in sources [14][15][16] include under-sampling, over-sampling, using a combination of both under-sampling and over-sampling, and synthetically generating data. For the experiments of this section, the Random Forest model was used to predict asthma. Resampling methods are deployed while keeping hyperparameters constant to observe the isolated effects of this resampling on balanced accuracy and sensitivity. Overall accuracy is given less importance here as the balanced accuracy provides a better insight by providing the average accuracy of each class.

The unsampled training set ignoring class imbalance achieves 50.99% balanced accuracy and only 2.17% sensitivity on the test data. This low value of sensitivity makes it evident that the classes are skewed.

Each method is briefly described below:

1. Over-sampling is when the minority class has its values replicated at random to match the size of the majority class.
2. Under-sampling randomly deletes instances from the majority class to match the size of the minority class.
3. When using a combination of both methods described above, the minority class is first over-sampled to reach a predetermined probability of occurrence with given dataset size. The majority class is then under-sampled to the same size.
4. There are many algorithms available to synthetically generate data. The ROSE function [17] is used here with default arguments which adopts a smoothed-bootstrap approach.

The resulting class sizes are displayed in Table 2.4.1. The combination and ROSE methods produce a dataset of the same size as the unsampled data while the under-sampled dataset is 73% smaller and the over-sampled dataset is 72% larger than the original.

	0	1	Total
Unsampled	24310	3826	28136
Undersampled	3789	3826	7615
Oversampled	24310	24141	48451
Under + Over	14162	13974	28136
ROSE	14057	14079	28136

Table 2.4.1: Class sizes for resampling methods

These newly generated subsets produce metrics depicted by graphs in the following Table 2.4.2.

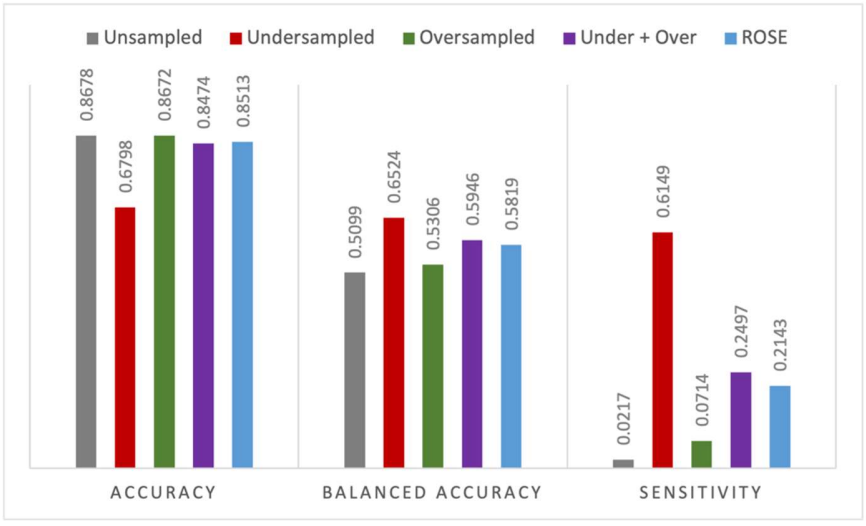


Table 2.4.2: Training results for resampled datasets

Clearly, under-sampling works best for the chosen dataset as the balanced accuracy is higher than that of the other methods and the sensitivity is significantly higher. The overall accuracy drops by about 10%.

However, it is not the best indicator of the quality of performance. Under-sampling may not always be the preferred way to solve an imbalance problem as data loss is generally not ideal. Nonetheless, since the size of the BRFSS dataset is immense, it is an appropriate tactic here. Additionally, training is about five times faster than the original dataset using this approach due to the reduced size. This will make the subsequent tests more time efficient and reduce the overall project time and computational resources used.

3. Model Preparation and Tuning

Decision trees and tree-based models were first proposed in the 1970s and have been the foundation of many algorithms since then [18]. For a binary classification problem, the idea is that all the predictor variables are recursively used to classify the target variable (a disease to be predicted) into the output classes (1 and 0) using a tree-like structure resembling a flow-chart. At each step of the tree, a decision is made regarding the classification using one predictor variable before moving on to the next. Decision trees are great for feature selection, easily interpretable, and robust [18] which is why they are still used especially on medical datasets [19],[20]. Three decision tree models were used to get optimal predictions for each disease including: Random Forests, C5.0, and XGBoost. Furthermore, a neural network was also implemented to compare the performance metrics of decision trees to a more complex approach.

3.1 Random Forest

Random Forest is an ensemble supervised machine learning algorithm used for classification meaning that it requires labelled data for training several decision trees. It has been described in detail in several books like [21] and [22]. Instead of creating one decision tree utilizing all the predictor variables of the dataset, the Random Forest algorithm generates an ensemble of decision trees with a set number of predictor variables used at random in each tree. The average of the outputs of all the trees is taken as the final classification prediction probability for the target variable. For example, if 65 out of 100 trees vote for class 1, then the class 1 probability would be 0.65. If the threshold is set as 0.5, the final prediction would still be class 1. However, if the threshold is set as 0.7, then it would be class 0.

Random Forests have many variations in mechanisms applied to randomize the selection of predictor variables driving the growth of trees. The 'randomForest' R package used for this project employs the algorithm developed by Breiman in 2001 [23] which builds on his previous work. This model is robust and includes the calculation of variable importances. Thus, the project chose this version of the Random Forest algorithm due to its ease of implementation, ready availability, and proven performance on medical data.

In the random forest algorithm, there are three main parameters that were altered to maximise performance. The first of these is 'ntree' which corresponds to the number of decision trees generated which are then used to vote on the class. 'mtry' determines the number of variables that are considered at each split in the decision trees. The value of 'mtry' should be low enough that there is enough randomness between the decision trees but not too small that the trees are unable to select important features. The final parameter we changed was the threshold, which determines what probability random forest needs to predict before assigning a class; in this report, what probability needs to be achieved before a person is predicted to have a disease.

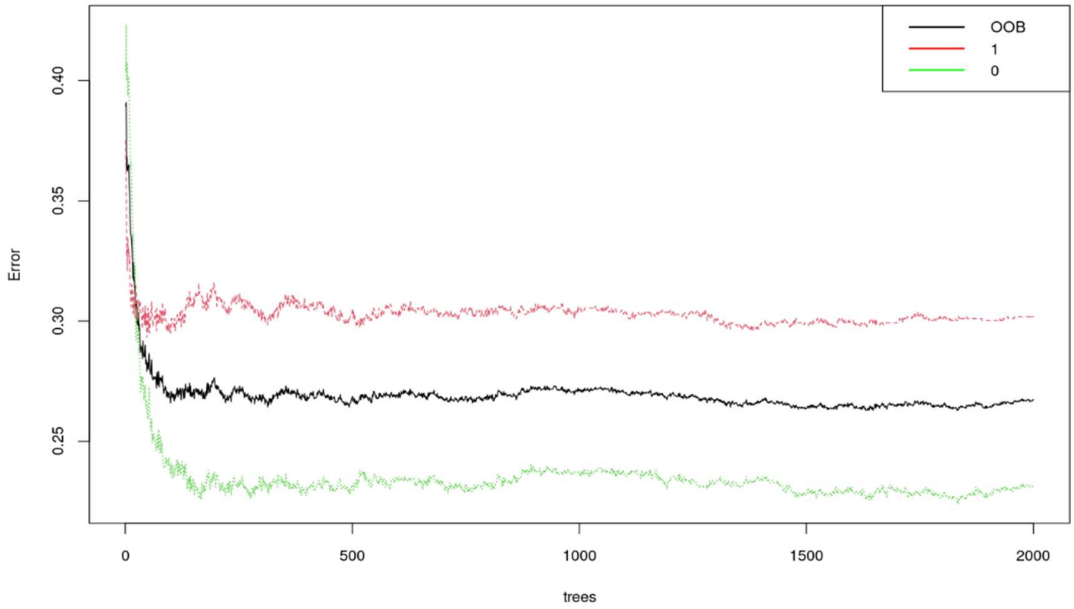


Figure 3.1.1: A plot of Out-Of-Bag error rate and the individual error rates for both classes of output field against number of trees in the random forest model

As trialling parameters requires the model to be trained multiple times, a subset which was a randomly sampled 10% of the original dataset was used to infer which values for the parameters would also give the best results on the full dataset. This was possible due to the fact the original dataset was very large. First, overall Out-Of-Bag error rate and the individual error rates for both classes were plotted against number of trees. It can be seen in the plot in Figure 3.1.1 that the error trend flattens past 200 trees and increasing the number of trees beyond 500 has a very low impact on the error but does significantly increase the training time for the model. For this reason, 'ntree' was set at 500 for all random forest models.

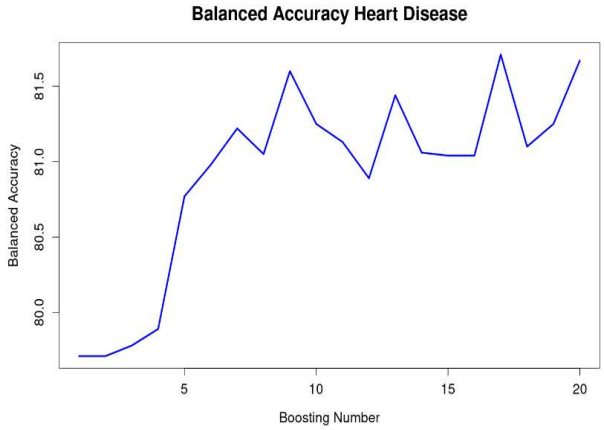
For each disease 7 different values of 'mtry' were trialled from 1 to 12, these values were based on the fact that 'mtry' is usually the square root of the number of features in classification problems [24]. The value of 'mtry' that gave the highest balanced accuracy for each disease was selected; the reasoning behind balanced accuracy being chosen here is explained in the evaluation methods section of the methodology. Plots of 'mtry' against balanced accuracy can be found in Appendix A.2.1

3.2 C5.0

To determine whether the patients included in the dataset were likely to have one of the 7 diseases previously mentioned a boosted decision tree classifier was employed called C50::C5.0. This algorithm

takes the role of a decision support system and through the application of a structure of decision trees, classifies the data into categories [25]. The classifier also has the capacity to be boosted. Boosting is a form of ensemble methodology. The training dataset is initially processed by a base classifier and then another classifier is put in place to target the falsely classified instances (in our case patients) by the first classifier. This is repeated until the number of boosts/trials inputted by the user is met. [26]

Figure 3.2.1

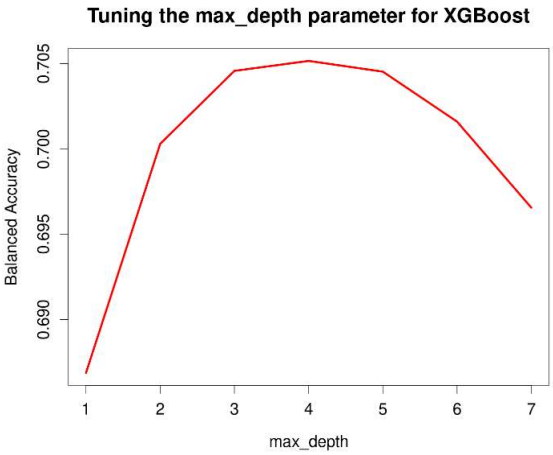


Before going extensively into the metrics for evaluation purposes, this algorithm was implemented on a smaller sample of the dataset so that the hyperparameters of C5.0 were tuned to the ideal setting. The hyperparameter examined for this model was boost (trials) number against balanced accuracy shown in Figure 3.2.1. The trend of the balanced accuracy displayed a significant increase in the range of boosting 0-7. After that, the balanced accuracy almost formed a plateau due to the model overfitting on the data. The highest balanced accuracy was obtained at a boost number of 17 and was later used to explore more metrics of the algorithm.

3.3 XGBoost

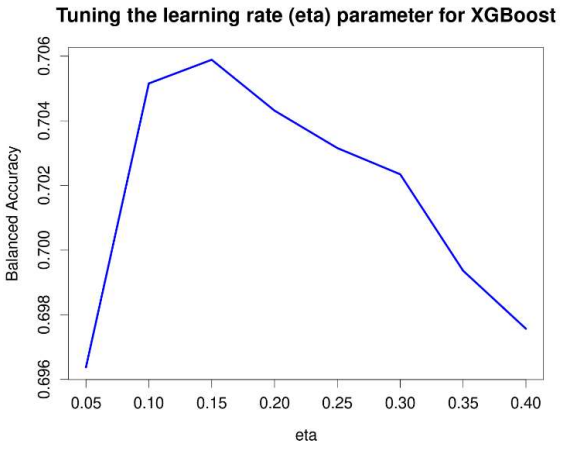
Xgboost (extreme gradient boosting) is a more sophisticated version of the gradient descent boosting technique (C5.0), which can increase the speed and efficiency of computation of the algorithm. Like Random Forest, it is an ensemble learning algorithm consisting of multiple decision trees. However, while Random Forest uses bagging to build full decision trees independently, XGBoost works by building decision trees one after another. Like C5.0, this makes XGBoost much faster than RandomForest (5 mins training time for XGBoost vs 40 mins training time for RandomForest).

Compared to C5.0, hyper-parameter tuning is much more complex and important with XGBoost. For this project, tuning was focused on 3 main parameters: max_depth: The maximum depth of each decision tree, eta: The learning rate and nrounds: The number of boosting rounds (basically the number of decision trees) [27]. The best way to select these was by visualising how they impact the most important metric for this business case: the balanced accuracy. All testing was done on a randomly sampled 10% subset of the original dataset.



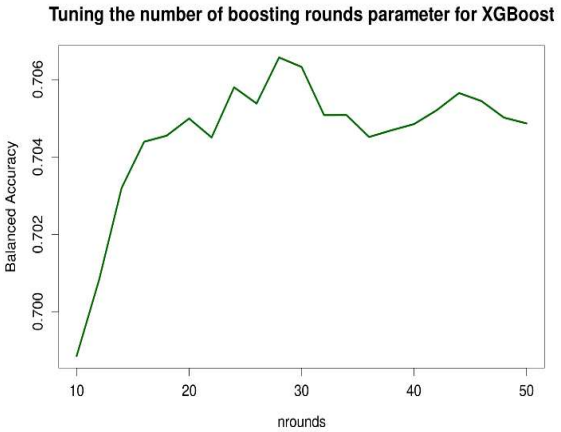
Here the optimal max_depth is 4.

Shallower trees tend to have poor performance because they capture few details of the problem. Deeper trees tend to overfit on the train dataset therefore limiting their ability to correctly predict on new data. The higher the max_depth, the higher the chance of overfitting.



Here the optimal eta (learning rate) is 0.15.

The learning rate represents how quickly the model converges to a good decision tree. So, the higher the better until it starts to overshoot (we can see that here this happens after 0.15) and it can no longer find a good solution/decision tree.



