ST494 Final Project- Thoracic Surgery

Cole Hanniwell – 180320780

Einstein Oyewole - 180517070

Table of Contents

Executive Summary

Our project was focused on the ability to predict whether a patient would die within a year after having thoracic surgery. We wanted to be able to predict whether a patient would die as well as find the most important variables for the predictions. Our dataset contained many categorical variables and the majority of those being dummy variables with values 1 or 0 (True or False) and a few numerical variables. Our dataset was quite small in terms of number of observations (470) and the majority of those also having the response variable be of the value survived after surgery. We wanted to increase the number of observations and the minority class in the response variable (died after surgery) so we used a technique called SMOTE. SMOTE allowed us to create observations of the minority class by synthesizing from the already known minority class observations. Our dataset also had problems with outliers in a numeric variable as well as scaling the dataset for PCA. One more problem we had was choosing which type of error we would prefer to minimize for the model (false positives or false negatives). We decided that both errors were extremely bad and that both should be minimized so we just tried to achieve the model with the highest accuracy therefore minimizing both types of errors. After outlining the goals and fixing the problems with the dataset, we split our dataset into 70% training data and 30% test data and then began testing models. We did use cross validation for some of the models to ensure they were robust.

The first model type we tried to use on the data was clustering. We used hierarchical clustering with single, complete and average as well as top-down clustering. All the clustering methods were unsuccessful producing misclassification rates greater than 30% for all methods. We then attempted to use LDA and QDA to try and split the data. LDA and QDA both performed better than clustering but the results were still not good producing misclassification rates greater than 25%. We then attempted to use linear and radial SVM. The linear SVM did not perform great but did perform better than LDA so we concluded the data was not linear. However, the radial SVM performed excellent with a test misclassification rate of just 10.5%. We then attempted to use bagging method random forest on the data which performed just okay. The random forest had a good training misclassification rate of 7% but was overfitting as the test misclassification rate was about 20%. We then created a new dataset where we transformed the data onto its first PCA components. Using the PCA data we ran all the same models from the original data set. The models from the PCA data set all performed just slightly worse than the original dataset achieved. We concluded by choosing the radial SVM model from the original dataset. We then looked at the weights of the PCA loading as well as the importance chart from the random forest models. From both it was obvious there was 4 variables that were significantly more important than the others. These 4 variables were Age, FEV1, FVC and the size of the tumor.

Problem

Our data set is about Thoracic Surgery and the results after the operation. Thoracic surgery is surgery on organs in the chest area (Heart, Lungs, Esophagus). Some examples of Thoracic Surgery are lung cancer surgeries, heart transplants and anti-reflux surgeries. After having thoracic surgery, there are many common complications. Some of these are atelectasis (complete or partial collapse of the entire lung), pulmonary oedema (excess fluid in the lungs), atrial fibrillation (quivering or irregular heartbeat), haemorrhage (escape of blood from a ruptured blood vessel), wound infection, persistent air leak, pneumonia, and respiratory failure. The main risk factor that leads to early mortality post surgery is the Age of the patient. The older you get the less likely you survive post surgery. Other risk factors are current smoking, underlying carcinoma, and chronic lung disease. The data can be found at this site: https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data#.

Data Description

Categorical Variables (Distribution of each variable can be seen in Appendix A)

- **Diagnosis:** Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any. This is a nominal variable with levels (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1). We assigned each diagnosis in our data set to the number at the end of the diagnosis (DGN1 = 1).
- **Performance Status:** Uses the Zubrod Scale. In medicine, performance status is an attempt to quantify cancer patients' general well-being and activities of daily life. Also known as ECOG or WHO score. This is an ordinal variable with 3 levels PRZ2, PRZ1 and PRZ0. These are sequential so in our data set we have their values at 2, 1, 0. A site with more detail on the Zubrod scale: https://en.wikipedia.org/wiki/Performance_status.
- **Pain before Surgery:** A dummy variable to determine if the patient was in pain right before the surgery occurred. 1 for in pain and 0 for not in pain.
- **Haemoptysis before Surgery:** A dummy variable for whether the patient coughs blood before surgery. 1 for coughing blood and 0 for not.
- **Dyspnea before Surgery:** A dummy variable that declares if the patient had difficult or laboured breathing before surgery. 1 for difficult or laboured breathing and 0 for not.
- **Cough before Surgery:** A dummy variable for if the patient was coughing before the surgery. 1 for coughing before surgery and 0 for not.
- **Weakness before Surgery:** A dummy variable to determine if the patient felt weak before the surgery occurred. 1 for patient feeling weak and 0 for not.
- **Type 2 DM (Diabetes Mellitus):** Diabetes mellitus is a disease that prevents your body from properly using the energy from the food you eat. In more detail, the pancreas makes insulin, but it either doesn't produce enough, or the insulin doesn't work properly. This is a dummy variable where 1 indicates the patient has diabetes melllitus and 0 indicates they do not.

