ASSIGNMENT 2

Data Modelling and Presentation

MINH HOANG NGUYEN – S3712611
QUACH PHUOC NHAT – S3580007

Executive Summary

The goal of this report is to develop a model which can successfully predict the decision of a donor on whether he/she would donate their blood in March 2007.
The dataset that was used contained 5 columns and 748 rows entries, the dataset provides that there were 570 “no” entries and 178 “yes” entries for the blood donations in March 2007.
Every attribute was examined and observed before the model was developed using the <.corr> function as well as multiple figures including histogram, bar chart, density, and boxplots.
The final model was decided based on its performance, including the analysis of overall accuracy, confusion matrix and classification report.
Each of the model was split with 3 different training and test set ratios of 80/20, 60/40 and 40/60 and the crucial parameters were tuned automatically.
Finally, the Decision Tree model with ratio split 60/40 was proven to be the best model to predict people who donate their blood in March 2007 to predict a large data set.

Table of contents

1 Introduction

Blood donation is the only source of blood supply since blood cannot be created in the laboratory, also, there are not any other alternative. According to the Australian Red Cross website (Australian Red Cross Lifeblood, n.d.), every blood donation can help save three lives. Blood can only be stored in the refrigerators at 6ºC and must be used within 42 days of donation (What Happens to Donated Blood, n.d.). Therefore, it is crucial to maintain and preserve adequate blood supply in case of major surgery or an extreme condition.

1.1 Project Goal

The project was conducted to help accurately predict whether a blood donor would re-donate as well as study how certain features would affect his or her decision, using two Machine Learning Classification models named K Nearest Neighbour and Decision Tree.

1.2 Data Set

The data set that was used for the prediction in this project is the "*Blood Transfusion Service Center Data Set*" collected from the UCI Machine Learning Repository database (UCI Machine Learning Repository: Data Set, 2008). The original purpose of this dataset was for Professor I-Cheng Yeh at Chung-Hua University to develop the RFMTC (Recency, Frequency, Monetary, Time, Churn) marketing model (UCI Machine Learning Repository: Data Set, 2008). To do so, the professor selected 748 random blood donors from the donor database; each of the data consists of 5 attributes which are:

- R (Recency): - months since last donation
- F (Frequency): - total number of donations
- M (Monetary): - total blood donated in c.c.
- T (Time): - months since first donation
- a binary value representing whether the donor donated blood in March 2007 (1 stands for donating blood and 0 stands for not donating)

Only the binary value in this case can be considered categorical data, whereas the other attributes are numerical.

2 Methodology

2.1 Data Retrieving

Uploading the File - the data was downloaded, uploaded to Anacondas, and named as "blood_transfusion_filename". Before checking the first 5 head and 5 last rows, check for the shape of the dataframe to see how many rows and columns in this data file and if it uploaded correctly. The data show that there were 748 rows and 5 columns, the Recency (months), Frequency (times), Monetary (c.c. blood), Time (months) and whether he/she donated blood in March 2007 was examined carefully and ensured that each of them match to the data set that was given in the original source.
Initial Data inspection - the data set was inspected for more information to double check if it matches with the original data that downloaded from the source. After inspection, data was described to see more basic statistical detail. Next when finished describing the data, the data was plotted in histogram to see the distribution of the data of 5 each column in the data set. But when reaching the column "whether he/she donated blood in March 2007" the column needs to be checked by value count to see each binary 1 and 0 exactly how many there were.

2.2 Feature Analysis

Correlation Coefficient - In this step after data retrieval, the data set were split up into columns. The correlation coefficient of these features were checked using the <.corr()> function This section is one of the most crucial step

as it provided insight on the relationship between each features and help confirm or reject the null hypothesis that we will have when plotting these features together.

Column investigation - next step after data exploration, each column was broken down furthermore to understand the dataset of each column by using the <.describe()> function it. It helped closely observe the minimum, maximum and the average for each column.
The data for each of these columns were then plotted and investigated using different plots such as histogram, density, bar and pie chart.

Relationship Investigation - the final step, after investigating each column was to investigate the relationship between these. The relationship between the columns were investigated from right to left to avoid redundant. Each column was plotted out to check if it needed another column to plot out, was it readable and was it easy to understand. The way we check relationships from left to right were:

- R (Recency - months since last donation),
- F (Frequency - total number of donation),
- M (Monetary - total blood donated in c.c.),
- T (Time - months since first donation),
- a binary variable representing whether he/she donated blood in March 2007 (1 stands for donating blood and 0 stands for not donating blood).

❖ R to F, M, T and whether he/she donated blood in March 2007.
❖ F to M, T, and whether he/she donated blood in March 2007.
❖ M to T and whether he/she donated blood in March 2007.
❖ T and whether he/she donated blood in March 2007.