Here the optimal nrounds is 28.

This represents the number of 'weak learners' added to the model. The accuracy changes for each new tree added. The graph helps to see after which tree number to stop training.

The full hyper-parameter tuning plots for every disease can be found in appendix A.4

3.4 Neural Network

In the scope of this project, it was decided to create a neural network model that retrains itself iteratively for each disease and its balanced training dataset. In each training of the NN model, the model was tested with corresponding test dataset and the results were stored in a list to create confusion matrices once all iterations are done.

For the creation of the neural network, the 'h2o' library was used. The network architecture had three hidden layers and a hundred nodes in each layer. For the activation function selection of the model, the proposed model trains itself with backpropagation and gradient-based learning method. Therefore, the vanishing gradient problem might be a problem during training of the model, and it could stop training once the layer number increased [28]. Since the neural network modelled had three hidden layers, the vanishing gradient problem couldn't be considered in the scope of this project, and it was decided to use Tanh activation function.

The other parameter that had direct effect on both model's training performance and its complexity was the epoch count, how many times would the model backpropagate in each dataset iteration. The training performance could significantly improve per epoch but initializing a stopping round was necessary to detect the model that stopped learning already and stop it. As a result, the maximum epoch was indicated as 100 and after 10 epochs with no improvement the model would stop.

During the search of other hyperparameters which adjusting could be necessary for the model performance, it was observed that 'h2o' library doesn't score all training and validation samples as default, these parameters were defaulted to ten-thousand samples [29]. Scoring all samples was crucial for the model's real performance, therefore these parameters were adjusted for scoring all samples in the training dataset.

3.5 Model Threshold Adjustment

The models were then trained on the full dataset, not including the testing data. Once this was complete, the threshold was adjusted from the default of 0.5. Provost writes how using the standard value for threshold in classification problems with imbalanced data, like the data used in this report, can be a critical mistake [30]. It is possible to achieve high accuracy on imbalanced datasets by setting a high threshold so nearly all true negative cases are identified correctly, but this defeats the point of creating a model which can predict positive cases for a disease. It is therefore important a threshold is selected that correctly identifies as many true positives as possible without the false positive rate becoming too high. To do this the threshold was selected that maximised the True Positive Rate (TPR) and minimised the False Positive Rate (FPR). This was done in the form of minimising a Euclidean distance function, shown by the equation below:

$$Euclidean\ distance = \sqrt{(1 - TPR)^2 + FPR^2}$$

4. Performance Metrics

The metrics chosen to judge and measure the model performance were specificity, sensitivity, threat score, miss rate and balanced accuracy.

4.1 Balanced accuracy

It is noted that "Accuracy refers to a measure of the degree to which the predictions of a model match the reality being modelled." [31]. It can be a useful measure if we have similar balance in the dataset. However, as we were dealing with an imbalanced dataset, we wanted to avoid cases where only one class can be predicted by the model. K. H. Brodersen et al. observed that while under sampling the large class may under some circumstances prevent a classifier from being biased, it does not provide generic safeguards against reporting an optimistic accuracy estimate [32]. With this observation, a better approach would be to use "a different generalizability measure: the balanced accuracy, which can be defined as the average accuracy obtained on either class." [31] . This can be seen in (1) below:

$$balanced\ accuracy\ =\ \frac{sensitivity\ +\ specificity}{2} \qquad (1)$$

4.2 Sensitivity (TPR)

Sensitivity is also known as the true positive rate (TPR). In this context, it refers to the classifier's ability to correctly identify individuals with a high risk of getting a disease. It can be expressed as the ratio between the number of true positives and the sum of the number of true positives and false negatives. This is illustrated in (2) below [33]:

$$sensitivity\ =\ \frac{TP}{TP\ +\ FN}\ =1-FNR \qquad (2)$$

4.3 Specificity (TNR)

Specificity is also known as the true negative rate (TNR). This is a measure of how well the classifier can identify true negatives, in this case, correctly identify low risk individuals among the total number of those in low risk as seen in (3) below [33]:

$$specificity\ =\ \frac{TN}{TN\ +\ FP}=1-FPR \qquad (3)$$

4.4 Threat Score

This is the ratio between the number of correctly predicted positive samples against the sum of correctly predicted positive samples and all incorrect predictions [34]. It considers false alarms and missed events in a balanced way and excludes only the correctly predicted negative samples [34]. This illustrated in (4) below:

$$threat\ score\ =\ \frac{TP}{TP\ +\ FN+\ FP} \qquad (4)$$

4.5 Miss rate

This is the ratio between false negatives against the sum of false negatives and true positives [34] (5).

$$miss\ rate\ =\ \frac{FN}{TP\ +\ FN}=\ 1\ -\ TPR \qquad (5)$$

5. Results

5.1 Performance Metrics

Tables were produced for each of the four models with their performance in each of the metrics for each disease.

Disease	Sensitivity	Specificity	Miss Rate	Threat Score	Balanced Accuracy
Asthma	0.6389	0.6747	0.3611	0.2071	0.6568
Skin Cancer	0.7658	0.7028	0.2342	0.1905	0.7343
Heart Disease	0.8071	0.7957	0.1929	0.1835	0.8014
Kidney Disease	0.7787	0.7222	0.2213	0.0940	0.7504
Arthritis	0.7492	0.7293	0.2508	0.4660	0.7392
Lung Disease	0.8031	0.7930	0.1969	0.2329	0.7980
Diabetes	0.7811	0.7305	0.2189	0.2776	0.7558
Other Cancers	0.7477	0.6730	0.2523	0.1754	0.7104

Table 5.1.1: Random Forest performance metrics across all eight diseases

Disease	Sensitivity	Specificity	Miss Rate	Threat Score	Balanced Accuracy
Asthma	0.6399	0.6781	0.3601	0.2089	0.6590
Heart Disease	0.8058	0.7998	0.1942	0.1861	0.8028
Kidney Disease	0.7556	0.7378	0.2444	0.0960	0.7467
Arthritis	0.7525	0.7271	0.2475	0.4667	0.7398
Lung Disease	0.8094	0.7838	0.1906	0.2275	0.7966
Diabetes	0.7684	0.7425	0.2316	0.2812	0.7554
Skin Cancer	0.7647	0.7076	0.2353	0.1926	0.7362
Other Cancers	0.7368	0.6901	0.2632	0.1801	0.7134

Table 5.1.2: C5.0 performance metrics across all eight diseases

Disease	Sensitivity	Specificity	Miss Rate	Threat Score	Balanced Accuracy
Asthma	0.6554	0.6636	0.3446	0.2076	0.6595
Heart Disease	0.8016	0.8018	0.1984	0.1865	0.8017
Kidney Disease	0.7633	0.7279	0.2367	0.0939	0.7456
Arthritis	0.7552	0.7297	0.2448	0.4700	0.7425
Lung Disease	0.8006	0.7866	0.1994	0.2272	0.7936
Diabetes	0.7648	0.7427	0.2352	0.2800	0.7538
Skin Cancer	0.7686	0.7041	0.2314	0.1919	0.7363
Other Cancers	0.7553	0.6704	0.2447	0.1761	0.7128

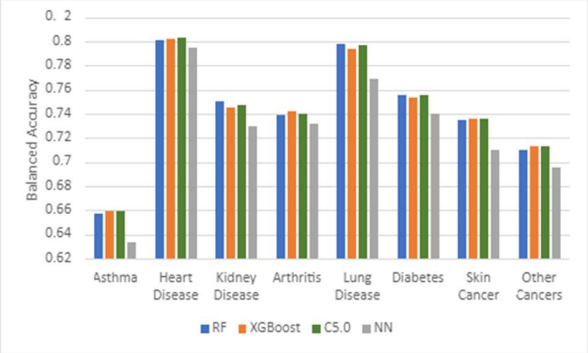
Table 5.1.3: XGBoost performance metrics across all eight diseases

Disease	Sensitivity	Specificity	Miss Rate	Threat Score	Balanced Accuracy
Asthma	0.5565	0.7104	0.4435	0.1948	0.6335
Skin Cancer	0.7507	0.6696	0.2493	0.1723	0.7102
Heart Disease	0.7899	0.7995	0.2101	0.1822	0.7947
Kidney Disease	0.7211	0.7382	0.2789	0.0917	0.7296
Arthritis	0.7195	0.7444	0.2805	0.4571	0.7320
Lung Disease	0.7470	0.7910	0.2530	0.2151	0.7690
Diabetes	0.7671	0.7131	0.2329	0.2618	0.7401
Other Cancers	0.7367	0.6546	0.2633	0.1657	0.6956

Table 5.1.4: Neural Network performance metrics across all eight diseases

Two further tables and corresponding bar plots were produced to better allow comparison between models in each disease.

Disease	RF	XGBoost	C5.0	NN
Asthma	0.6568	**0.6595**	0.6590	0.6335
Heart Disease	0.8014	0.8017	**0.8028**	0.7947
Kidney Disease	**0.7504**	0.7456	0.7467	0.7296
Arthritis	0.7392	**0.7425**	0.7398	0.732
Lung Disease	**0.7980**	0.7936	0.7966	0.769
Diabetes	**0.7558**	0.7538	0.7554	0.7401
Skin Cancer	0.7343	**0.7363**	0.7362	0.7102
Other Cancers	0.7104	0.7128	**0.7134**	0.6956



Table/Figure 5.1.5: Comparison of balanced accuracy for each model in each disease

Disease	RF	XGBoost	C5.0	NN
Asthma	0.6389	**0.6554**	0.6399	0.5565
Heart Disease	**0.8071**	0.8016	0.8058	0.7899
Kidney Disease	**0.7787**	0.7633	0.7556	0.7211
Arthritis	0.7492	**0.7552**	0.7525	0.7195
Lung Disease	0.8031	0.8006	**0.8094**	0.747
Diabetes	**0.7811**	0.7648	0.7684	0.7671
Skin Cancer	0.7658	**0.7686**	0.7647	0.7507
Other Cancers	0.7477	**0.7553**	0.7368	0.7367

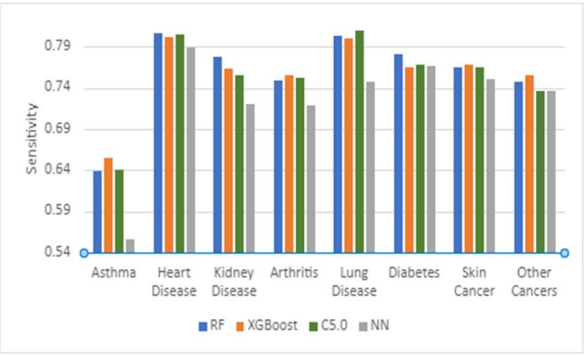


Table /Figure 5.1.6: Comparison of sensitivity for each model in each disease

5.2 Feature Importances

After obtaining the evaluation metrics, the features that played a key role in determining the desired output, which is whether a patient has a disease or not, were determined to introduce an alternative method in testing the robustness of our algorithms. Some research was then performed on the causal factors of a disease according to a certified medical organization, NHS, to examine their correlation.

Each of the different classifiers applied included a different method of calculating the importance for all the features. Then the top five most important features selected from each algorithm were compared with the most important features according to NHS displayed in Figure 5.2.1.

Figure 5.2.1

Pulmonary Disease	Arthritis: Osteoarthritis	Arthritis: Rheumatoid	Asthma	Skin Cancer	Kidney Disease	CVD	Diabetes
Smoking	joint injury	being a woman	allergies (eczema, food allergy or hay fever)	previously damaged your skin through sunburn or radiotherapy treatment	high blood pressure	High blood pressure (hypertension), High blood cholesterol, Diabetes	age
Fumes and dust at work	secondary arthritis	family history	family history (asthma or atopic conditions)	close relative who's had melanoma skin cancer	diabetes	Smoking,Excessive alcohol consumption	family history
Air pollution	age	smoking	bronchiolitis (childhood lung infection)	pale skin that does not tan easily	acute kidney injury	Poor diet, Lack of exercise, Being overweight	overweight or obese
close relative with the condition (family history of lung problems)	family history	age	exposure to tobacco smoke as a child	red or blonde, hair blue eyes, several freckles	cardiovascular disease	Stress	ethnicity
body mass index (BMI) using your weight and height	obesity		mother smoking during pregnancy	a condition that suppresses your immune system, such as diabetes or you take medicines that suppress your immune system	conditions that can affect the kidneys (kidney stones, an enlarged prostate or lupus)	Family history	
	being a woman		born prematurely	risk of developing skin cancer also increases with age	family history of advanced CKD	Ethnicity	

[35] Starting with the Random Forest classifier (RF), one could see, by the first two bar charts Figure 5.2.2 and Figure 5.2.3, there are two metrics of feature importance inbuilt into the algorithm: the mean decrease in accuracy and the mean decrease in Gini coefficient. The mean accuracy decrease is calculated through executing random permutations of each feature's values to obtain the percentage increase in the misclassification rate and deducting from it the original test set error of that feature [23]. The mean decrease in Gini is a measure of how much a variable on average decreases class impurity in the nodes of the trees it is present in. There is an evident correlation between the features in the table and the two bar charts. Smoking is present in all figures and body mass index (BMI) is strongly related with the general health (GENHLTH). Some of the other causal factors mentioned by NHS like air pollution or family history were not included in the survey. Therefore, the features(variables) the algorithm deems significant in the determination of the output coincide with the causal disease factors by NHS, validating the good performance of RF.

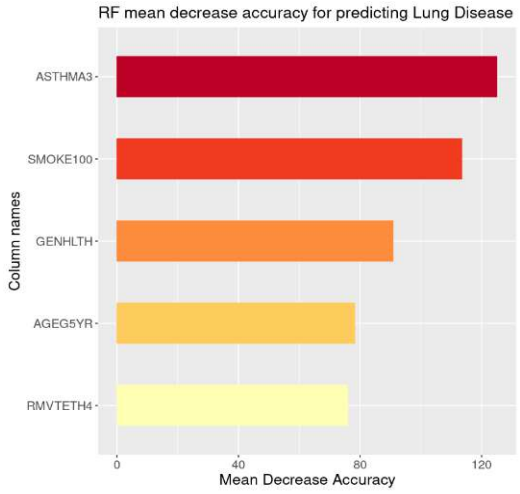


Figure 5.2.2

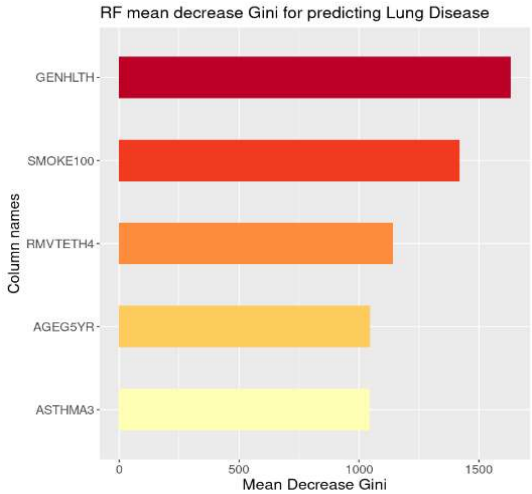


Figure 5.2.3

Following that, XGboost importance metric 'xgb.importance' was employed to explore the key features for predicting Arthritis (Figure 5.2.4). The in-built importance function uses a measure called gain. That measure reveals the way the importance of a specific feature in the dataset contributes in making the branches of the decision trees used in the model purer. Since this is a boosting algorithm, it considers each gain of each feature of each tree. Finally, to get a complete view of the model an average of the gains per featured is calculated [36].