- **Tumor Size:** The TNM system is the most widely used cancer staging system. Most hospitals and medical centers use the TNM system as their main method for cancer reporting. This variable corresponds to T in TNM system which refers to the primary tumor. This is an ordinal variable which refers to the size of the main tumor. The levels for this variable are (OC11, OC14, OC12, OC13) where OC11 is the smallest and OC14 is the largest. In our data they are assigned 1, 2, 3, 4 from smallest to largest. A site giving more detail on tumor sizes and their class: https://www.cancer.gov/about-cancer/diagnosis-staging/staging.
- **MI (myocardial infarction) up to 6 months:** A dummy variable that indicates whether they have had a heart attack in the last 6 months. 1 indicates the patient has had a heart attack in the last 6 months and 0 indicates the patient has not.
- **Peripheral Arterial diseases:** PAD is a common circulatory problem in which narrowed arteries reduce blood flow to your limbs. This is a dummy variable where 1 indicates the patient has PAD and 0 indicates they do not.
- **Smoking:** This is a dummy variable which indicates if the patient was a smoker. 1 indicates that the patient was and 0 indicates the were not.
- **Asthma:** Asthma is a respiratory condition in which your airways narrow and swell and may produce extra mucus (difficulties in breathing). This is a dummy variable where 1 indicates the patient has asthma and 0 indicates they do not.

Numerical Variables (Distribution of each variable can be seen in Appendix B)

- **Age:** The age of the client at the time of the surgery. This is a numerical value with a mean of 62.5 and a standard deviation of 8.7.
- **FVC (Forced Vital Capacity):** The amount of air that can be forcibly exhaled from your lungs after taking the deepest breath possible, as measured by spirometry. This is a numeric variable with mean 3.28 and standard deviation 0.87.
- **FEV1 (Forced Expiratory Volume):** Volume that has been exhaled at the end of the first second of forced expiration. In lay mans terms, it measures how much air a person can exhale during a forced breath. This is a numeric variable with mean 11.77 and standard deviation 4.57. (Including the outliers mentioned later).

Target Variable

- **Risk of Death after 1 year:** The target variable contains 400 non deaths (class 0/False) and 70 deaths (class 1/True).

Project Goals

- Our first goal of this project is to be able to accurately predict if a patient will die after surgery.
- Another goal after being able to accurately predict the response variable is to explore the model we use and find the most impactful variables. We want to be able to determine which factors are the most important in determining the response variable.

Results

Dataset Problem

Our dataset has a large number of predictors (22) relative to the number of observations (470). Another problem with our dataset is that the distribution of our response variable (Death after 1 year of surgery) is very unbalanced. About 14% of the observations are True and the remaining 86% are False. We used two solutions to solve this problem: SMOTE and extra evaluation metrics for the models.

With our dataset being this unbalanced we decided to use the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset. Using SMOTE we are able to oversample the minority class by creating new observations that are synthesized from the already existing observations. With our oversampled class we are able to get the response variable to have about 50% of the observations in each class. The oversampled dataset will allow us to use the algorithms like normal and evaluate like normal.

To be sure that after using SMOTE our models are still accurate on the original dataset, we decided to use other evaluation metrics. If we used accuracy or misclassification rate it would be quite easy to get good scores due to the imbalance in our dataset. If everything were to be predicted class 0 then the accuracy would be 86% and misclassification rate 14%. We decided to use other metrics as well when testing our models to confirm that this was not the case. The evaluation metrics we used were accuracy/misclassification rate, precision, recall and the F1 score.

Outlier Problem

The FEV1 variable contained many outliers which were significantly affecting the scaling of our variables as well as some models. There were 14 significant outliers in the FEV1 variable which typically ranged from 0-20. The outliers were all greater than 40 so we elected to remove them from the data set entirely. Since we are using SMOTE to create a dataset that is 50% both classes, we did not mind removing the outliers as only 1 outlier belonged to the minority class. Removing from the larger class allowed the SMOTE to still be able to create new observations from many datapoints and limit the number of duplicates.

Type 1 and Type 2 Error Problem

The type 1 error or a false positive would implicate that the model predicts the patient will die from the surgery when they will not. The type 2 error or a false negative would implicate that the model predicts the patient will not pass away from the surgery when in fact they will. We were debating about limiting which error we attempt to remove from the model. Both errors are extremely bad and we were unable to decide which one was worse. This led to us attempting to just have the highest accuracy possible to lower the total amount of error instead of increasing one by minimizing the other.

Train/Validate/Test Set Decision

With our dataset having roughly 732 observations after we applied SMOTE to the minority class. We decided to not use a validation set but rather create a stratified test and train set. The training set included 70% of the data and the test set contained the remaining 30%. For some models we elected to use cross validation to make sure the results were accurate and robust.

Scaling Problem

We attempted to scale the variables for PCA when trying to perform analysis on the PCA data. The scaling however made the PCA capture significantly less variance in the beginning components and created even worse results once that data was tested on the models. We think our model was not able to scale well for PCA due to the number of dummy variables in our model.

PCA Method

We decided to still perform PCA analysis on the data with the unscaled variables. Using the unscaled variables our first component was able to capture approximately 96% of the variance in the dataset. After transforming the data with PCA we then ran models on both the original dataset as well as the PCA transformed dataset. We used the loadings to investigate the variables that held significant weights for the PCA. We saw that AGE accounts for most of component 1 as it has the largest weight. For component 2 we saw large weights for FVC and FEV1 with weights 0.76 and 0.66 respectively. The size of tumor also had a large weight and combining these four variables and components they captured 98.8% of variance in the data. The loading graph, bar chart of the components and their variance captured as well as the plot of the data on the first two components can be seen in Appendix C.