2.3 Model Development

In this model development step, data was modelled using two Machine Learning model which is the K Nearest Neighbours and the Decision Tree Model.
The data was split into training set and test set with the ratio of 80/20, 60/40 and 40/60 to avoid overfitting. The parameters for each of these models were tuned accordingly – K for KNN and Criterion, maximum depth and min sample leafs for Decision Tree – and the result of the Confusion atrix <confusion_matrix()> and Classification Report <classification_report()> were observed and analysed to seek the best performing model. Hill Climbing method was then used to determine which parameters have the most impact on the model.

3 Results

3.1 Single Attribute Exploration

3.1.1 Recency

Figure 1: Recency Distribution

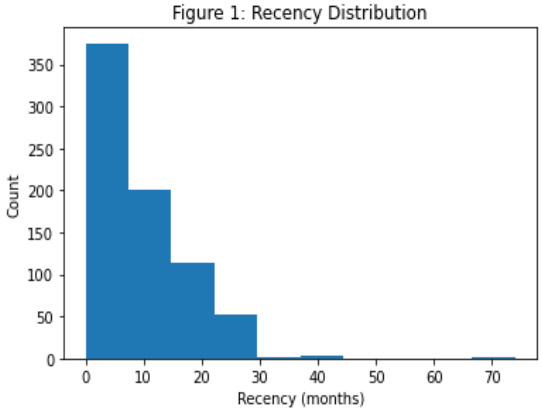
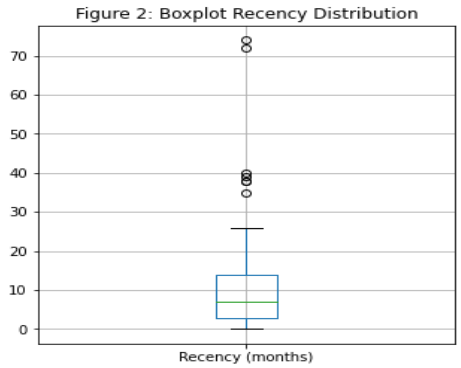
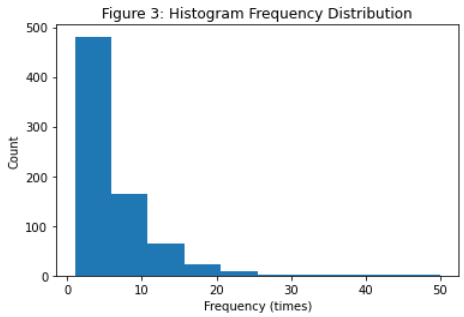


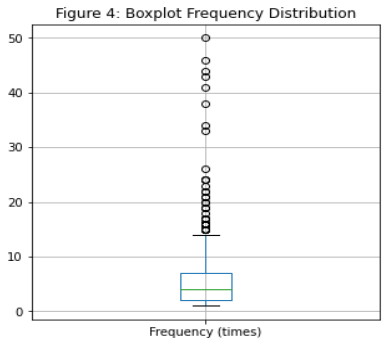
Figure 2: Boxplot Recency Distribution



The "Recency" column has 748 data, with the lowest value (minimum) at 0 months and the highest (maximum) at 74 months and an average of about 9.5 months. From Figure 1 and 2 above, it was observed that the data in this column heavily skewed toward the left, which means that most of the sample in this dataset has donated blood recently (within 24 months or 2 years). There are only 9 donors who have not donated within the last 2 years.

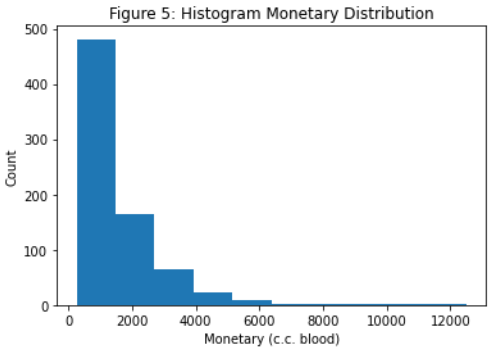
3.1.2 Frequency

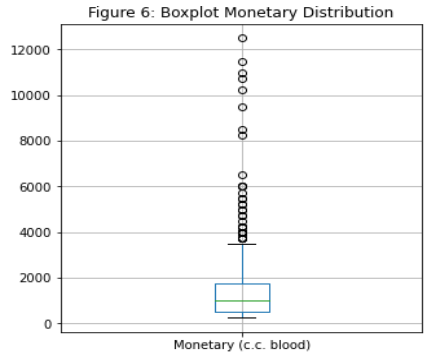




The "Frequency" column has 748 data. There are donors who only donated their blood 1 time (minimum) before, while there are also donors who have already donated 50 times (maximum) prior to when the data was collected. The average frequency that a donor donated his/her blood in this dataset is about 5.5 times. As can be seen from Figure 4 and 5, the skewness of the data in this column is to the left and that most of the donors donated less frequently than the average (481 values).