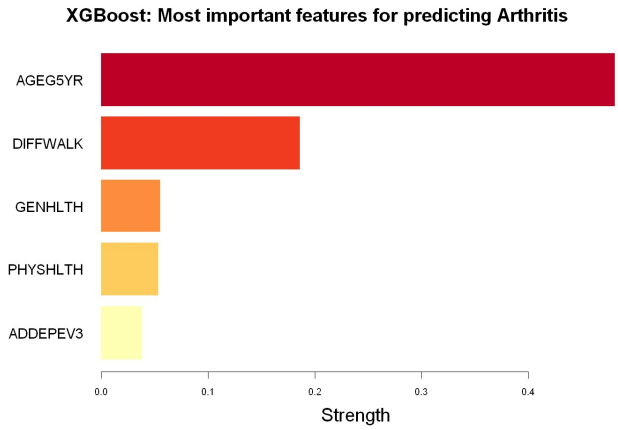


Figure 5.2.4

The performance of the XGboost model can be tested through the comparison of the two columns of the NHS table, Arthritis: Osteoarthritis and Rheumatoid, with the bar chart generated by the importance metric in Figure 5.2.4. The age feature plays a key role in both the table columns, and the bar chart. The causal factor of smoking pointed out by NHS can be linked to general health (GENHLTH). Additionally, the physical health (PHYSHLTH) feature in Figure 5.2.4 can involve the joint injury and obesity causal factors in the table. Therefore, this comparison proves the robustness of the XGboost algorithm, in patient classification.

In the case of the C50::C5.0 decision tree classifier, the C5imp metric was used to obtain the importance of features in determining whether a patient had Asthma, Figure 5. For each of the features in the dataset C5imp calculates the percentage of training set samples that are used in the terminal nodes of the decision trees of the model after splitting [37] and that percentage was displayed in Figure 5.2.5. The metric revealed that all the features it considered as important, chronic obstructive pulmonary disease (CHCCOPD2), age (AGEG5YR), general health (GENHLTH), body mass index (BMI5) and depressive disorder (ADDEPEV3), were not related to the causal factors determined by NHS. However, that does depreciate C5.0's performance capacity since the questions asked in the survey to obtain this dataset were not targeted towards the asthma causal factors. Thus, factors like allergies, family history or bronchiolitis did not take part in the training of any

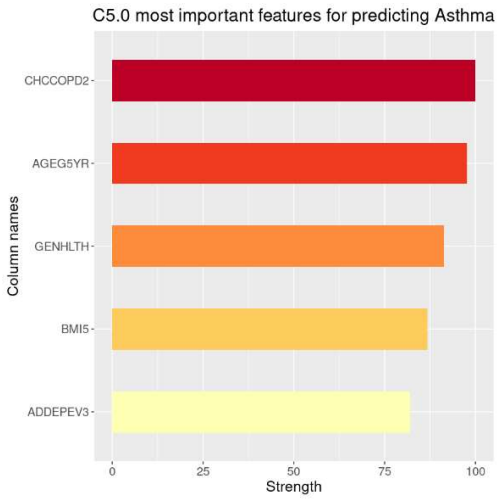


Figure 5.2.5

of the classifiers tested, leading to their lower performance in classifying patients with asthma.

With a more complete picture of the overall performance of the classifiers tested, we have the capability of deploying said algorithms to the real business world. When applied in hospitals or insurance companies, the models can help improve the fiscal, managerial, and service side of the business. More on that was investigated in the following section.

6. Discussion

6.1 Model Interpretation

6.1.1 Arthritis

For both balanced accuracy and sensitivity, XGBoost stands out as the best performing model.

To understand why it does better here than Random Forest and the C5.0 boosted decision tree, we can investigate which features it determined to have the most impact on predicting Arthritis:

1. Age
2. Difficulty walking

Arthritis most often develops in people in their mid-40s or older [38], so age is a reliable indicator of arthritis risk. Knowing the individual's age will therefore be crucial to insurance companies and hospitals.

However, difficulty walking is more of a consequence of having arthritis than a predictor. Perhaps this field will not be useful for insurance companies and hospitals as the individuals reporting difficulty walking will most likely already have an arthritis diagnosis.

6.1.2 Skin Cancer

Both types of boosted decision tree algorithms, XGBoost and C5.0, had a high balanced accuracy when fitted to the testing dataset. However, C5.0 achieved slightly poorer sensitivity compared to XGBoost.

Sensitivity being a key metric for our business case, we select the XGBoost classifier. This model indicates that the most important features in the prediction of Skin Cancer are:

1. Age
2. Ethnicity

This is strongly supported by medical research [39], [13]. Cell DNA damage accumulates over time spent exposed to the sunlight, the elderly are therefore most at risk of skin cancer. Fair-skinned people are also more at risk as they have less of the protective pigment called *melanin.*

Clearly, knowledge of an individual's age and ethnicity is vital to be able to accurately assess their risk of skin cancer. Insurance companies or hospitals should collect these details.

6.1.3 Other Cancers

All models performed well at predicting other cancers as can be seen by a consistent balanced accuracy of around 71%, fitted to the testing dataset. However, XGBoost achieves the best sensitivity.

The model identifies the following most important features:

1. Age
2. General Health (the individual's personal opinion)

Once again, age is the most decisive factor in predicting the risk of other cancers for the same reason: Cell DNA damage accumulates over time. The impact of perceived General Health is more interesting as it indicates that this question was answered reliably. Individuals have a good sense of their overall health. It is therefore worth it for insurance companies or hospitals to ask an individual how they perceive their overall health.

6.1.4 Asthma

All models performed poorly at predicting Asthma as can be seen by a consistent balanced accuracy of around 66%, fitted to the testing dataset.

XGBoost was marginally better than the other models across both balanced accuracy and sensitivity.

To understand why the metrics are so low, it is useful to investigate which attributes the model found to have most impact on the risk of asthma:

1. Having any kind of pulmonary disease
2. Age
3. Having a depressive disorder

According to medical research, the most common causal factor of asthma is genetic [40]. Therefore, a more suitable question to ask individuals would be "Is there any history of asthma in your family?". Information such as age and whether they have a depressive disorder is not very useful in the prediction of asthma.

As for whether they have any kind of pulmonary disease, this field should perhaps be removed when training the model. Asthma is a type of pulmonary disease, the 2 are highly correlated and therefore we shouldn't use it to train a model to predict asthma.

6.1.5 Heart Disease

All models performed well at predicting heart disease as can be seen by a consistent balanced accuracy of around 80%, however, random forest achieved the best specificity as it was best able to identify positive cases. The random forest model found that the most important features for identifying heart disease were whether the individual had ever suffered from a heart attack or stroke, their age, their BMI, individuals' opinions of their own health and their sex. Heart attacks and strokes, like heart disease, are caused by restricted blood flow so this importance agrees with previous literature [41]. Age, gender and BMI have also been shown by the National Center for Chronic Disease Prevention and Health Promotion to be strongly linked to heart disease [42]. Individuals' opinions of their own health are something that has been identified by some research as having value for both clinical use [43] and subsequent mortality [44] but its importance here and in many of the other diseases confirms it has value as a predictor of chronic diseases.

6.1.6 Lung Disease

Random forest, XGBoost and C5.0 all performed strongly at predicting lung disease with balanced accuracies and sensitivities of 80%, however, C5.0 was marginally better than the other models across

both these metrics. Asthma, smoking, age were all features that were identified as being important for classification by the model which was to be expected as the American Lung Association list smoking as the biggest risk factor as well as identifying the link between other respiratory conditions [45]. Individuals' opinion of their general health again was an important feature but most interestingly the number of teeth removed due to decay or gum disease was one of the most powerful predictors. It is unlikely to be just that smoking also causes gum diseases because if these factors were strongly correlated removal of one would not have had such an impact on the mean decrease in accuracy for random forest. This is not listed by either the WHO [46] or the American Lung Association [45] as a risk factor but Bansal, Khatri and Taneja in a review of the potential role of periodontal infection in respiratory diseases found that bacteria in the mouth can be aspirated into the lungs triggering inflammation [47].

6.1.7 Kidney Disease

Random forest, XGBoost, and C5.0 performed similarly on balanced accuracies for predicting kidney disease; Random Forest taking the lead with 75%. However, in terms sensitivity, Random Forest was much higher at almost 78% than the other four models. Overall, the model has a similar performance when predicting kidney disease as compared to the other diseases. Upon plotting the model importances, it is seen that kidney disease is dependent on the individual's general health, age, BMI, and whether or not they have diabetes. Overall, this aligns with research from the NHS [35] as they identified diabetes as one of the leading factors correlated to kidney disease; another factor is if the individual has high blood pressure which was not asked as a part of the survey.

6.1.8 Diabetes

Again, the three decision tree models had comparable performance on balanced accuracy (within 0.2%) to predict diabetes at about 76%. The Random Forest however, was marginally higher on sensitivity values at 78%. The model estimated general health, age, and BMI as variables of the highest importance as per the Gini coefficient and balanced accuracies. General health is just each individuals' opinions about their own health and is a more qualitative variable here. However, age and BMI are also identified as the main causal factors of diabetes by the NHS [35]. According to them, individuals over 40 and those that are overweight are more susceptible to the disease. They also claim that ethnicity is correlated to diabetes; contrary to their claim, the Random Forest model did not deem this variable as important.

6.2 Business Case

"The likelihood that patients with a particular condition such as heart failure or diabetes will use expensive health care resources such as hospital care increases substantially with the presence of other comorbidities" [48]. This corresponds with research conducted by Menzin et al. which shows that the most significant factors influencing the cost of treating a patient for heart disease are those associated with rehospitalization and outpatient care [49]. Menzin et al. also found that new treatment options that reduce subsequent resource utilization could potentially offer a substantial overall cost savings [49]. Our models can be used in reducing these costs for the medical insurance companies and medical practices.

A higher balanced accuracy means that the classifier would be able to correctly classify the risk of disease, and this would help the insurance companies to assess the financial risks associated with insuring their potential customers and adjust their premiums and policies accordingly. "The care of individuals with chronic conditions is estimated to account for 78% of health expenditures in the United States. Patients with more than 1 chronic condition are estimated to account for 95% of all Medicare spending" [48]. With

this intelligence, insurance companies could charge more premiums to the customers classified as high risk, as this would significantly reduce the costs associated with covering high-risk diseases should an event occur. Medical practices would also use this information to adequately plan out their resources to cope with both increasing and decreasing demand. They would also be able to offer early intervention programmes targeting high risk patients and reduce the number of patients that will require inpatient care in the long term, reducing both manpower and cost. This is in line with Tafazzoli et al. who found that multi-cancer early detection testing shifted cancer diagnoses to earlier stages decreasing per cancer treatment costs by $5421 [50].

A high miss rate of 36% of asthma, would indicate that the classifier would be unable to confidently classify the risk of this disease. This is whereby the model classifies an individual as being of low risk, when they are indeed at high risk of contracting the disease. Such implications would prove to be costly to the insurance companies as this puts them at risk of loss when they insure such a customer with a low premium. Similarly, medical practices will be at risk of straining their resources if a high-risk patient is misclassified. The importance observed from the models can be used to inform the survey questions that the insurance companies and the medical practices would ask their clients and patients respectively. Consequently, the models' accuracies can be improved, and this means better classification of the risk of serious illness.

One weakness in all models in the number of false positives identified. This may be skewed due to at risk individuals who haven't yet been diagnosed but is also the case that the models will identify individuals incorrectly. However, in the business context it is more important that the number of false negatives is reduced as a medical practice failing to identify an at risk individual is far more dangerous than them offering treatment to someone who was not going to develop the disease.

As well as offering medical practices and insurance companies the ability to identify patients or customers who are at risk of developing a chronic disease, the models also offer insight into which risk factors are most important. For example, the fact that gum disease is highly correlated with lung disease as discussed earlier. This would allow insurance companies to better tailor their pre-offer questionnaires to best identify the risk to their business when taking on a customer. It would also allow medical practices to identify the key causal factors amongst their patients, so they best know where to target early intervention. For example, BMI was one of the most correlated factors for heart disease, so targeted campaigns at weight loss would be highly effective at reducing cases of this disease. They would also know that men are most likely to develop heart disease, and so can tailor campaigns like this even further.

7. Conclusion

7.1 Summary of Key Findings

As seen in Table 5.1.5 the models were able to predict lung disease and heart disease with a balanced accuracy of approximately 80% and kidney disease, diabetes, arthritis, skin cancer and other cancers with a balanced accuracy of ranging from 70-76%. They were less successful at predicting asthma, this may be because the causes of this are mostly genetic, but this could be tested with further data. The models were also able to identify the causal factors of each disease which agreed with prior research as well as identify some interesting risk factors that were not as well publicised.

The models created here would be able to be implemented by an insurance company to identify the risk new customers pose, when deciding whether to offer them insurance and what premium to charge.

They would also help insurance companies identify what questions to ask potential customers to best inform these decisions.

The models created here would be able to be implemented by a medical practice to identify which patients are at risk of these chronic diseases and therefore know who to target early intervention treatment towards. For example, prescribing patients aspirin to reduce the risk of them developing heart disease. This would save the practice cost in the long run as patients who eventually develop these diseases are very costly [3]. The models would also allow the medical practice to identify which are the key causal factors for these diseases and which groups of individuals are most at risk and, therefore, target their interventions accordingly.

7.2 Future Directions

The accuracy and outcomes of this report were limited by the dataset as the individuals interviewed could have been asked different questions more relevant to specific diseases. The models could be improved if worked on in collaboration with the data gathering. This would allow different questions to be tested to see their impact on the accuracy of the models. It would also allow for data to be gathered on other diseases to create models for their prediction as well. Working with medical practices would also open up the potential of bringing in further data sources such as medical testing allowing for information on individuals' cholesterol levels or blood pressure which are established risk factors for heart disease [51]

In particular, the models here struggled to predict Asthma, it may be that this has strong genetic links and is affected less by environmental factors, research by Mukerjee and Zhang found that asthma is triggered by a complex association of genetic and environmental factors [52]. Gathering more and different data would allow for this to be investigated further.

Given more time and computing power, the model hyperparameters could be further tuned to boost performance. For example, more complex neural networks could be trialled or further boosting could be completed on the decision trees. As discussed in the introduction, other studies have found success using different algorithms. With more time it would be possible to create models using techniques such as Support Vector Machines to see if these could produce similar or better results.