Clustering Results

We began by attempting to cluster the result on the original dataset. Our dataset containing all the dummy variables we thought that there would be some obvious clusters in the data that our model would be able to find. We attempted to use hierarchical clustering to find the clusters in the data. We used single, complete and average methods to perform this clustering. All the clusters were unsuccessful in achieving good results. The single linkage results on the original dataset:

	Training
Misclassification rate	48.2%
F1 Score	68.3%
Precision	99.6%
Recall	51.9%

We then attempted the top-down approach which also performed poor as we expected like the other clustering methods. We then ran the same clustering methods on the PCA transformed dataset. The PCA transformed clustering also had poor results which were similar to the original dataset for all the clustering methods.

LDA and QDA Results

Our next model attempted was to use LDA to split the data. We hoped the data would be linear with the numerous amounts of dummy variables. We first ran LDA on the original data and the results were similar to clustering.

	Training	Testing
Misclassification rate	28.3%	33.3%
F1 Score	72.7%	68.9%
Precision	72.7%	67.5%
Recall	72.9%	70.4%

We then tried to use QDA too, these results were also poor and quite similar to the LDA results. We then ran the LDA and QDA on the PCA transformed data which led to worse results than the LDA and QDA on the original dataset.

SVM Results

We used linear SVM to start to determine if the data was linear or not. The linear SVM did not perform great but it did perform better than LDA. We tuned the linear SVM by changing the cost for the constraint violation value for each iteration, the cost values were in the range 0.01 to 10. We used the value of 5.84 and had better results than the previous methods. The results were better but not great which allowed us to conclude the data was not linear and to not try logistic regression. We also did not try the linear SVM on the PCA transformed data due to the poor results on the original. We then tried to use SVM with the radial kernel on the original dataset. We also tuned this model with a constraint violation value of 11.43 and gamma value of 1. The radial kernel performed exceptionally well on the training data and the testing data.

	Training	Testing
Misclassification rate	0.3%	10.5%
F1 Score	99.6%	90.6%
Precision	99.3%	93.3%
Recall	100%	88%

We then ran the radial kernel SVM on the PCA transformed data.

	Training	Testing
Misclassification rate	0.78%	14.6%
F1 Score	99.3%	86.7%
Precision	98.9%	91.2%
Recall	99.6%	82.5%

The original data set performed better on both the training and testing set and was an option for our final model.

Bagging/Random Forest Results

We chose to use the bagging method random forest to try to predict the response variable. With the high number of dummy variables, we believed that random forest would be able to have a high accuracy predicting the response variable. We believed that this would also lead to better performance than a boosting method on the dataset. We began by running the algorithm on the original dataset. After running the algorithm on the original dataset, the algorithm was overfitting with a 20% difference between test and train misclassification rates. We then tuned the model to reduce the complexity and hopefully solve the overfitting problem. To reduce the complexity, we tried different combinations of the hyper parameters. The hyperparameters we tuned were the number of trees to grow, the number of variables randomly sampled, the size of sample to draw, the minimum size of the terminal nodes and the maximum number of terminal nodes. The results after tuning all the hyper parameters on our original dataset were.

	Training	Testing
Misclassification rate	7%	19.6%
F1 Score	93.3%	82.8%
Precision	95.9%	86.7%
Recall	90.8%	79.4%

We then tried to run the random forest on the PCA transformed data.

	Training	Testing
Misclassification rate	14.8%	27.1%
F1 Score	85.4%	74%
Precision	83.5%	70.5%
Recall	87.5%	77.8%

The original dataset also did better here on the training and testing compared to the PCA transformed model. The two models appear to be overfitting the data. The misclassification rate for both is about 13% higher on the testing data set than the training data set.

Important Variables

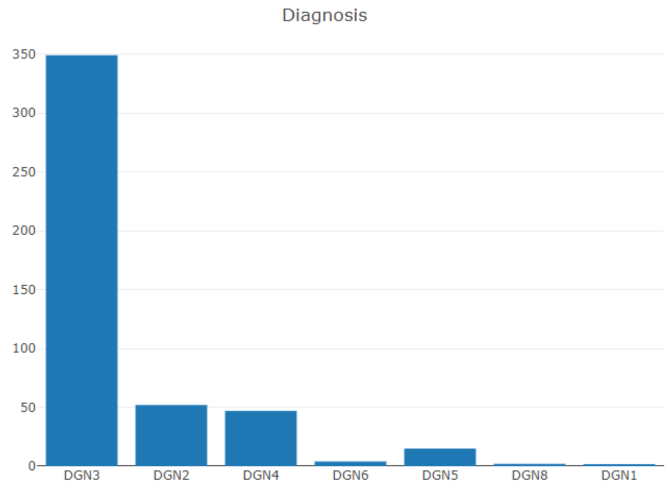
After testing the models and their algorithms on both the original and PCA transformed data and getting an accurate prediction model we wanted to find the important variables in prediction. We noticed from both the PCA transformation that only the Age, FVC, FEV1 and the size of tumor had large weights. We also noticed that from bagging the same four variables were the top four in the importance of the forests, this can be seen in Appendix D. We hypothesized that age would be a very important factor in our analysis for the models given its importance in other surgeries and overall health. The results from the random forest variable importance and the PCA loadings suggest that this is in fact true. Looking at the age from the original dataset patients who died after surgery had a larger average age then those who did not die. It is clear why the Forced Vital Capacity and Forced Expiratory Volume variables are also important in the prediction. They are strong indicators of the lung strength of the patients and determine how much impact the surgery can have on their lung's performance. The lower the FVC and FEV1 is the less strength the patient's lungs have and this aligns with our data as the patients who die have lower average values than those who do not. The last important variable the size of tumor is also clear why it is so important. With the distribution of the variable having the majority in size 1 or 2 and very few in size 3 or 4. Inspecting the data the size 3 or 4 tumor patients have a much higher percent chance of dying with a 57% chance compared to the size 1 or 2 tumor patients with a 15% chance of dying.