3.1.3 Monetary

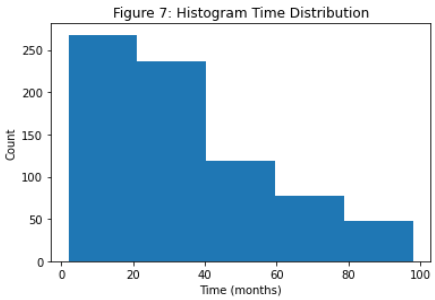


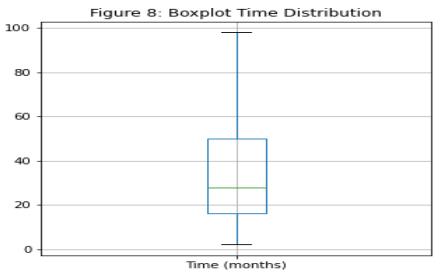


The "Monetary" column has 748 data. The lowest c.c blood (minimum) that a donor donated is 250, and the highest (maximum) is 12500 c.c. blood. The average blood that the donor donated in this dataset is about 1378 c.c. blood and the skewness, which can be observed from Figure 5 and 6 above, is also to the left. There also seems to be a relationship between the amount of blood a donor donated and the frequency his/her went to donate, as there are also 481 donors, who donated less than the average. This relationship will be further inspected in section

The inspection of the "Monetary" column also reveals that the values of the data in this column is much higher than the values of the data in the other columns. As a result of this, "Monetary" value can carry more weight in certain Machine Learning models, the authors recommended that a normalization should take place at this stage in a real project to achieve higher prediction accuracy.

3.1.4 Time



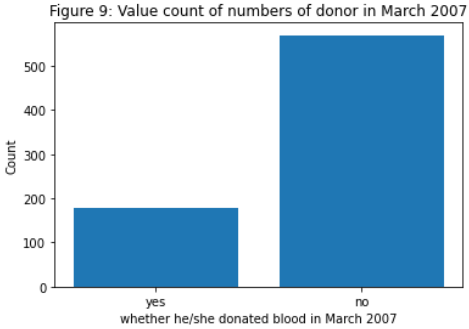


The "Time" column has 748 data. According to both figure 7 and 8 the highest number of people donating their blood is 98 months away from their first donation and the newest donors are 2 months. The average month since the first donation is about 34 months.

3.1.5 Whether he/she donated blood

According to Figure 9 and 10, the number of donors saying no to donation in March 2007 is much higher than the number of donors who will donate. There are only 178 donors (23.7%) said that they will donate their blood, while there are 570 donors (76.2%) said that they would not. Indeed, this difference also means that there would be more samples to train the Machine Learning Classification model for those who would not donate than those who would donate, which could result in the negative result being better recognized.

Figure 9: Value count of numbers of donor in March 2007

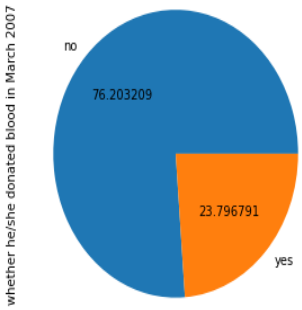


3.2 Relationship between Attributes Exploration

Due to the small number of features available and a bigger number of samples in this dataset, feature selection was not considered a very big concern. However, it is still important to exercise features relationship exploration so as to get a better understanding of the dataset.

A null hypothesis will be developed for each pair of numerical attributes, and this null hypothesis will consider the correlation of those involving features. A relationship that has an absolute value of correlation coefficient of 0.7 or higher would show that the pair has a strong relationship, a value between 0.5 and 0.7 would be moderate and lower than 0.5 would be considered a weak relationship (Moore, 2013). In the program, the function <.corr> was used to determine the correlation coefficients.

Figure 10: Pie Chart of Percentage of numbers of donor in March 2007



For the categorical data in column “whether he/she donated blood in March 2007”, a box plot plotting this value against the other columns will be presented.

Figure 11: Scatterplot showing the relationship between Recency and Frequency

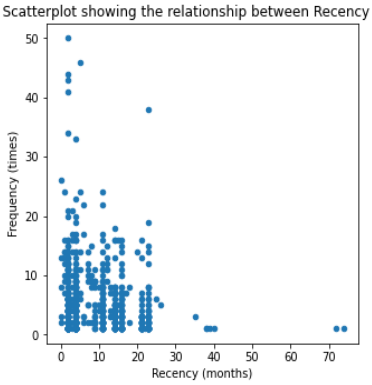
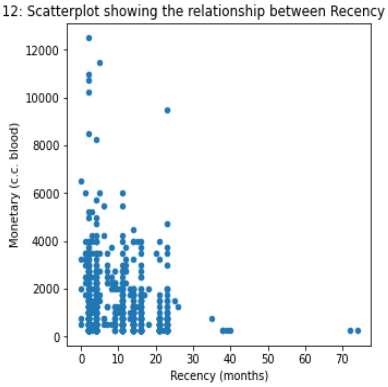


Figure 12: Scatterplot showing the relationship between Recency and Monetary



3.2.1 Recency (months) and Frequency (times)

H0 - the higher the Recency the lower the Frequency.