Another limitation of this report was the details on costs associated with the treatment of chronic diseases and the costs of early intervention treatments. Working in collaboration with a medical practice would allow for a detailed assessment of the exact cost-benefit of implementing these models into day-to-day practice for clinicians. For example, one application could be to further classify patients into risk groups (low/medium/high) for better resource management and reducing patient waiting times. Working in conjunction with an insurance company would allow us to implement models with cost data to potentially provide the insurance company with accurate cost estimates for individual customers. These cost estimates could further be used to classify individuals into several premium brackets and minimise losses to the company.

References

[1] Columbia Public Health (2022) *Risk Prediction.* Available at: https://www.publichealth.columbia.edu/research/population-health-methods/risk-prediction (Accessed 23 November 2022)

[2] World Health Organization June 2021, *Cardiovascular diseases (CVDs),* viewed 2nd November 2022

Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[3] Nichols, G.A., Bell, T.J., Pedula, K.L. and O'Keeffe-Rosetti, M., 2010. Medical care costs among patients with established cardiovascular disease. *The American journal of managed care*, *16*(3), pp.e86-e93.

[4] Centres for Disease Control and Prevention (2020) '2020 BRFSS Data'. Available at: https://www.cdc.gov/brfss/annual_data/annual_2020.html (Accessed October 2022)

[5] Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction - Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), pp.43-48.

[6] Uddin, S., Khan, A., Hossain, M. *et al.* Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform* Decis Mak 19, 281 (2019). https://doi.org/10.1186/s12911-019-1004-8

[7] Ayon, S.I., Islam, M.M. and Hossain, M.R., 2022. Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. *IETE Journal of Research*, *68*(4), pp.2488-2507.

[8] Hossain, M.E., Khan, A., Moni, M.A. and Uddin, S., 2019. Use of electronic health data for disease prediction: A comprehensive literature review. *IEEE/ACM transactions on computational biology and bioinformatics*, *18*(2), pp.745-758.

[9] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in *IEEE Access*, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.

[10] Centres for Disease Control and Prevention (2020) '2020 BRFSS Codebook CDC'. Available at: https://www.cdc.gov/brfss/annual_data/2020/pdf/codebook20_llcp-v2-508.pdf (Accessed October 2022)

[11]: Laniado-Laborín, Rafael. "Smoking and chronic obstructive pulmonary disease (COPD). Parallel epidemics of the 21 century." International journal of environmental research and public health vol. 6,1 (2009): 209-24. doi:10.3390/ijerph6010209 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2672326/

[12] Steyn K, Damasceno A. Lifestyle and Related Risk Factors for Chronic Diseases. In: Jamison DT, Feachem RG, Makgoba MW, et al., editors. Disease and Mortality in Sub-Saharan Africa. 2nd edition. Washington (DC): The International Bank for Reconstruction and Development / The World Bank; 2006. Chapter 18. Available from: https://www.ncbi.nlm.nih.gov/books/NBK2290/

[13] Garcovich, Simone et al. “Skin Cancer Epidemics in the Elderly as An Emerging Issue in Geriatric Oncology.” Aging and disease vol. 8,5 643-661. 1 Oct. 2017, doi:10.14336/AD.2017.0503 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5614327/

[14] T. Fawcett, “Learning from Imbalanced Classes - Silicon Valley Data Science,” Silicon Valley Data Science, Aug. 25, 2016. https://www.svds.com/learning-imbalanced-classes/ (accessed Nov. 23, 2022).

[15] A. Singh, R. K. Ranjan, and A. Tiwari, “Credit Card Fraud Detection under Extreme Imbalanced Data: A Comparative Study of Data-level Algorithms,” Journal of Experimental & Theoretical Artificial Intelligence, vol. 34, no. 4, pp. 571–598, Apr. 2021, doi: 10.1080/0952813x.2021.1907795.

[16] G. Menardi and N. Torelli, “Training and assessing classification rules with imbalanced data,” Data Mining and Knowledge Discovery, vol. 28, no. 1, pp. 92–122, Oct. 2012, doi: 10.1007/s10618-012-0295-5.

[17] N. Lunardon, G. Menardi, and N. Torelli, “ROSE: a Package for Binary Imbalanced Learning,” The R Journal, vol. 6, no. 1, pp. 79–89, 2014, doi: 10.32614/rj-2014-008.

[18] G. Louppe, “Classification and Regression Trees,” in Understanding Random Forests: From Theory to Practice, University of Liege, 2015. Accessed: Nov. 24, 2022. [Online]. Available: https://arxiv.org/pdf/1407.7502.pdf

[19] M. Bhagat and B. Bakariya, “Prediction of Heart Disease Through KNN, Random Forest, and Decision Tree Classifier Using K-Fold Cross-Validation,” in Artificial Intelligence and Sustainable Computing, Nov. 2022, pp. 67–75. doi: 10.1007/978-981-19-1653-3_6.

[20] M. M. Ghiasi and S. Zendehboudi, “Application of decision tree-based ensemble learning in the classification of breast cancer,” Computers in Biology and Medicine, vol. 128, p. 104089, Oct. 2021, doi: 10.1016/j.compbiomed.2020.104089.

[21] G. Louppe, “Random Forests,” in Understanding Random Forests: From Theory to Practice, University of Liege, 2015. Accessed: Nov. 24, 2022. [Online]. Available: https://arxiv.org/pdf/1407.7502.pdf

[22] C. Zhang and Y. Ma, Ensemble Machine Learning. Boston, MA: Springer US, 2012. doi: 10.1007/978-1-4419-9326-7.

[23] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/a:1010933404324.

[24] Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H., 2009. *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

[25] J. Han, J. Pei, and H. Tong, Data Mining. Morgan Kaufmann, 2022.

[26] J. Brownlee, “C5.0: An Informal Tutorial,” Rulequest.com, 2019. https://www.rulequest.com/see5-unix.html

[27] XGBoost Developers, "*XGBoost Parameters*". Official XGBoost Documentation.

Available at: https://xgboost.readthedocs.io/en/stable/parameter.html

[28] S. Basodi, C. Ji, H. Zhang and Y. Pan, "Gradient amplification: An efficient way to train deep neural networks," in Big Data Mining and Analytics, vol. 3, no. 3, pp. 196-207, Sept. 2020, doi: 10.26599/BDMA.2020.9020004.

[29] Deep Learning (Neural Networks) (2016) Available at: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html?highlight=deep%20learning (Accessed: 27th November 2022)

[30] Provost, F., 2000, July. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets* (Vol. 68, No. 2000, pp. 1-3). AAAI Press.

[31] C. Sammut and G. I. Webb, "Accuracy," in *Encyclopedia of Machine Learning*, Boston, MA: Springer US, 2010, pp. 9–10. doi: 10.1007/978-0-387-30164-8_3.

[32] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *2010 20th International Conference on Pattern Recognition*, Aug. 2010, pp. 3121–3124. doi: 10.1109/ICPR.2010.764.

[33] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2020, doi: 10.1186/s12864-019-6413-7.

[34] Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, Sravanthi Parasa, "On evaluation metrics for medical applications of artificial intelligence," Apr. 2021, doi: 10.1101/2021.04.07.21254975.

[35] NHS Choices, "Health A-Z," NHS, 2019. https://www.nhs.uk/conditions/ (accessed Nov. 24, 2022).

[36] Xgb. Developers, "xgb.importance function - RDocumentation," www.rdocumentation.org.

https://www.rdocumentation.org/packages/xgboost/versions/0.6.4.1/topics/xgb.importance

[37] John Ross Quinlan, C4.5 programs for machine learning. San Mateo, Calif. M. Kaufmann, 1993.

Accessed: Dec. 17, 2019. [Online]. Available: https://doi.acm.org/10.1145/152181

[38]: Shane Anderson, A, and Richard F Loeser. "Why is osteoarthritis an age-related disease?." Best practice & research. Clinical rheumatology vol. 24,1 (2010): 15-26. doi:10.1016/j.berh.2009.08.006 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2818253/

[39]: CDC. "Incidence Rates of Skin Cancer in The U.S. in 2019, by Race/Ethnicity (per 100,000 Population)." Statista, Statista Inc., 10 Nov 2022, https://www.statista.com/statistics/663907/skin-cancer-incidence-rate-in-us-by-ethnicity/

[40] Bijanzadeh, M., Mahesh, P. A., & Ramachandra, N. B. (2011). An understanding of the genetic basis of asthma. The Indian journal of medical research, 134(2), 149–161. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181014/

[41] S Mendis, D Webber et al. (2005). *Avoiding heart attacks and strokes: don't be a victim-protect yourself*. World Health Organization. ISBN 92 4 154672 7

[42] National Center for Chronic Disease Prevention and Health Promotion (September 8, 2022) *Heart Disease and Stroke*. Available at: https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm#:~:text=Leading%20risk%20factors%20for%20heart,unhealthy%20diet%2C%20and%20physical%20inactivity. (Accessed 25/11/22)

[43] Lorem, G., Cook, S., Leon, D.A. *et al.* Self-reported health as a predictor of mortality: A cohort study of its relation to other health measurements and observation time. *Sci Rep* **10**, 4886 (2020). https://doi.org/10.1038/s41598-020-61603-0

[44] Burström B, Fredlund P. Self rated health: Is it as good a predictor of subsequent mortality among adults in lower as well as in higher social classes? J Epidemiol Community Health. 2001 Nov;55(11):836-40. doi: 10.1136/jech.55.11.836. PMID: 11604441; PMCID: PMC1763304.

[45] American Lung Association (November 17, 2022) *COPD Causes and Risk Factors*. Available at: https://www.lung.org/lung-health-diseases/lung-disease-lookup/copd/what-causes-copd (Accessed 25/11/22)

[46] World Health Organization M. Kokic (2022) *Chronic respiratory diseases*. Available at: https://www.who.int/health-topics/chronic-respiratory-diseases#tab=tab_1 (Accessed 25/11/22)

[47] Bansal, M., Khatri, M. and Taneja, V., 2013. Potential role of periodontal infection in respiratory diseases-a review. *Journal of medicine and life*, *6*(3), p.244.

[48] Vogeli, C., Shields, A.E., Lee, T.A., Gibson, T.B., Marder, W.D., Weiss, K.B. & Blumenthal, D. (2007) Multiple Chronic Conditions: Prevalence, Health Consequences, and Implications for Quality, Care Management, and Costs. *Journal of General Internal Medicine*. 22 (3), 391–395. doi:10.1007/s11606-007-0322-1.

[49] Menzin, J., Wygant, G., Hauch, O., Jackel, J. & Friedman, M. (2008) One-year costs of ischemic heart disease among patients with acute coronary syndromes: findings from a multi-employer claims database*. *Current Medical Research and Opinion*. 24 (2), 461–468.

[50] Tafazzoli, A., Ramsey, S.D., Shaul, A., Chavan, A., Ye, W., Kansal, A.R., Ofman, J. & Fendrick, A.M. (2022) The Potential Value-Based Price of a Multi-Cancer Early Detection Genomic Blood Test to Complement Current Single Cancer Screening in the USA. *PharmacoEconomics*. 40 (11), 1107–1117. doi:10.1007/s40273-022-01181-3.

[51] Bogers RP, Bemelmans WJE, Hoogenveen RT, et al. Association of Overweight With Increased Risk of Coronary Heart Disease Partly Independent of Blood Pressure and Cholesterol Levels: A Meta-analysis of 21 Cohort Studies Including More Than 300 000 Persons. *Arch Intern Med.* 2007;167(16):1720–1728. doi:10.1001/archinte.167.16.1720

[52] Mukherjee, A.B. and Zhang, Z., 2011. Allergic asthma: influence of genetic and environmental factors. *Journal of Biological Chemistry*, *286*(38), pp.32883-32889.