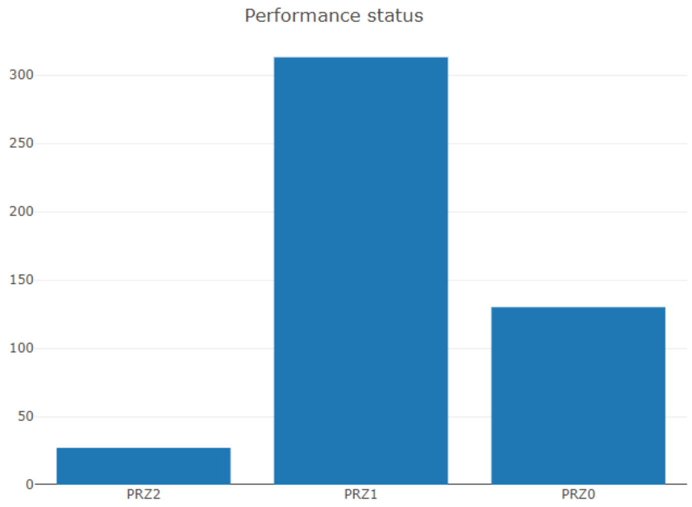
Final Model Selection

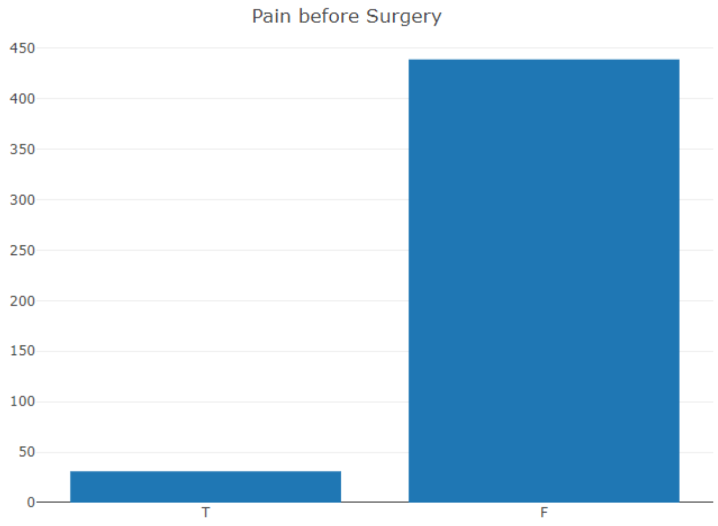
We elected to use the radial SVM model from the original data set as our final model. This model had the best misclassification rate out of any model we tried for both the training and test data. It also had outstanding F1, precision and recall scores on the test data of 90.6%, 93.3% and 88% respectively. We believe that this is an excellent model for predicting the death of a patient after surgery. Having maximized the accuracy of the model to limit the total number of type 1 and 2 errors mentioned earlier.