It is very difficult to tell whether an increase in the time since last donation would lead to a lower value in Frequency, however, when looking at a correlation coefficient of -0.18, this was confirmed. The relationship is a negative relationship, so the null hypothesis is accepted, but this is a very weak relationship as the coefficient is much lower than 0.5.

3.2.2 Recency (months) and Monetary (c.c. blood)

H0 - the higher the Recency the lower the Monetary (c.c. blood).

The scatterplot between Recency and Monetary is very similar to the scatterplot between Recency and Frequency. Therefore, this is a weak negative relationship, which means that the null hypothesis is accepted. The correlation coefficient for this relationship is also -0.18.

Figure 13: Scatterplot showing the relationship between Time (since first donation) and Recency (since last donation)

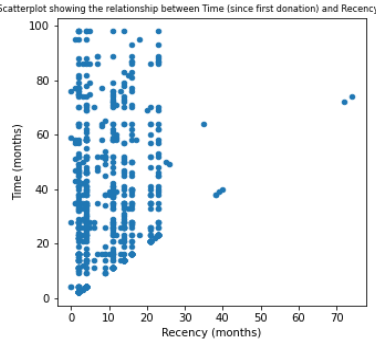


Figure 14: Scatterplot showing the relationship between Monetary and Frequency

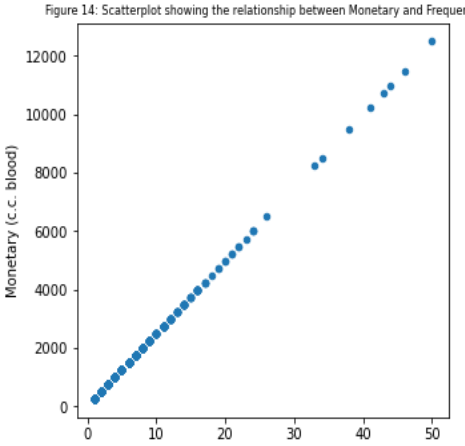


Figure 15: Scatterplot showing the relationship between Time (since first donation) and Frequency

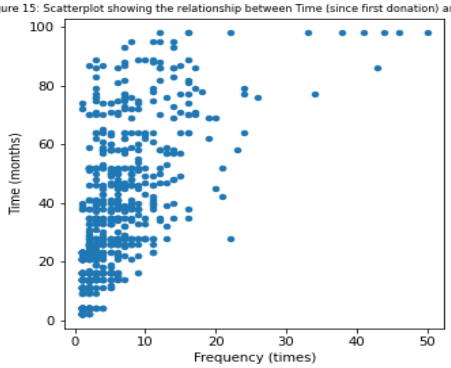
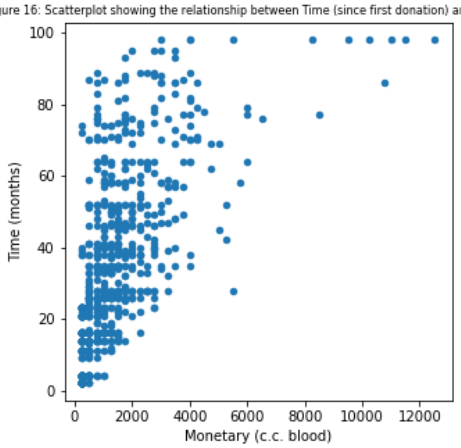


Figure 16: Scatterplot showing the relationship between Time (since first donation) and Monetary



3.2.3 Recency (months) versus Time (months)

H0 - the higher the Recency since the last donation the lower the Time since the first donation.
It can be seen from Figure 13, that there is an increase in the Time since the first donation as the Recency since the last donation increase. The null hypothesis is accepted; however, this is a very weak positive relationship, only with a correlation coefficient of 0.16.

3.2.4 Frequency (times) and Monetary (c.c. blood)

H0 - the higher the Monetary the higher the Frequency.
The relationship between the c.c. blood donation and the frequency a donator donated is a straight line. Indeed, this is a very strong relationship with a correlation coefficient of 1, which means that the donor who donate frequently always donate more blood than the donor who donate less frequently. For this reason, the null hypothesis is accepted of a strong positive linear relationship.
This kind of strong relationship is normally not utilised in a Machine Learning model, as these two features would have very similar effect on the result but would require more operating time. For the purpose of this assignment, both of these features were still kept.