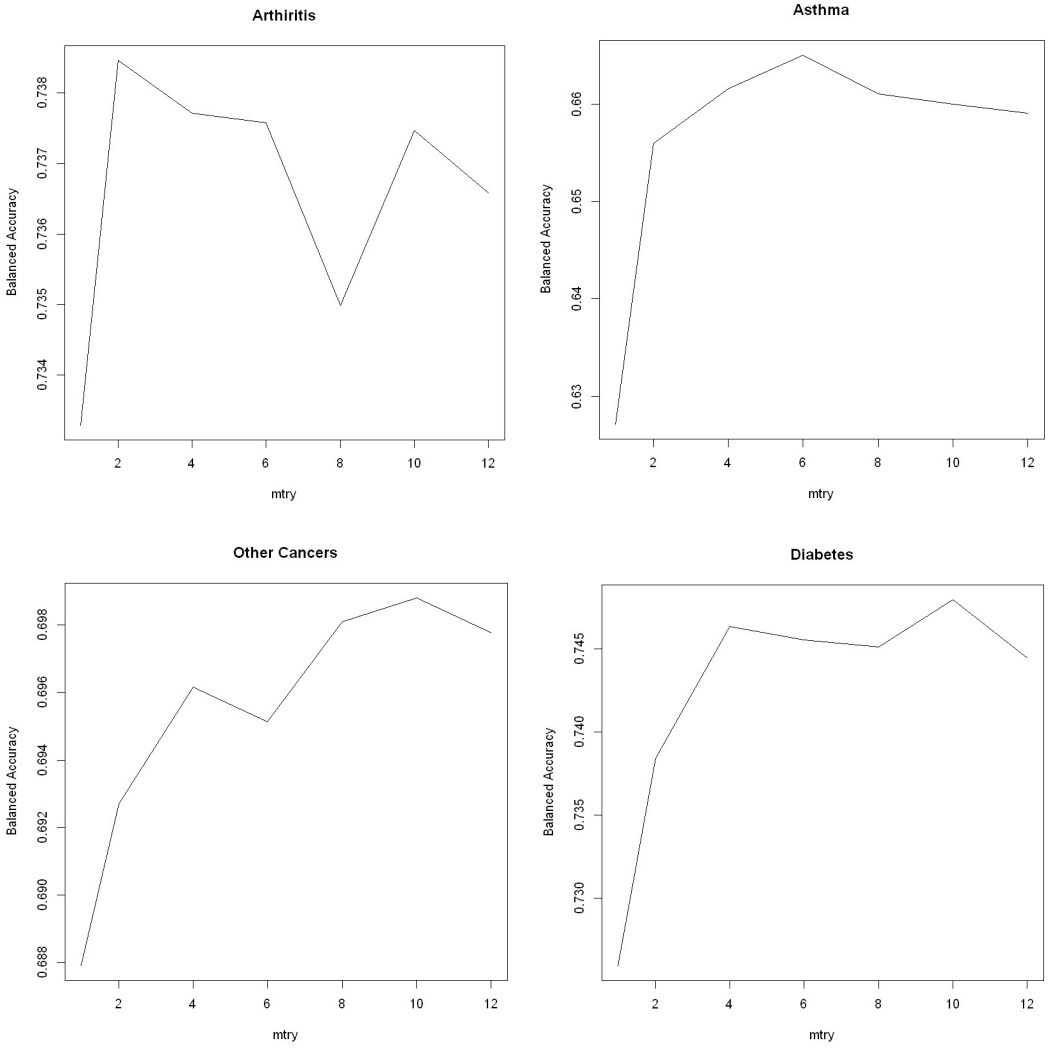
Appendix

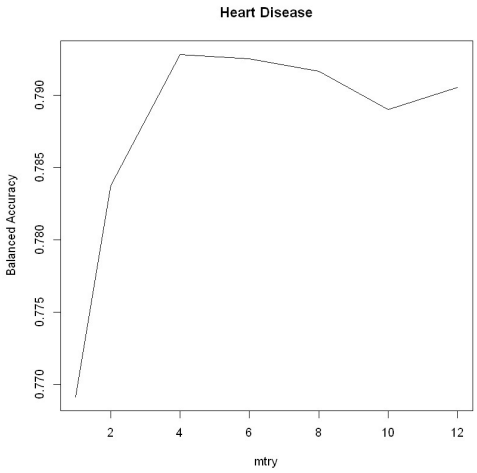
Appendix A.1 – Data Dictionary

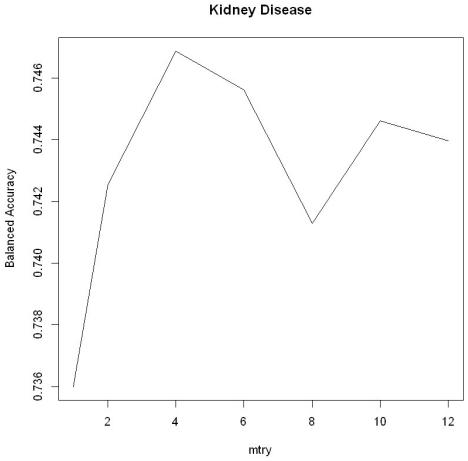
INPUT COLUMNS	EXPLANATION	RESPONSE TYPE
SEXVAR	Sex of Respondent	Discrete
GENHLTH	General respondent health in scale	Discrete
PHYSHLTH	# days during the past 30 days was physical health not good	Continuous
MENTHLTH	# days during the past 30 days was mental health not good	Continuous
HLTHPLN1	Having health care coverage	Discrete
PERSDOC2	# of personal doctors/health care providers	Discrete
MEDCOST	A time in the past 12 months that could not afford the cost of seeing doctor	Discrete
CHECKUP1	Last doctor visit for routine checkup	Discrete
EXERANY2	Participation in any physical activities during the past month	Discrete
SLEPTIM1	# hours of sleep in a 24-hour period	Continuous
CVDSTRK3	Ever had a stroke	Discrete
CVDINFR4	Ever had a heart attack	Discrete
ADDEPEV3	Ever had a heart depressive disorder	Discrete
LASTDEN4	Last dentist visit, scale of years	Discrete
RMVTETH4	# of teeth removed b.o. decay or disease	Discrete
MARITAL	Marital Status	Discrete
EDUCA	Highest grade or year of school	Discrete
RENTHOM1	Own or rent home	Discrete
CPDEMO1B	Having more than one telephone number in household	Discrete
VETERAN3	Ever served on duty in US Armed Forces	Discrete
EMPLOY1	Employment status	Discrete
CHILDREN	# of children less than 18 years	Continuous
INCOME2	Annual household income	Discrete
WEIGHT2	Weight in pounds	Continuous
DEAF	Deaf or serious difficulty hearing	Discrete
BLIND	Blind or serious difficulty seeing	Discrete
DECIDE	Because of any health condition, having serious difficulty making decisions	Discrete
DIFFWALK	Difficulty walking or climbing stairs	Discrete
DIFFDRES	Difficulty dressing or bathing	Discrete
DIFFALON	Because of any health condition, having serious difficulty doing errands alone	Discrete
SMOKE100	Ever smoked at least 100 cigarettes	Discrete
USENOW3	Current use of smokeless tobacco products	Discrete
ALCDAY5	During past 30 days, at least one drink of any alcoholic beverage	Continuous
FLUSHOT7	Ever had flu vaccine during the past 12 months	Discrete
PNEUVAC4	Ever had pneumonia shot	Discrete
FALL12MN	# of falls during the past 12 months	Continuous
SEATBELT	Frequency of use seat belts	Discrete
COLNSCPY	Ever had colonoscopy	Discrete
SIGMSCPY	Ever had sigmoidoscopy	Discrete
BLDSTOL1	Ever had blood stool test using a home kit	Discrete
STOOLDNA	Ever had blood stool test using a lab kit	Discrete
VIRCOLON	Ever had virtual colonoscopy	Discrete
HIVTST7	Ever been tested for H.I.V	Discrete
HIVRISK5	Ever exhanged money or drug for sex in the past year	Discrete
ECIGARET	Ever used an e-cigarette or other electronic vaping product	Discrete
QSTVER	Questionnaire Version Identifier	Discrete
QSTLANG	Language Identifier	Discrete
METSTAT	Metropolitan Status	Discrete
URBSTAT	Urban/Rural Status	Discrete
PRACE1	Preffered Race Category	Discrete
AGEG5YR	Fourteen-level age category	Discrete
HTM4	Heights in meters	Continuous
BMI5	Body Mass Index (BMI)	Continuous
DROCDY3_	Drink-occasions-per-day	Continuous

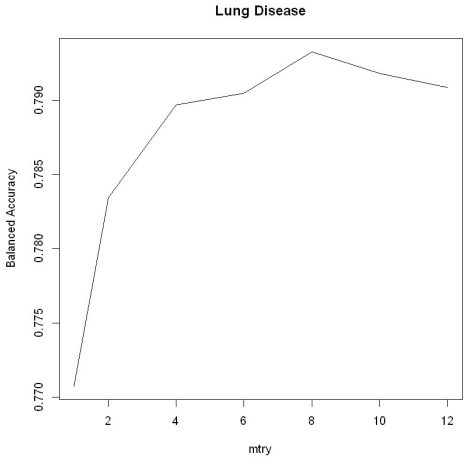
OUTPUT COLUMNS	EXPLANATION	RESPONSE TYPE
ASTHMA3	Ever had asthma	Discrete
CHCSCNCR	Ever had skin cancer	Discrete
CHCCOPD2	Ever had chronic obstructive pulmonary disease	Discrete
CHCKDNY2	Ever had kidney disease	Discrete
CVDCRHD4	Ever had coronary heart disease	Discrete
CHCOCNCR	Ever had any other types of cancer	Discrete
HAVARTH4	Ever had some form of arthritis	Discrete
DIABETE4	Ever had diabetes	Discrete

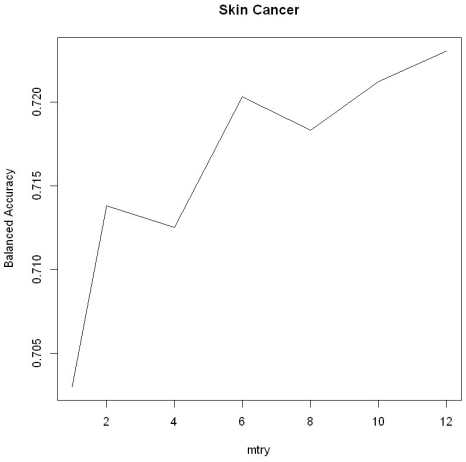
Appendix A.2.1 - Random Forest mtry hyperparameter tuning



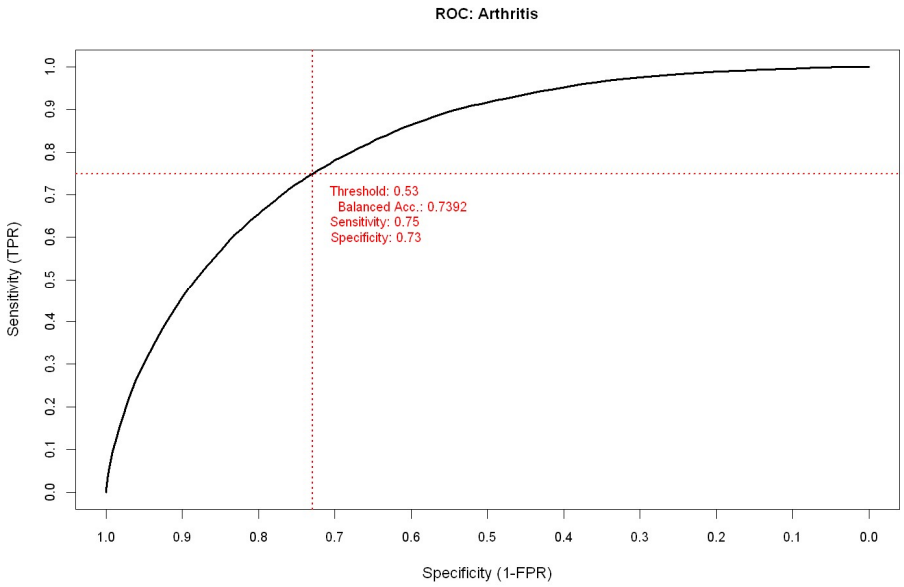

Heart Disease
Balanced Accuracy
0.770
0.775
0.780
0.785
0.790
2
4
6
8
10
12
mtry

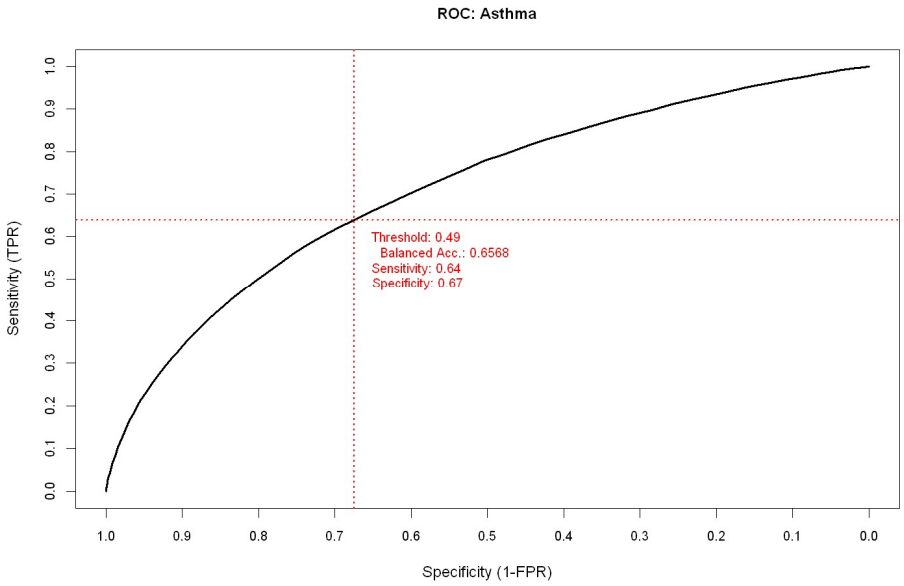

Kidney Disease
Balanced Accuracy
0.736
0.738
0.740
0.742
0.744
0.746
2
4
6
8
10
12
mtry