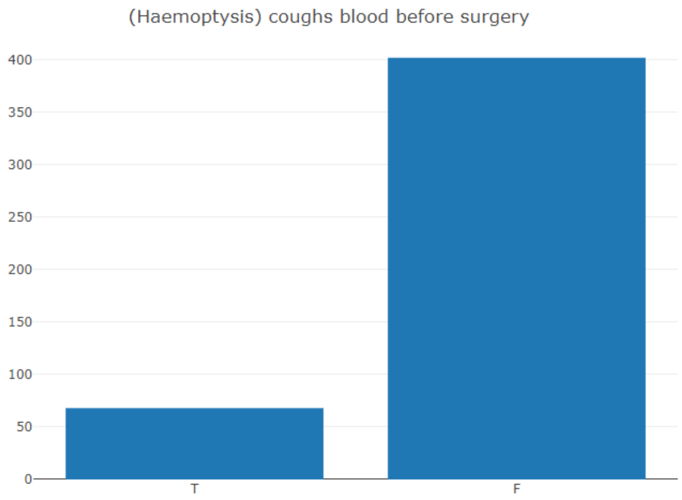
Appendices

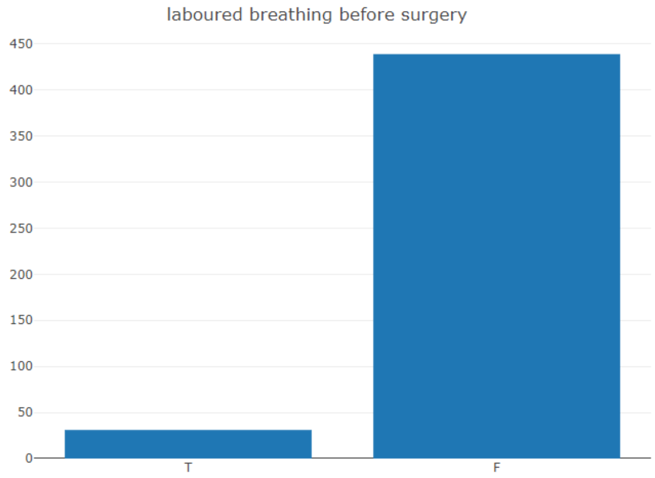
Appendix A

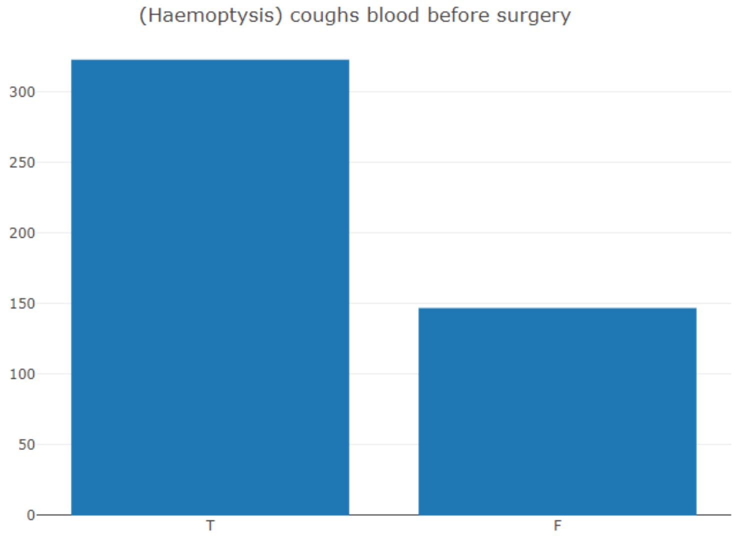

Diagnosis

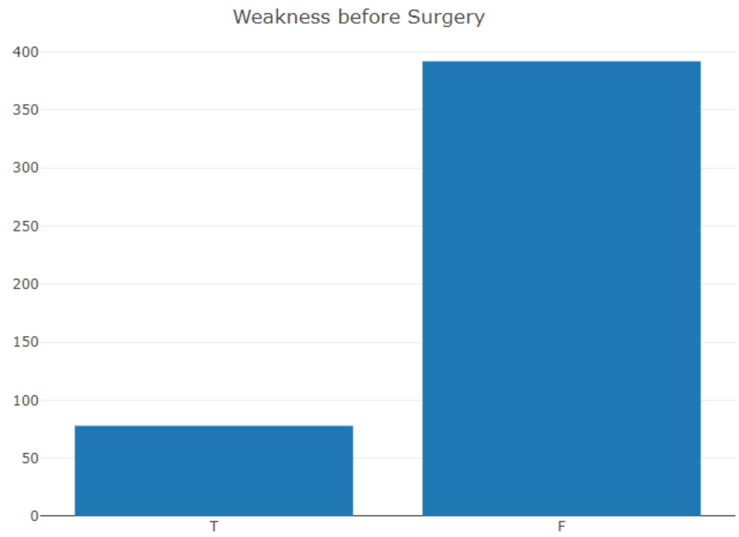

Performance status

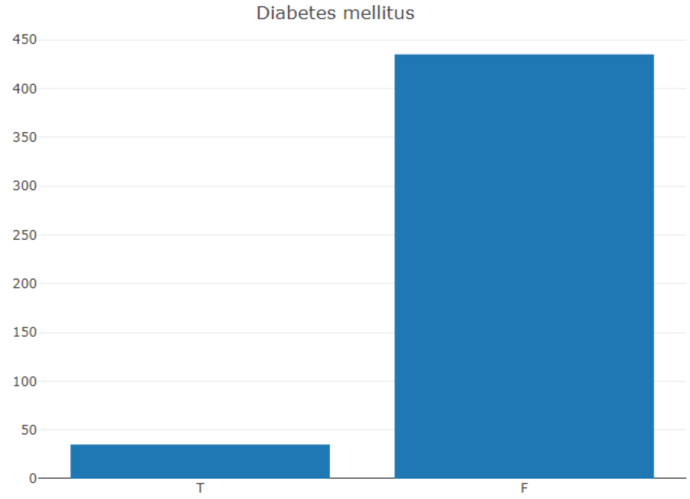

Pain before Surgery

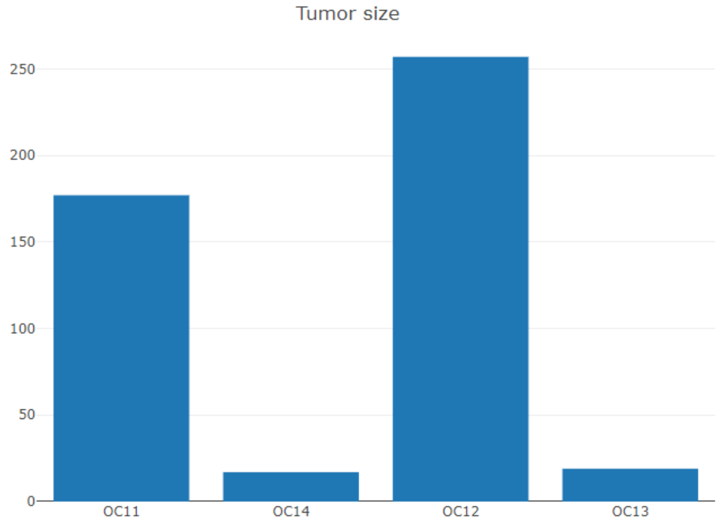

(Haemoptysis) coughs blood before surgery
400
350
300
250
200
150
100
50
0
T
F

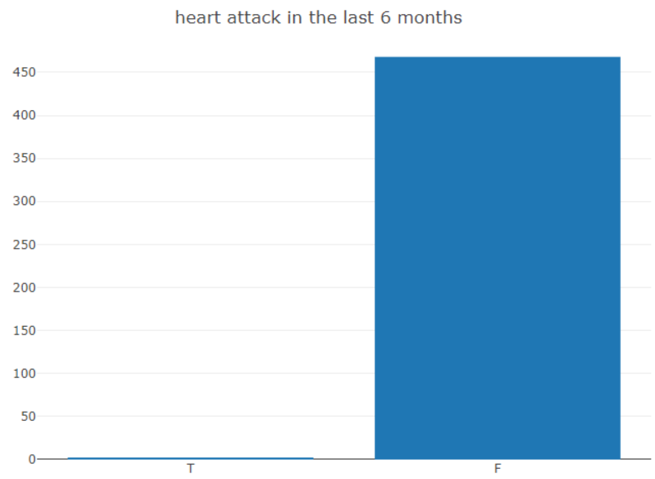

laboured breathing before surgery
450
400
350
300
250
200
150
100
50