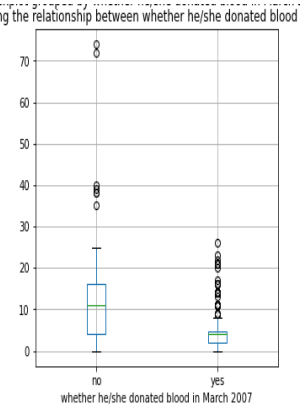
3.2.5 Frequency (times) and Time (months)

H0 - the higher the Frequency the lower the Time since the first donation.
It can be observed from Figure 15 that as the frequency increases the time since the first donation also increases, so the null hypothesis was rejected. The relationship has a correlation coefficient of 0.63, which means that these two features rather have a moderate positive relationship.

3.2.6 Monetary (c.c. blood) and Time (months)

H0 - the higher the Monetary the lower the Time since the first donation.
Figure 16 shows that there is also a positive relationship between the Monetary and Time since the first donation, so the null hypothesis is rejected. The relationship is the same as the relationship between the Frequency and the Time and has a correlation coefficient of 0.63. This relationship is classified as a moderate positive relationship.

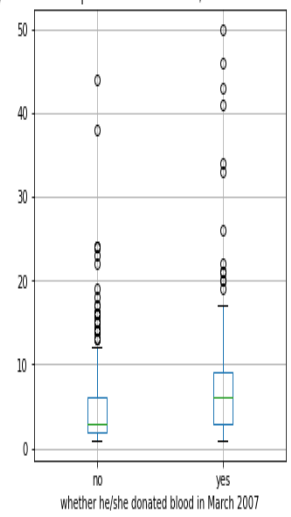
Figure 17: Boxplot showing the relationship between whether he/she donated blood in March 2007 and Recency



3.2.7 "Whether he/she donated in March 2007" by Recency (months)

Figure 17, "whether he/she donated in 2007 by Recency" shows that people who were not recently donated blood tend to say "no" to donating their blood in March 2007. Most donors who say "yes" donated very recently, often within less than 30 months. The average Recency value of people who say "no" is also higher than that of those who say "yes". As can be seen from the figure, there are more people who say "no" than people who say "yes" when using Recency to compare. The correlation coefficient for this relationship is -0.27, which confirms that the Recency feature has a weak negative relationship with the number of "yes" answer in March 2007.

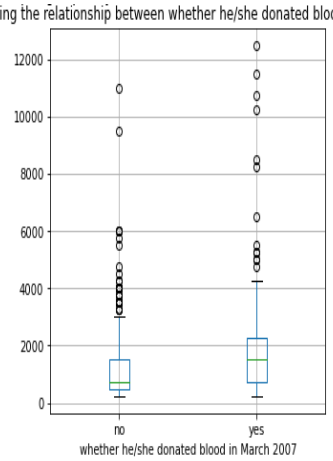
Figure 18: Boxplot showing the relationship between whether he/she donated blood in March 2007 and Frequency



3.2.8 "Whether he/she donated in March 2007" by Frequency (times)

Figure 18, "whether he/she donated blood in 2007 by Frequency" shows that people who frequently donated blood tend to say "yes" on March 2007, while those who say "no" tend to less frequently do blood donation. It is denoted that from the figure that the average Frequency donating blood of those donors who say "yes" is around 5.5 times, whereas of those who say "no" only around 3 times. From this it was concluded that there is a positive relationship between the frequency and the number of "yes" answer (the correlation coefficient is 0.21)

Figure 19: Boxplot showing the relationship between whether he/she donated blood in March 2007 and Monetary

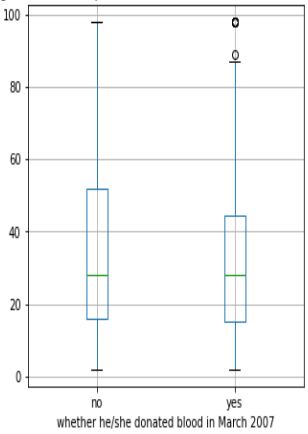


3.2.9 "Whether he/she donated in March 2007" by Monetary (c.c. blood)

Figure 19, "whether he/she donated in 2007 by Monetary", shows that people who donate more blood prior to March 2007 tend to say "yes" more than the people who donate less blood. So, there is a positive relationship between the monetary of blood a donor donates and the number of "yes" answer. This relationship. is the same with the relationship in section 3.2.8; this is understandable as Frequency and Monetary has a strong positive relationship.

The average blood donation for those who say "yes" is roughly 1800, while the average for those who say "no" is around 700.