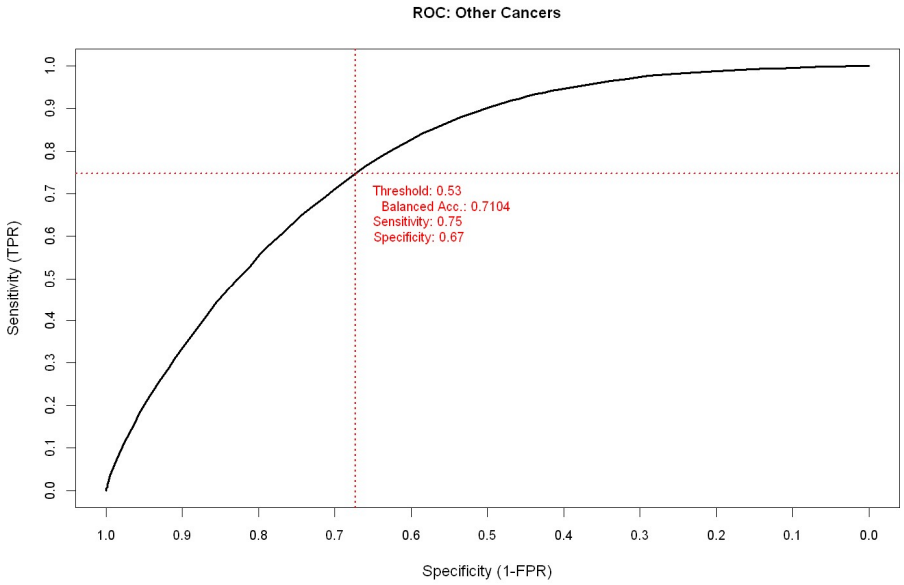

Lung Disease
Balanced Accuracy
0.770
0.775
0.780
0.785
0.790
2
4
6
8
10
12
mtry

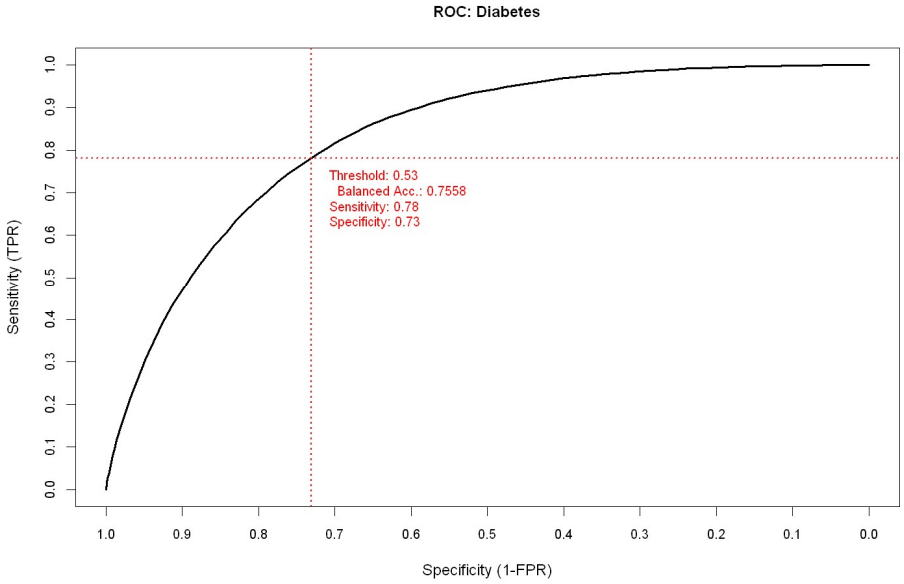

Skin Cancer
Balanced Accuracy
0.705
0.710
0.715
0.720
2
4
6
8
10
12
mtry

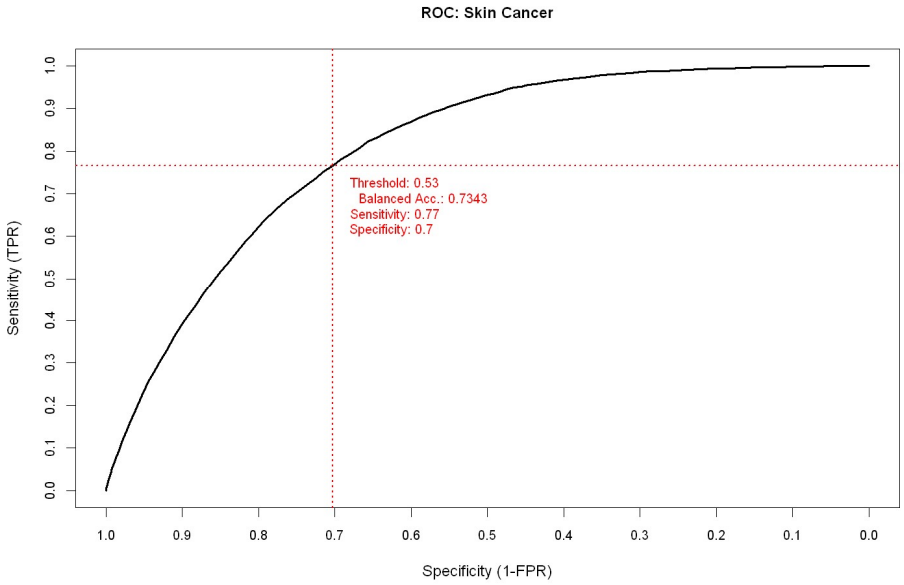
A.2.2 - Random Forest ROC graphs

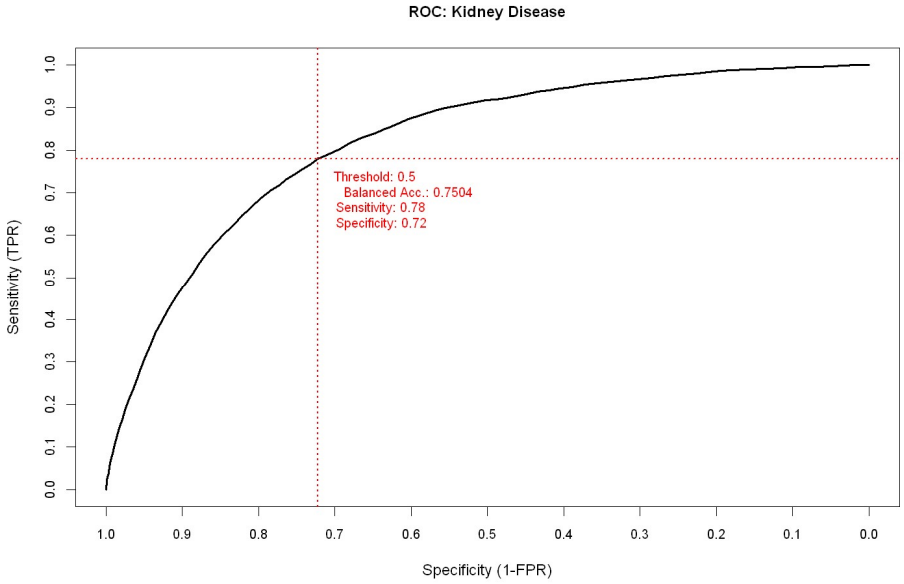


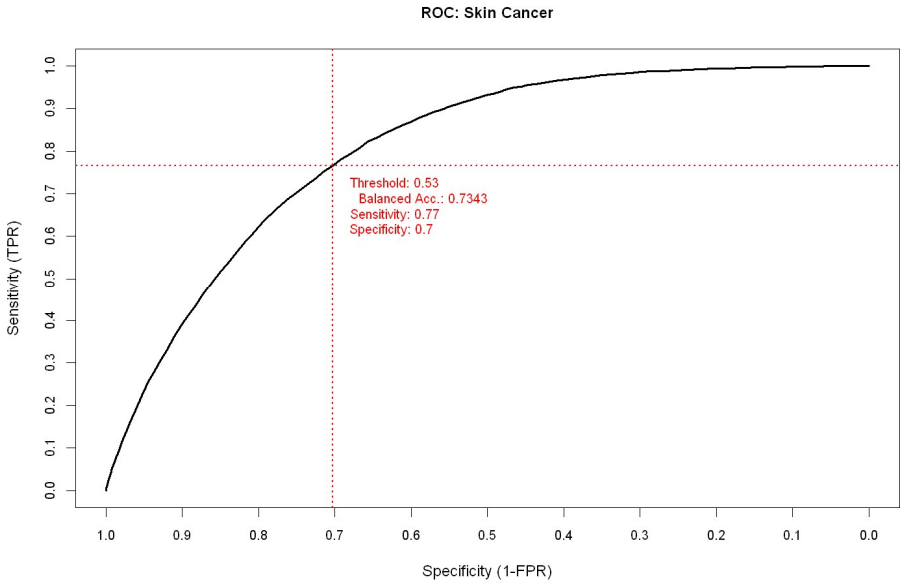


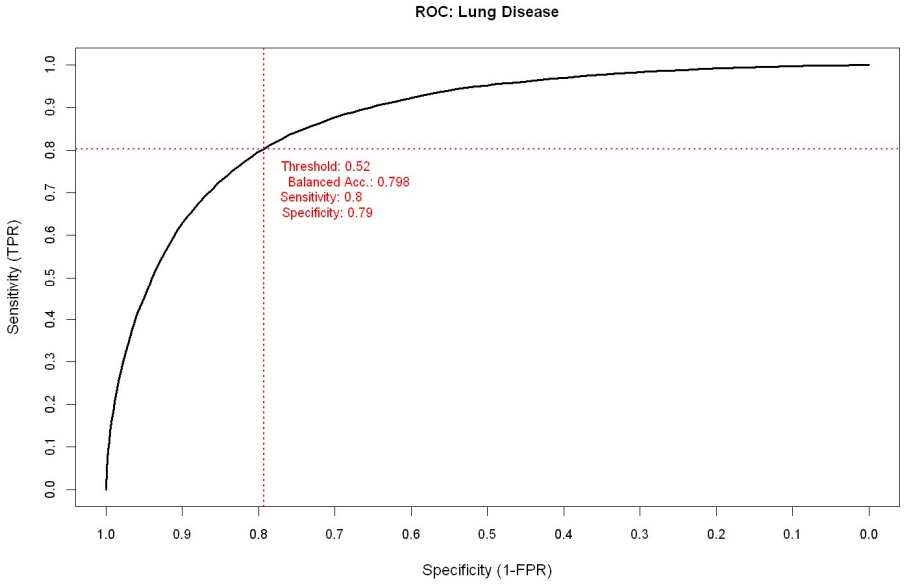

ROC: Other Cancers
Threshold: 0.53
Balanced Acc.: 0.7104
Sensitivity: 0.75
Specificity: 0.67
Sensitivity (TPR)
Specificity (1-FPR)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0


ROC: Diabetes
Threshold: 0.53
Balanced Acc.: 0.7558
Sensitivity: 0.78
Specificity: 0.73
Sensitivity (TPR)
Specificity (1-FPR)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

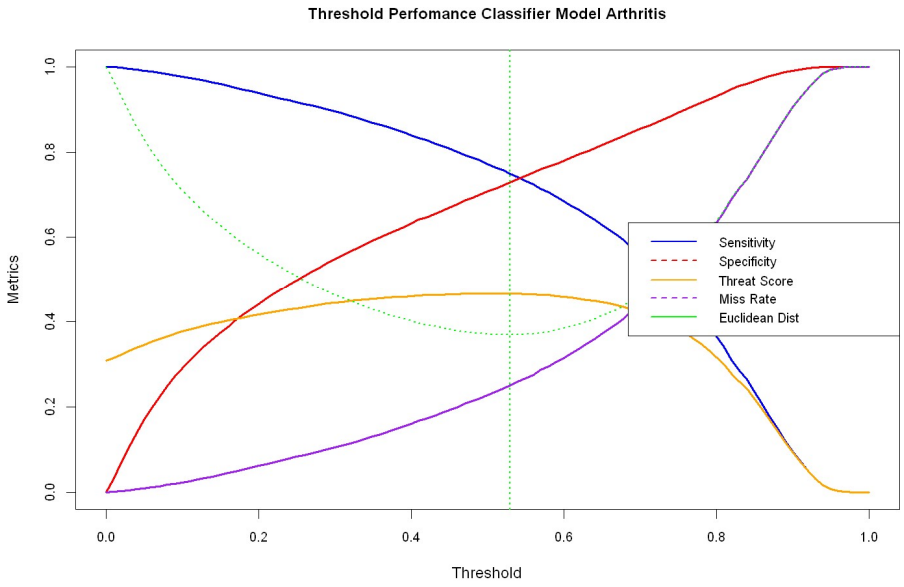

ROC: Skin Cancer
Sensitivity (TPR)
Specificity (1-FPR)
Threshold: 0.53
Balanced Acc.: 0.7343
Sensitivity: 0.77
Specificity: 0.7

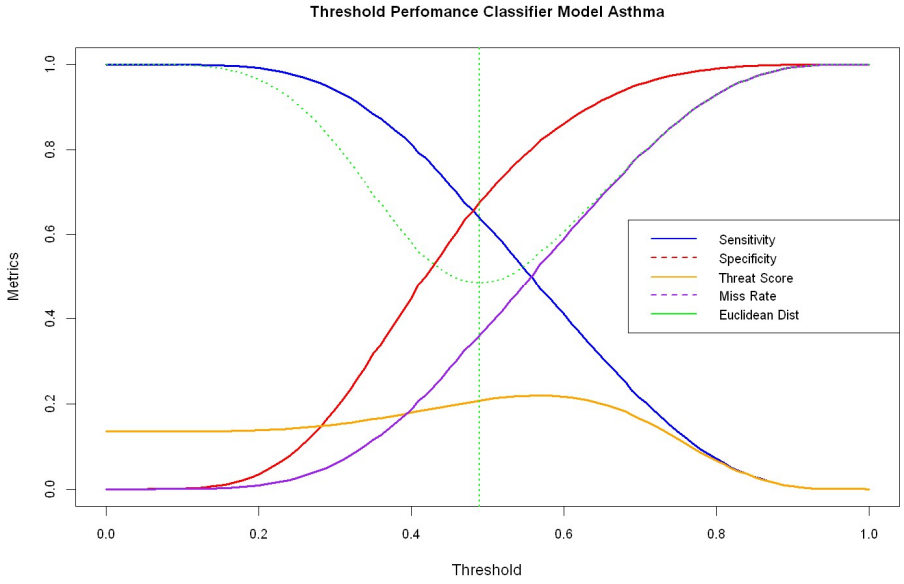

ROC: Kidney Disease
Sensitivity (TPR)
Specificity (1-FPR)
Threshold: 0.5
Balanced Acc.: 0.7504
Sensitivity: 0.78
Specificity: 0.72

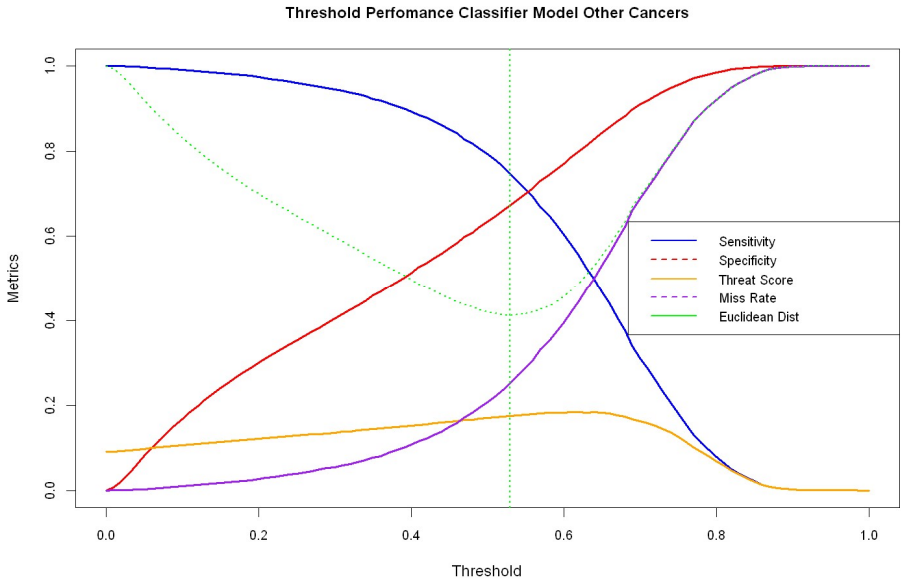

ROC: Skin Cancer
Sensitivity (TPR)
Specificity (1-FPR)
Threshold: 0.53
Balanced Acc.: 0.7343
Sensitivity: 0.77
Specificity: 0.7

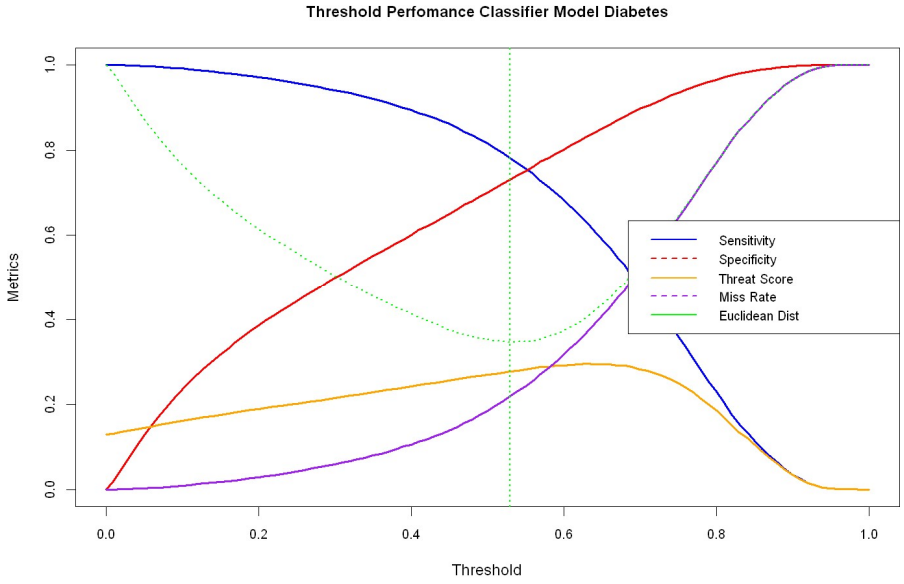

ROC: Lung Disease
Sensitivity (TPR)
Specificity (1-FPR)
Threshold: 0.52
Balanced Acc.: 0.798
Sensitivity: 0.8
Specificity: 0.79