0
T
F

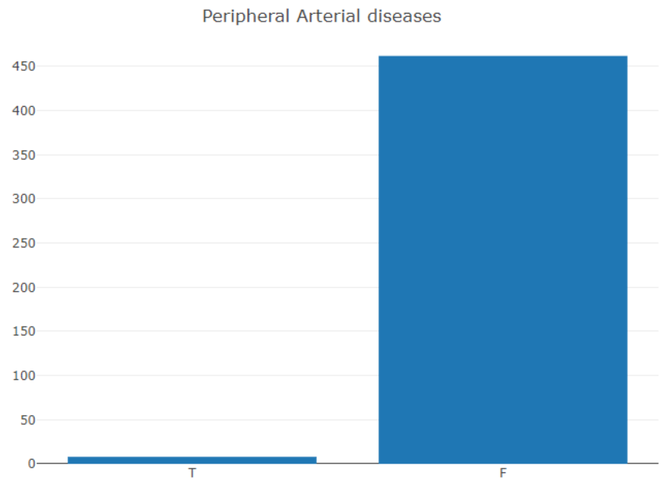

(Haemoptysis) coughs blood before surgery
300
250
200
150
100
50
0
T
F

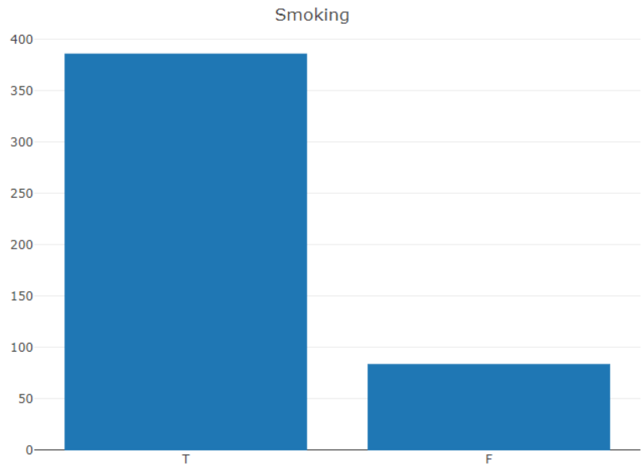

Weakness before Surgery

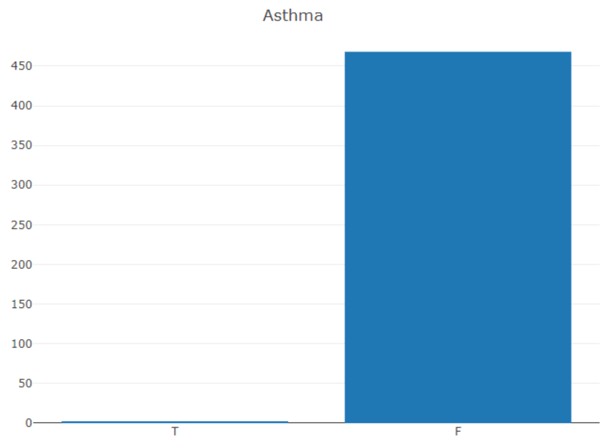

Diabetes mellitus


Tumor size


heart attack in the last 6 months

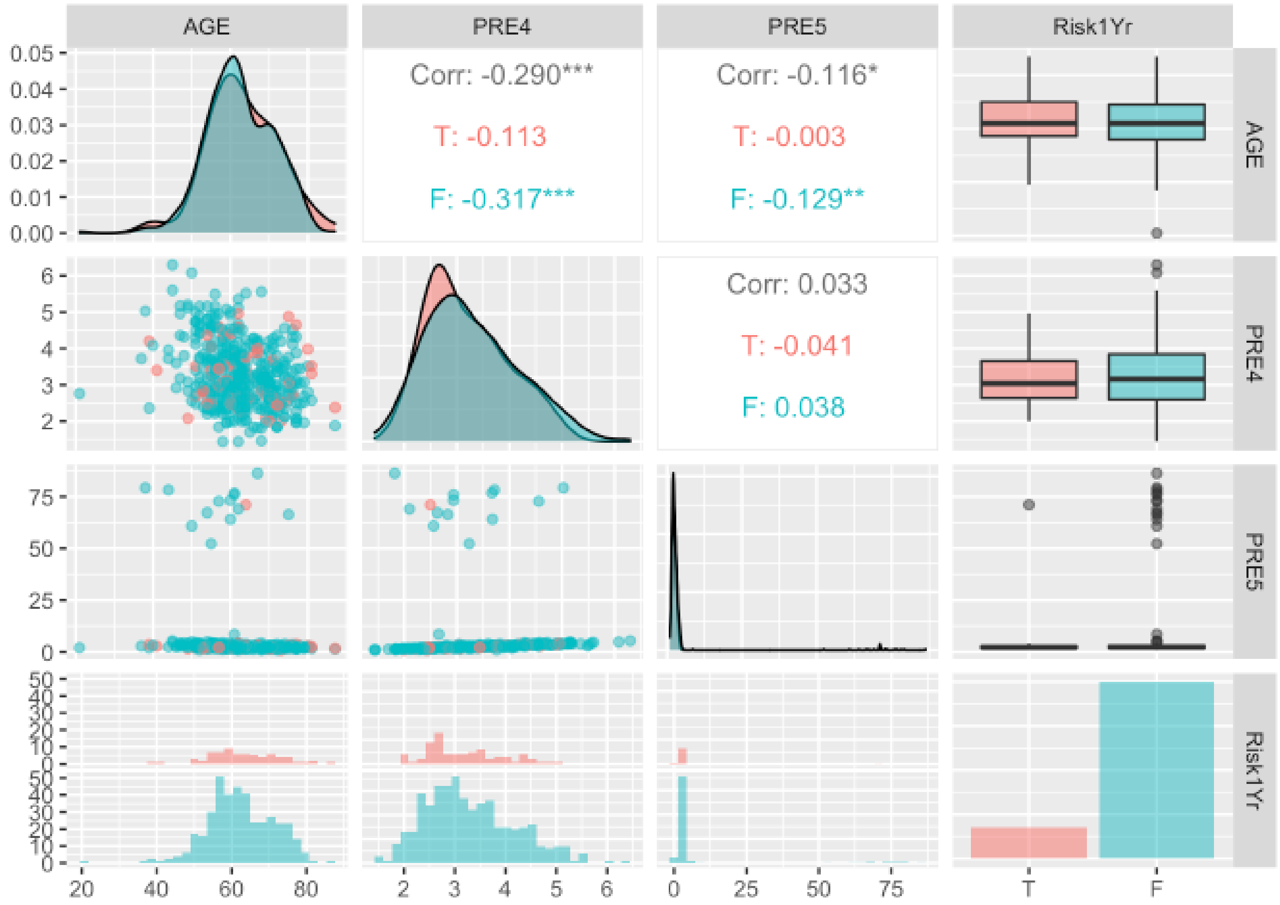

Peripheral Arterial diseases


Smoking

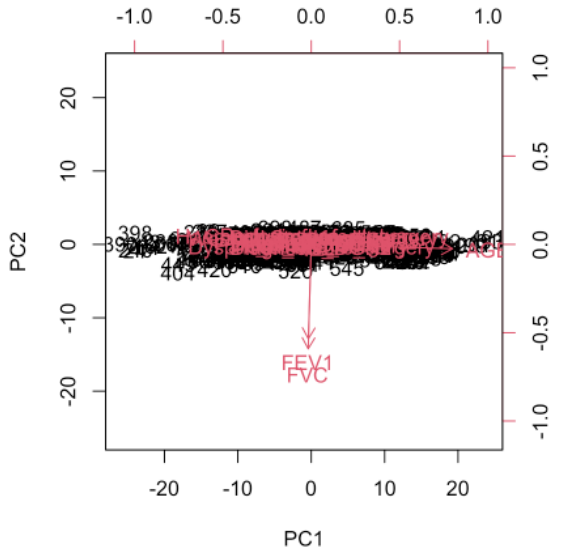


Appendix B

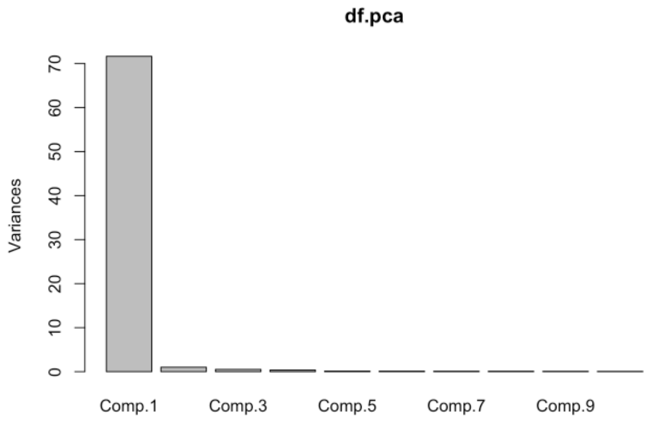
Where PRE4 = FVC and PRE5 = FEV1