Figure 20: Boxplot showing the relationship between whether he/she donated blood in March 2007 and Time



3.2.10 "Whether he/she donated in March 2007" by Time (months)

The relationship between Time since first donation and "whether he/she donated in March 2007" is shown in figure 20. The relationship, however, is not very clear, as the two boxplots in this figure are very similar. An average Time since first donation of around 28 months is available for both those who say "yes" and "no".

From this figure, it was concluded that Time since first donation does not have as strong relationship with the answer comparing to the other features. The correlation coefficient for this relationship is -0.03, which confirms a weak negative relationship.

3.3 Models

3.3.1 Model selection criteria

The aim of this project is to accurately predict which donor donated in March 2007, so it is equally important between predicting either "yes" or "no" answer. For this reason, the overall accuracy is the priority. Aside from the overall accuracy, confusion matrix and classification report were also used when comparing between each model. Specifically, precision, recall and F1 score on "yes" is crucial, since this prediction can be used to improve usage and preservation of donated blood.

Each model was evaluated according to this pipeline:

Overall Accuracy/Confusion Matrix/Precision on "yes"/Recall on "yes"/F1 score on "yes"

The following terms will be used in analysing the confusion matrix:

TP (True positive): Correctly predict that a donor would donate in March 2007.

TN (True negative): Correctly predict that a donor would not donate in March 2007.

FP (False positive): Incorrectly predict that a donor would donate in March 2007.

FN (False negative): Incorrectly predict that a donor would not donate in March 2007.

Finally, for all the selected models, "Hill Climbing Method" was used to analyse the impact of each parameter and whether using less parameter would improve the accuracy of the model.

3.3.2 KNN Model

3 models were developed using the K - Nearest Neighbour Classification. The difference between these models was how the train - test split between each model is done. Model 1 has a train - test split ratio of 80% and 20%. Model 2 has a train - test split ratio of 60% - 40%. And model 3 has a ratio of 40% - 60%. For each of these models, the K - Nearest Neighbour value was tuned to achieve higher overall accuracy. These models were then analysed using Confusion Matrix and Classification Report.

3.3.2.1 Model 1 (80/20 Train Test Split)

There are 598 entries for the training data and 150 entries for the testing data. The K value was repeatedly tuned for 30 times and the overall accuracy was recorded and compared. The value 30 was chosen because it is popular practice in real life project that a square root of the number of samples would be chosen as K value (Band, 2020). The best performing K value is at 13 neighbours (77% accuracy), therefore this model was chosen to analyse.

According to the confusion matrix:

TP = 10 *correctly predicted that the donor would donate blood.*

TN = 106 *correctly predicted that the donor would not donate blood.*

FP = 31 *incorrectly predicted that the donor would donate blood.*

FN = 3 *incorrectly predicted that the donor would not donate blood.*

This model has a very high False Positive (FP) and a very small False Negative (FN), which is not good for blood supply preparation since it would lead to over - optimistic when preparing as well as a smaller number of storages for the blood supply. On the other hand, this, however, can lead to less wastage as the blood supply chain would try to use up all the old blood since they would be expecting a large amount of blood donation.
Precision for "yes": 0.77*. The model accurately predicted 10 times out of 13 positive predictions.*
Recall for "yes": 0.24*. The model performs poorly on Recall for "yes" and only accurately predicted 10 out of 41 times*
F1 - score for "yes": 0.37*. The highest value of F1 - score is 1, due to a low in recall this model also has a low F1 - score.*
Overall accuracy - 77.3%

3.3.2.2 Model 2 (60/40 Train Test Split)

There are 448 training entries and 300 test entries for this model 2. In the same way that the K value was tuned above, it was tuned in this model as well. The best performing K value for this model is 10 (79.3% accuracy), therefore this model was chosen to analyse.
According to the confusion matrix:
TP = 13 *correctly predicted that the donor would donate blood.*
TN = 225 *correctly predicted that the donor would not donate blood.*
FP = 61 *incorrectly predicted that the donor would donate blood.*
FN = 3 *incorrectly predicted that the donor would not donate blood.*
This model has the same problems as the previous model, which has a very high False Positive (FP) and a very low False Negative (FN).
Precision for "yes": 0.93*. The model accurately predicted 13 times out of 14 positive predictions.*
Recall for "yes": 0.18*. The model performs poorly on Recall for "yes" and only accurately predicted 13 out of 54 times*
F1 - score for "yes": 0.3*. This model has a lower F1 - score compared to the previous model; this is due to a lower recall value.*
Overall accuracy - 79.3%