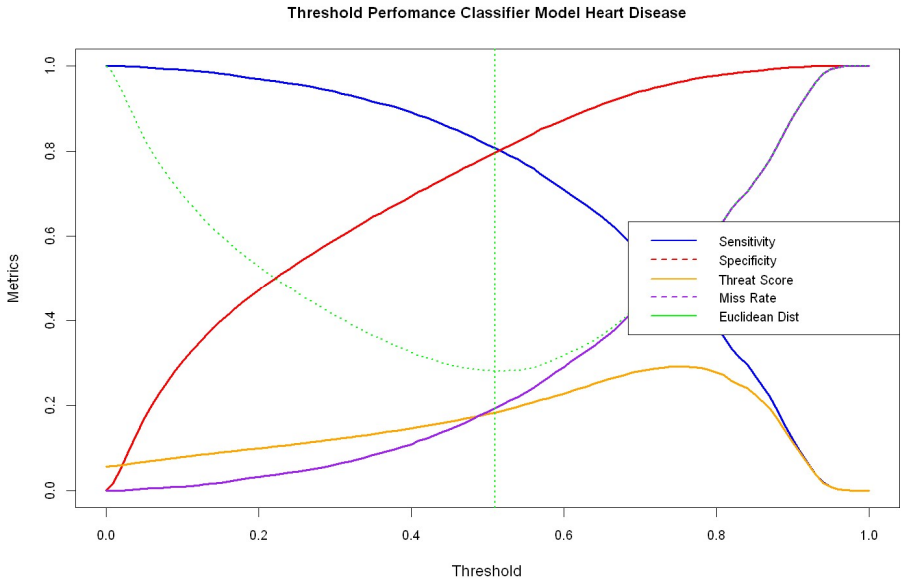
A.2.2- Random Forest Threshold graphs

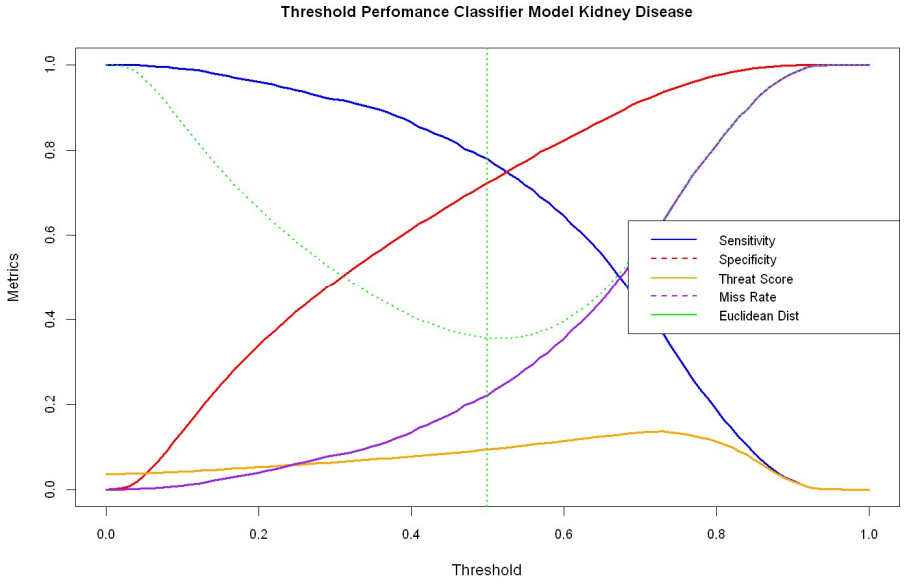


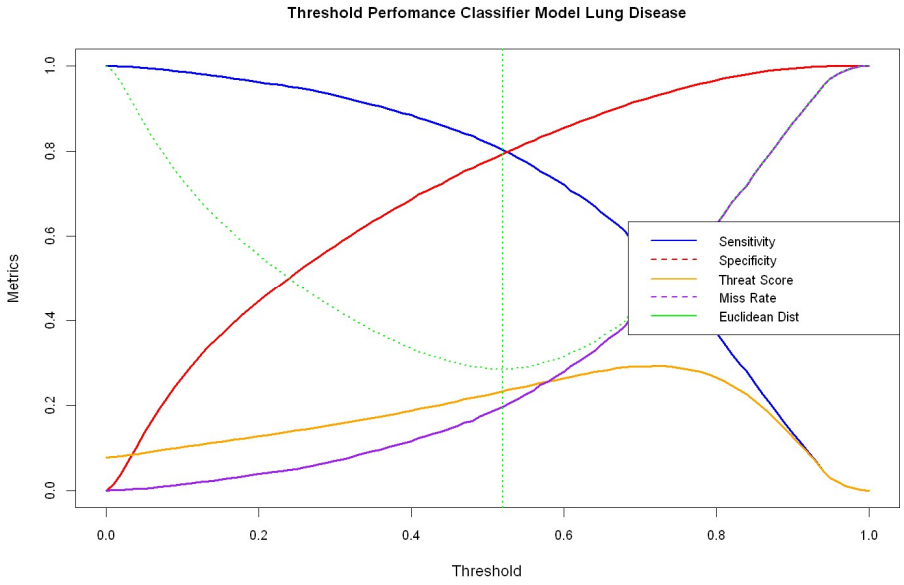



Threshold Perfomance Classifier Model Other Cancers
Metrics
Threshold
Sensitivity
Specificity
Threat Score
Miss Rate
Euclidean Dist
0.0
0.2
0.4
0.6
0.8
1.0

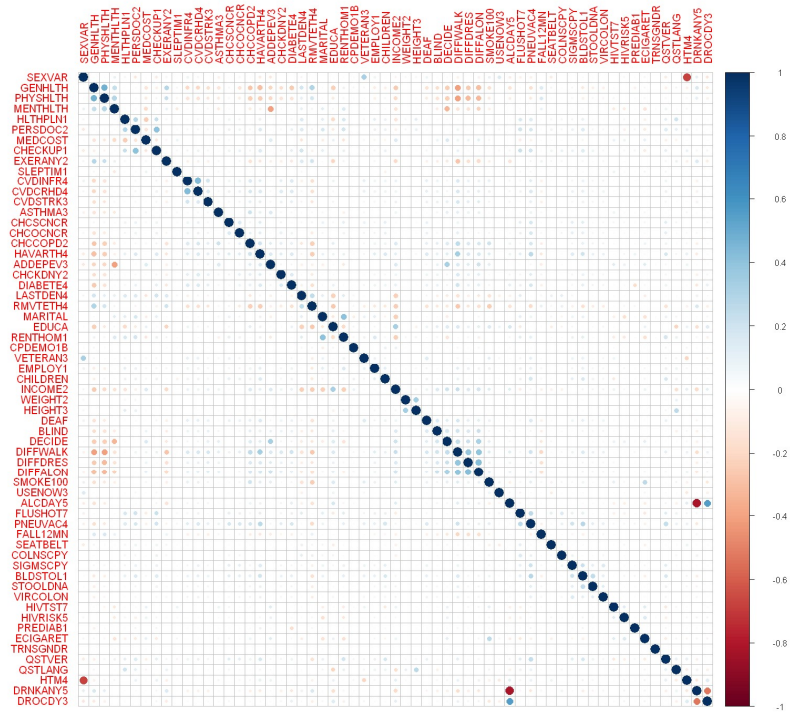

Threshold Perfomance Classifier Model Diabetes
Metrics
Threshold
Sensitivity
Specificity
Threat Score
Miss Rate
Euclidean Dist
0.0
0.2
0.4
0.6
0.8
1.0

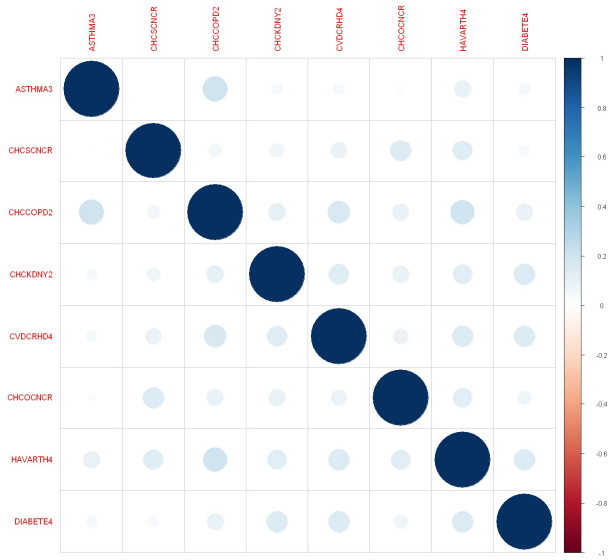

Threshold Perfomance Classifier Model Heart Disease
Metrics
Threshold
0.0
0.2
0.4
0.6
0.8
1.0
Sensitivity
Specificity
Threat Score
Miss Rate
Euclidean Dist

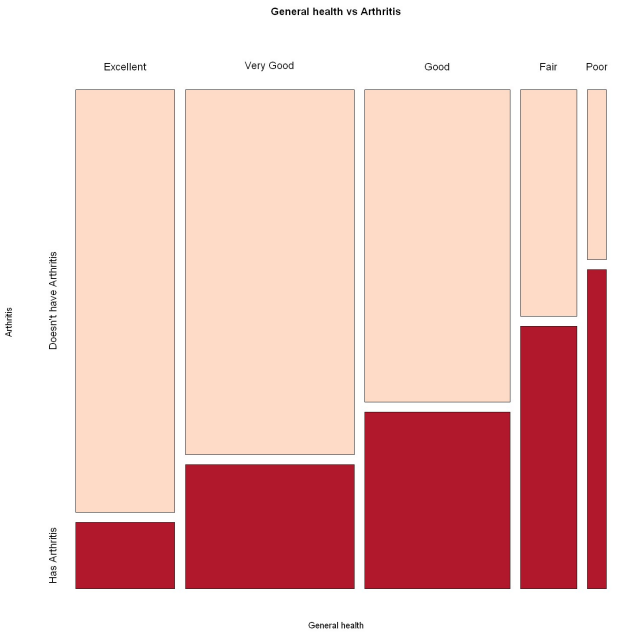

Threshold Perfomance Classifier Model Kidney Disease
Metrics
Threshold
0.0
0.2
0.4
0.6
0.8
1.0
Sensitivity
Specificity
Threat Score
Miss Rate
Euclidean Dist

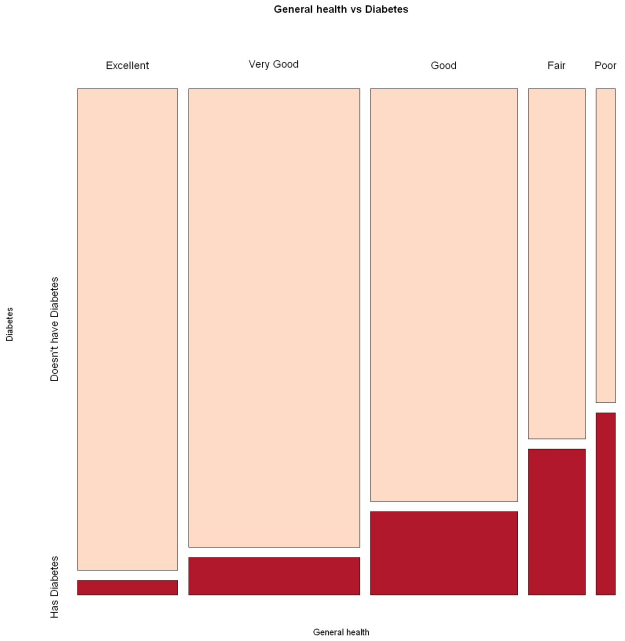

Threshold Perfomance Classifier Model Lung Disease
Metrics
Threshold
0.0
0.2
0.4
0.6
0.8
1.0
Sensitivity
Specificity
Threat Score
Miss Rate
Euclidean Dist

APPENDIX A.3

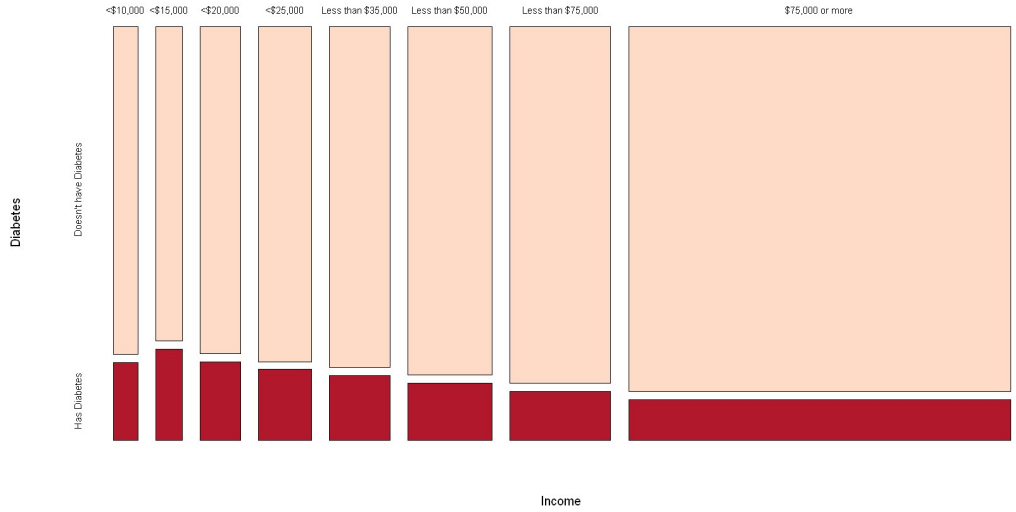

SEXVAR
GENHLTH
PHYSHLTH
MENTHLTH
HLTHPLN1
PERSDOC2
MEDCOST
CHECKUP1
EXERANY2
SLEPTIM1
CVDINFR4
CVDCRHD4
CVDSTRK3
ASTHMA3
CHCSCNCR
CHCOCNCR
CHCCOPD2
HAVARTH4
ADDEPEV3
CHCKDNY2
DIABETE4
LASTDEN4
RMVTETH4
MARITAL
EDUCA
RENTHOM1
CPDEMO1B
VETERAN3
EMPLOY1
CHILDREN
INCOME2
WEIGHT2
HEIGHT3
DEAF
BLIND
DECIDE
DIFFWALK
DIFFDRES
DIFFALON
SMOKE100
USENOW3
ALCDAY5
FLUSHOT7
PNEUVAC4
FALL12MN
SEATBELT
COLNSCPY
SIGMSCPY
BLDSTOL1
STOOLDNA
VIRCOLON
HIVTST7
HIVRISK5
PREDIAB1
ECIGARET
TRNSGNDR
QSTVER
QSTLANG
HTM4
DRNKANY5
DROCDY3
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


ASTHMA3
CHCSCNCR
CHCCOPD2
CHCKDNY2
CVDCRHD4
CHCOCNCR
HAVARTH4
DIABETE4
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