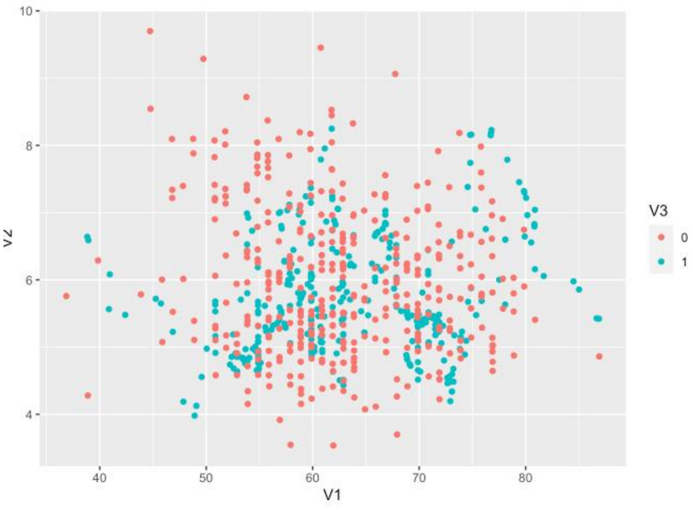
Appendix C



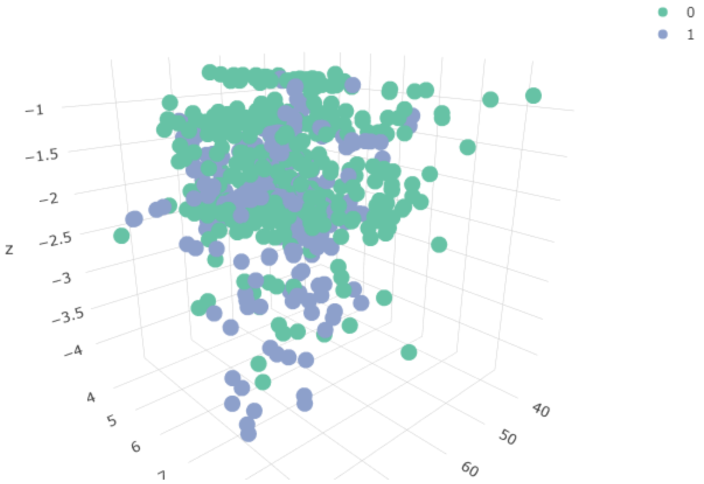
-1.0
-0.5
0.0
0.5
1.0
PC2
PC1
-20
-10
0
10
20
398
404
420
545
520
FEV1
FVC



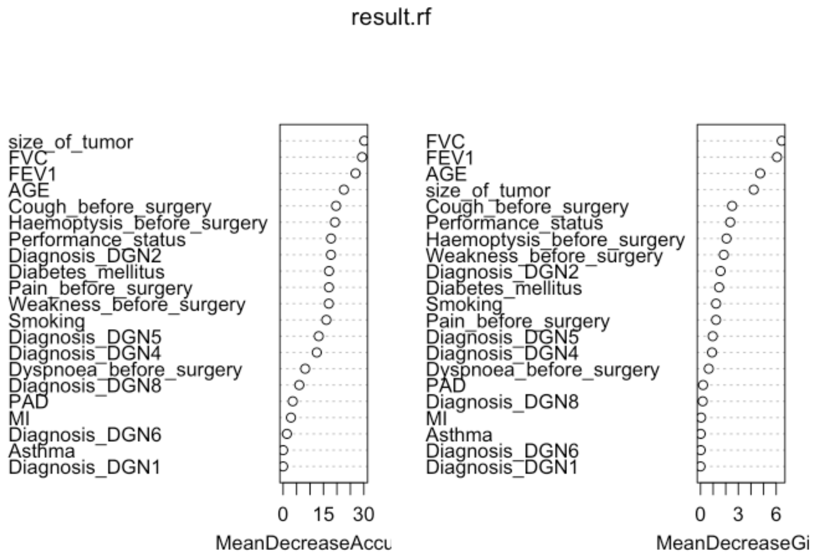
df.pca
Variances
0
10
20
30
40
50
60
70
Comp.1
Comp.3
Comp.5
Comp.7
Comp.9



V2
V1
4
6
8
10
40
50
60
70
80
V3
0
1


0
1
z
−1
−1.5
−2
−2.5
−3
−3.5
−4
4
5
6
7
40
50
60

Appendix D


result.rf
size_of_tumor
FVC
FEV1
AGE
Cough_before_surgery
Haemoptysis_before_surgery
Performance_status
Diagnosis_DGN2
Diabetes_mellitus
Pain_before_surgery
Weakness_before_surgery
Smoking
Diagnosis_DGN5
Diagnosis_DGN4
Dyspnoea_before_surgery
Diagnosis_DGN8
PAD
MI
Diagnosis_DGN6
Asthma
Diagnosis_DGN1
0
15
30
MeanDecreaseAccu
FVC
FEV1
AGE
size_of_tumor
Cough_before_surgery
Performance_status
Haemoptysis_before_surgery
Weakness_before_surgery
Diagnosis_DGN2
Diabetes_mellitus
Smoking
Pain_before_surgery
Diagnosis_DGN5
Diagnosis_DGN4
Dyspnoea_before_surgery
PAD
Diagnosis_DGN8
MI
Asthma
Diagnosis_DGN6
Diagnosis_DGN1
0
3
6
MeanDecreaseGi