3.3.2.3 Model 3 (40/60 Train Test Split)

There are 299 training entries and 449 test entries for this model 3. The best performing K value for this model is 26 (76.6% accuracy). Even with a very high number of K values, which means that there is less noise, the model still performs poorly compared to the previous 2 models. This can be expected as the training set is smaller than the test set. =
According to the confusion matrix:
TP = 0 *correctly predicted that the donor would donate blood.*
TN = 344 *correctly predicted that the donor would not donate blood.*
FP = 105 *incorrectly predicted that the donor would donate blood.*
FN = 0 *incorrectly predicted that the donor would not donate blood.*
The model does not make any yes prediction which leads to the following result in the Classification Report.
Precision for "yes": *0.00*
Recall for "yes": *0.00.*
F1 - score for "yes": *0.00*
Overall accuracy - 76.6%

3.3.3 Decision Tree Model

Like how the data was split for the KNN model, the data was also split for the Decision Tree Model. The three models include 80% - 20% train - test ratio, 60% - 40% ratio and 40% - 60% ratio. Three parameters were also tuned accordingly for these models, including the criterion, the maximum depth and the minimum sample leaf's. These parameters were chosen for tuning because they are the deciding factors in controlling the impurities of the decision tree as well as to avoid overfitting when fitting the test data (Yong Li 2021).
In these models, the criterion parameter was tuned manually between entropy and gini, while the value for maximum depth and minimum sample leaf's were tuned automatically between 0 and 10 at the same time to achieve maximum accuracy using both together.

3.3.3.1 Model 1 (80/20 Train Test Split)

Criterion - entropy: Best Parameter: max_depth = 5, min_samples_leaf = 9

According to the confusion matrix:
TP = 20 correctly predicted that the donor would donate blood.
TN = 99 correctly predicted that the donor would not donate blood.
FP = 21 incorrectly predicted that the donor would donate blood.
FN = 10 *incorrectly predicted that the donor would not donate blood.*
This model has a False Positive value of 21 and a False Negative of value of 0.
Precision for "yes": 0.67. *The model accurately predicted 20 times out of 30 positive predictions.*
Recall for "yes": 0.49. *The model performs slightly better on Recall for "yes" than comparing to KNN - model with a succeed rate of 20 out of 41 times*
F1 - score for "yes": 0.56. *This is the highest value of F1 - score achieved.*
Overall accuracy - 79.3%

Criterion - gini: Best Parameter: max_depth = 9, min_samples_leaf = 4
According to the confusion matrix:
TP = 16 *correctly predicted that the donor would donate blood.*
TN = 104 *correctly predicted that the donor would not donate blood.*
FP = 25 *incorrectly predicted that the donor would donate blood.*
FN = 5 *incorrectly predicted that the donor would not donate blood.*
According to the Classification Report:
Precision for "yes": 0.76. *This gini model accurately predicted positive 16 times out of 21 positive predictions, which is better than the entropy model.*
Recall for "yes": 0.39. *The model, however, performs poorly on recall with only 16 success predictions out of 41.*
F1 - score for "yes": *0.52.*
Overall accuracy – 80%

3.3.3.2 Model 2 (60/40 Train Test Split)

Criterion - entropy: Best Parameter: max_depth = 5, min_samples_leaf = 8
According to the confusion matrix:
TP = 34 *correctly predicted that the donor would donate blood.*
TN = 209 *correctly predicted that the donor would not donate blood.*
FP = 40 *incorrectly predicted that the donor would donate blood.*
FN = 17 *incorrectly predicted that the donor would not donate blood.*
According to the Classification Report:
Precision for "yes": 0.67. *The model accurately predicted 34 times out of 51 positive predictions.*
Recall for "yes": 0.46. *The model only accurately predicted positive values 34 times out of the total of 74 actual "yes" values.*
F1 - score for "yes": 0.54.
Overall accuracy - 81%

Criterion - gini: Best Parameter: max_depth = 4, min_samples_leaf = 2
According to the confusion matrix:
TP = 38 correctly predicted that the donor would donate blood.
TN = 203 correctly predicted that the donor would not donate blood.
FP = 36 incorrectly predicted that the donor would donate blood.
FN = 23 incorrectly predicted that the donor would not donate blood.
According to the Classification Report:
Precision for "yes": 0.62. *The gini model only accurately predicted positive 38 times out of 61 times that it actually predicted positive.*
Recall for "yes": 0.51. *The model, however, has the best performance on recall with over 50% (38 success attempts out of 74).*
F1 - score for "yes": 0.56 *which is also the highest F1 - score value.*
Overall accuracy - 80.3%

3.3.3.3 Model 3 (40/60 Train Test Split)

Criterion - entropy: Best Parameter: max_depth = 4, min_samples_leaf = 6
According to the confusion matrix:
TP = 41 *correctly predicted that the donor would donate blood.*
TN = 315 *correctly predicted that the donor would not donate blood.*
FP = 64 *incorrectly predicted that the donor would donate blood.*