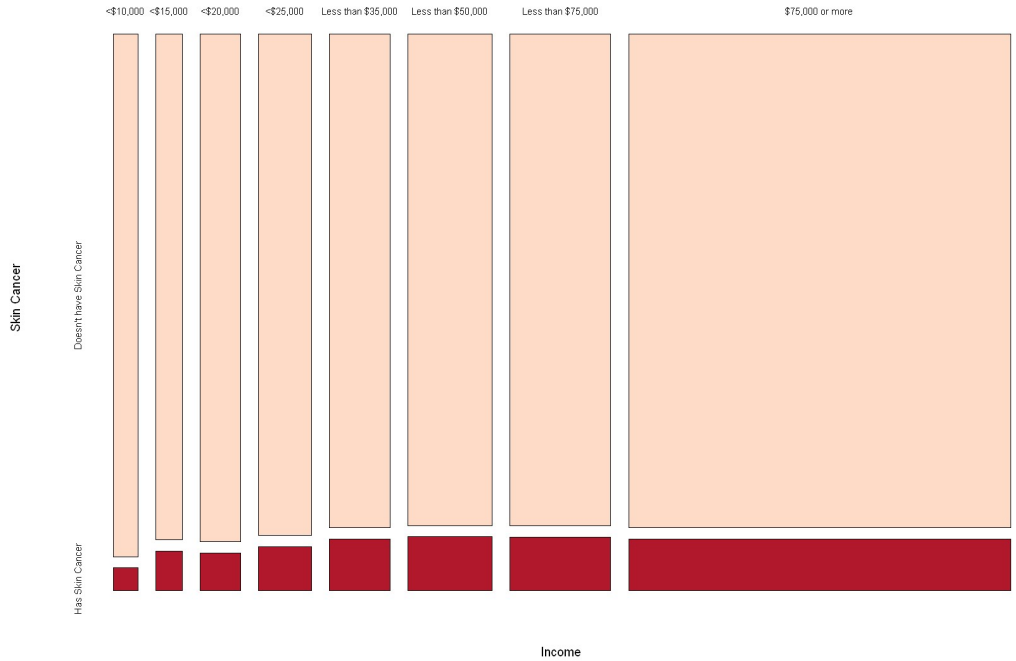

General health vs Arthritis
Excellent
Very Good
Good
Fair
Poor
Arthritis
Doesn't have Arthritis
Has Arthritis
General health


General health vs Diabetes
Excellent
Very Good
Good
Fair
Poor
Diabetes
Doesn't have Diabetes
Has Diabetes
General health

Income vs Diabetes

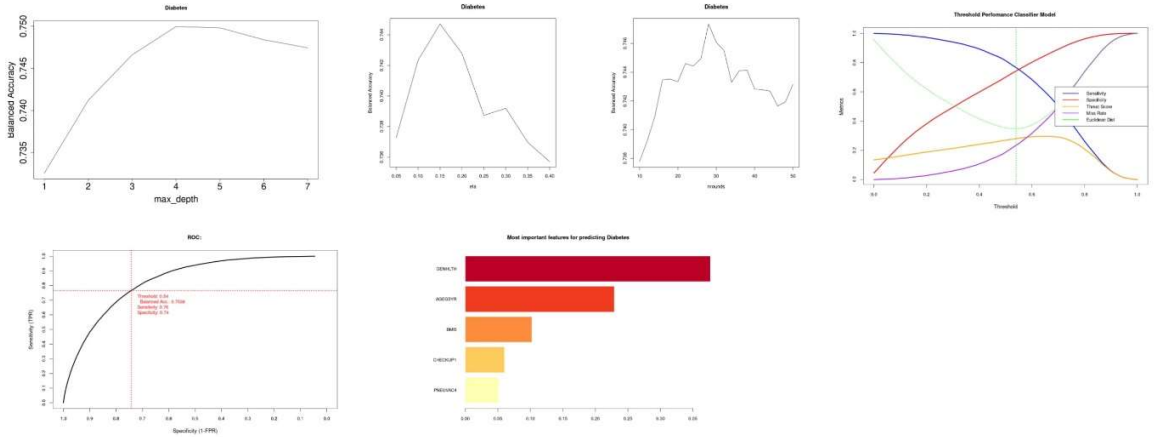


Income vs Skin Cancer

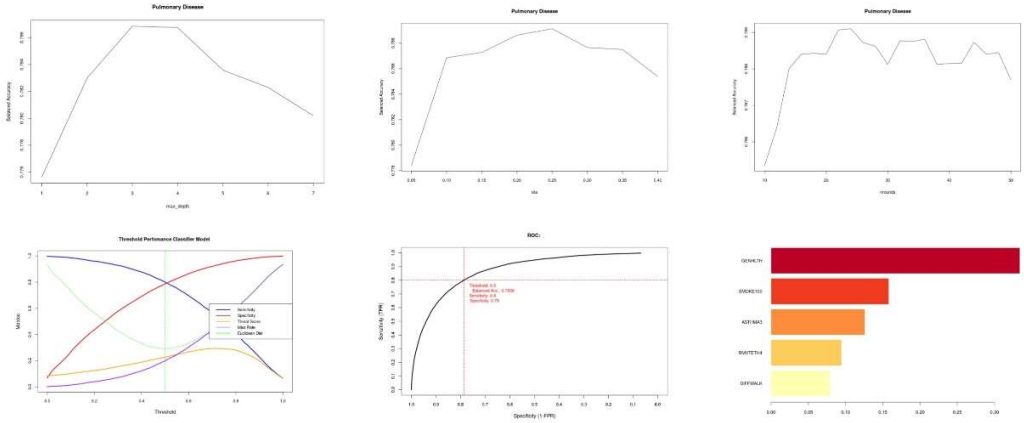


APPENDIX A.4

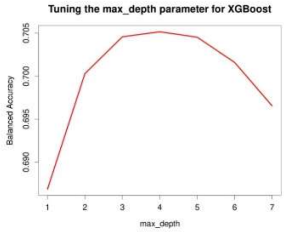
Hyper-parameter Tuning for XGBoost: Diabetes

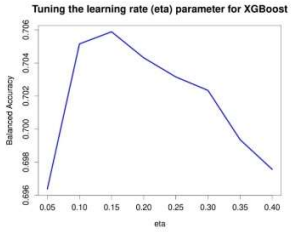


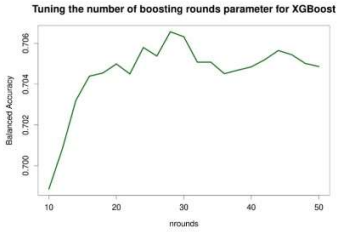
Hyper-parameter Tuning for XGBoost: Lung Disease

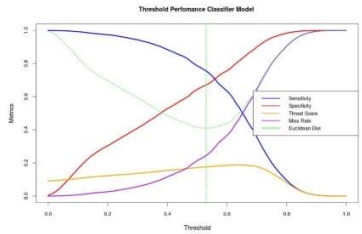


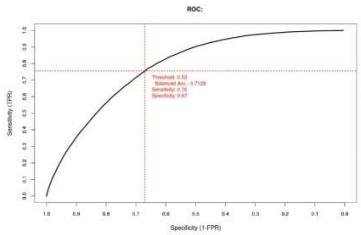
Hyper-parameter Tuning for XGBoost: Other Cancers

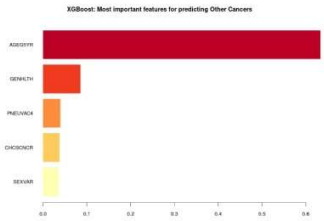




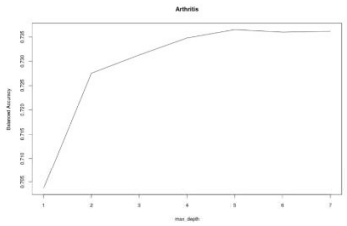


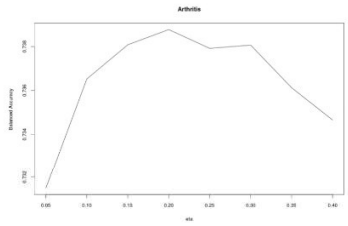


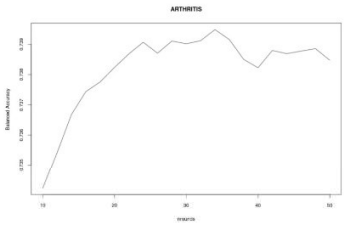


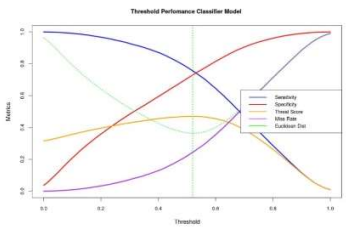


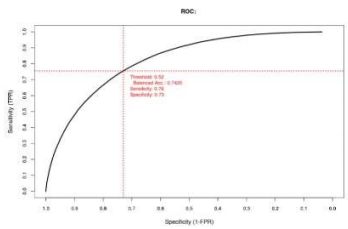
Hyper-parameter Tuning for XGBoost: Arthritis

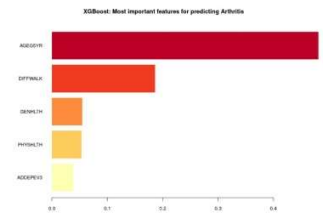




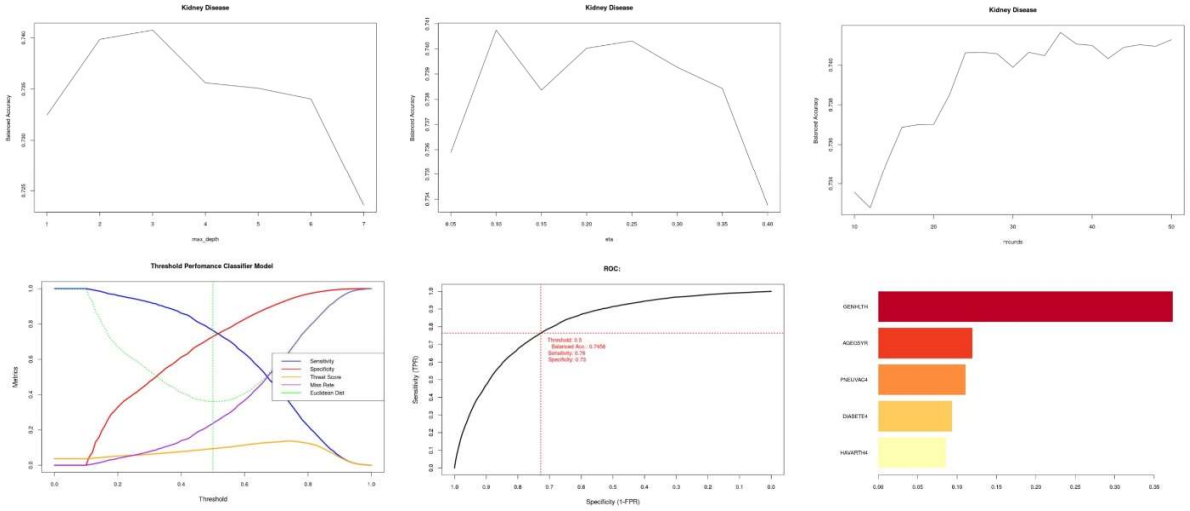




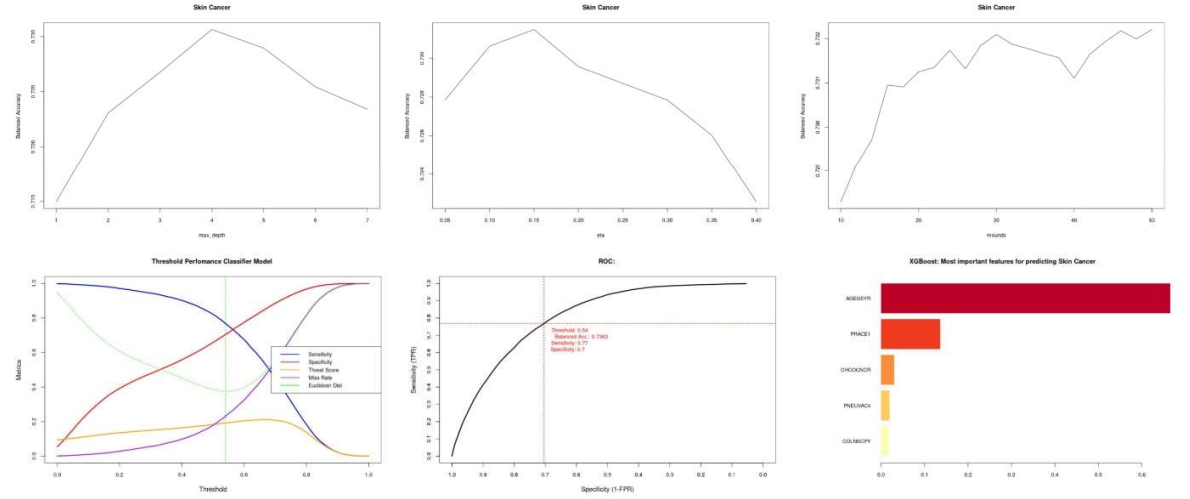




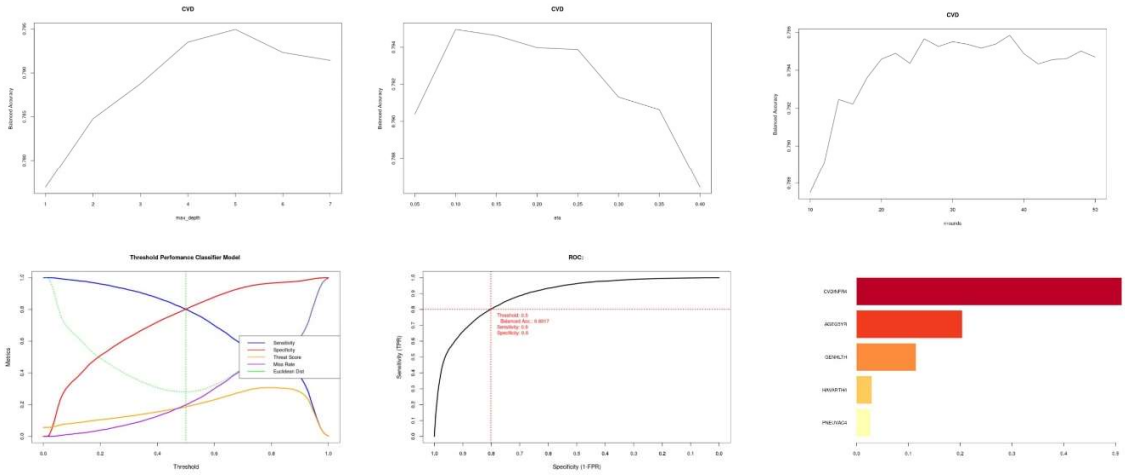
Hyper-parameter Tuning for XGBoost: Kidney Disease



Hyper-parameter Tuning for XGBoost: Skin Cancer



Hyper-parameter Tuning for XGBoost: CVD



Hyper-parameter Tuning for XGBoost: Asthma