FN = 29 *incorrectly predicted that the donor would not donate blood.*
According to the Classification Report:
Precision for "yes": 0.59. *The model made a total of 70 positive predictions and only predicted accurately 41 times.*
Recall for "yes": 0.39. *The model only accurately predicted positive values 41 times out of the total of 105 actual "yes" values.*
F1 - score for "yes": 0.47.
Overall accuracy - 79.3%

Criterion - gini: Best Parameter: max_depth = 4, min_samples_leaf = 6
According to the confusion matrix:
TP = 41 *correctly predicted that the donor would donate blood.*
TN = 315 *correctly predicted that the donor would not donate blood.*
FP = 64 *incorrectly predicted that the donor would donate blood.*
FN = 29 *incorrectly predicted that the donor would not donate blood.*
According to the Classification Report:
Precision for "yes": 0.59. *The model made a total of 70 positive predictions and only predicted accurately 41 times.*
Recall for "yes": 0.39*. The model only accurately predicted positive values 41 times out of the total of 105 actual "yes" values.*
F1 - score for "yes": 0.47.
Overall accuracy - 79.3%
There is no different in the performance of the two models in this case

3.3.4 Hill Climbing

The Hill Climbing method was applied to all of the models above to determine which one of the features or which combinations of features had the most impact on the decision to whether a donor would donate blood in March 2007. It was found out that using a single feature can greatly improve the overall accuracy of the model, however, it cannot be determined from this dataset which features would have the most effect on the decision.

4 Discussion

4.1 Overview

After closely observing the result, it was determined that, for this dataset, Decision Tree Classifiers tend to make more "yes" predictions than K - Nearest Neighbor Classifiers. Indeed, Decision Tree Classifiers tend to have better recall value for "yes" prediction than K - Nearest Neighbors. However, K - Nearest Neighbors often have higher precision in predicting "yes". As stated in the Introduction, blood donation management is extremely important so both the prediction of "yes" and "no" is crucial, with the prediction of "yes" slightly more to avoid over-optimism.

4.2 KNN Model

The model which was chosen is Model 1 with train - test split ratio of 80 - 20 and having a K value of 13. Although Model 2 has a slightly higher overall accuracy (79.3%) compared to the Model 1 (77%), it was the higher recall on predicting "yes" (24%) and the higher F1 score (37%) that made this model superior. The decision was made since, it is better to be prepared for not having enough blood supply than being too optimistic and only care about blood wastage.
Model 3 was not considered in the comparison as it does not make any "yes" prediction.

4.3 Decision Tree Model

The chosen Decision Tree Model is Model 2 with train - test split ratio of 60 - 40 ratio and having gini criterion, a maximum depth of 4 and a minimum sample leafs of 2. This model has a very high overall accuracy of 80.3% and a very high recall value for predicting "yes" (51%). This met the criteria of being pessimistic on the blood supply so as to be able to prepare for the worst case scenario of running out of blood.

4.4 Recommendation

As previously mentioned, the performance of Decision Tree Model on this dataset is better than KNN model, thus, the Decision Tree with the train - test split ratio of 60 - 40, with gini criterion, a maximum depth of 4 and a minimum sample leafs of 2 was concluded the best model.
It was also recommended by the author that more data is needed to improve the accuracy of the model for future uses. Once crucial acknowledgement would be the fact that this dataset has more people answering “no” (570 entries) than “yes” (178 entries), which makes the model perform better predicting donors who do not want to re-donate blood. The data is also not enough to experience which features have the most impact on the result of the model, which can be improved with a bigger sample size.

5 Conclusion

After multiple testing of different models, using both kNN Model and Decision Tree model to predict accuracy of the dataset. Decision Tree Model with the split 60/40 on the dataset should be used for prediction. This is because it was concluded to predict more accurately of people answering “no" (570 entries) and “yes” (178 entries) than KNN model can predict.

References

Archive.ics.uci.edu. 2008. *UCI Machine Learning Repository: Data Set*. [online] Available at: <https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center)> [Accessed 23 May 2021].

Band, A., 2020. *How to find the optimal value of K in KNN?*. [online] Medium. Available at: <https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb> [Accessed 23 May 2021].

Donateblood.com.au. n.d. *Learn how your donated blood is used | Australian Red Cross Lifeblood*. [online] Available at: <https://www.donateblood.com.au/learn#> [Accessed 23 May 2021].

Redcrossblood.org. n.d. *What Happens to Donated Blood*. [online] Available at: <https://www.redcrossblood.org/donate-blood/blood-donation-process/what-happens-to-donated-blood.html> [Accessed 23 May 2021].

Moore, D. S., Notz, W. I, & Flinger, M. A. (2013). The basic practice of statistics (6th ed.). New York, NY: W. H. Freeman and Company.

Li, Y (2021). Lecture Slide Week 6: Classification. RMIT Canvas.