**Assignment 1**

Logo, company name

Description automatically generated

**FACULTY OF INFORMATION TECHNOLOGY**

**ITRMA4-12**

**USING CUSTOMER SEGMENTATION AND PREDICTIVE MODELLING TECHNIQUES TO PREDICT BUILDING INSURANCE UPSELLS**

**Kyle Searle**

**EDUV4918080**

|  |  |  |
| --- | --- | --- |
| **NAME** | **SURNAME** | **STUDENT NUMBER** |
| Kyle | Searle | EDUV4918080 |
|  |  |  |
|  |  |  |

Assignment 1 submitted in partial fulfilment of the requirements for ITRMA4-12

**Supervisor:** M. Satekge

**Date:** 05 April 2024

Contents

[Question 1 3](#_Toc170229563)

[Question 2 5](#_Toc170229564)

[The following has been done to determine if there is any Null/None values 5](#_Toc170229565)

[Question 3 13](#_Toc170229566)

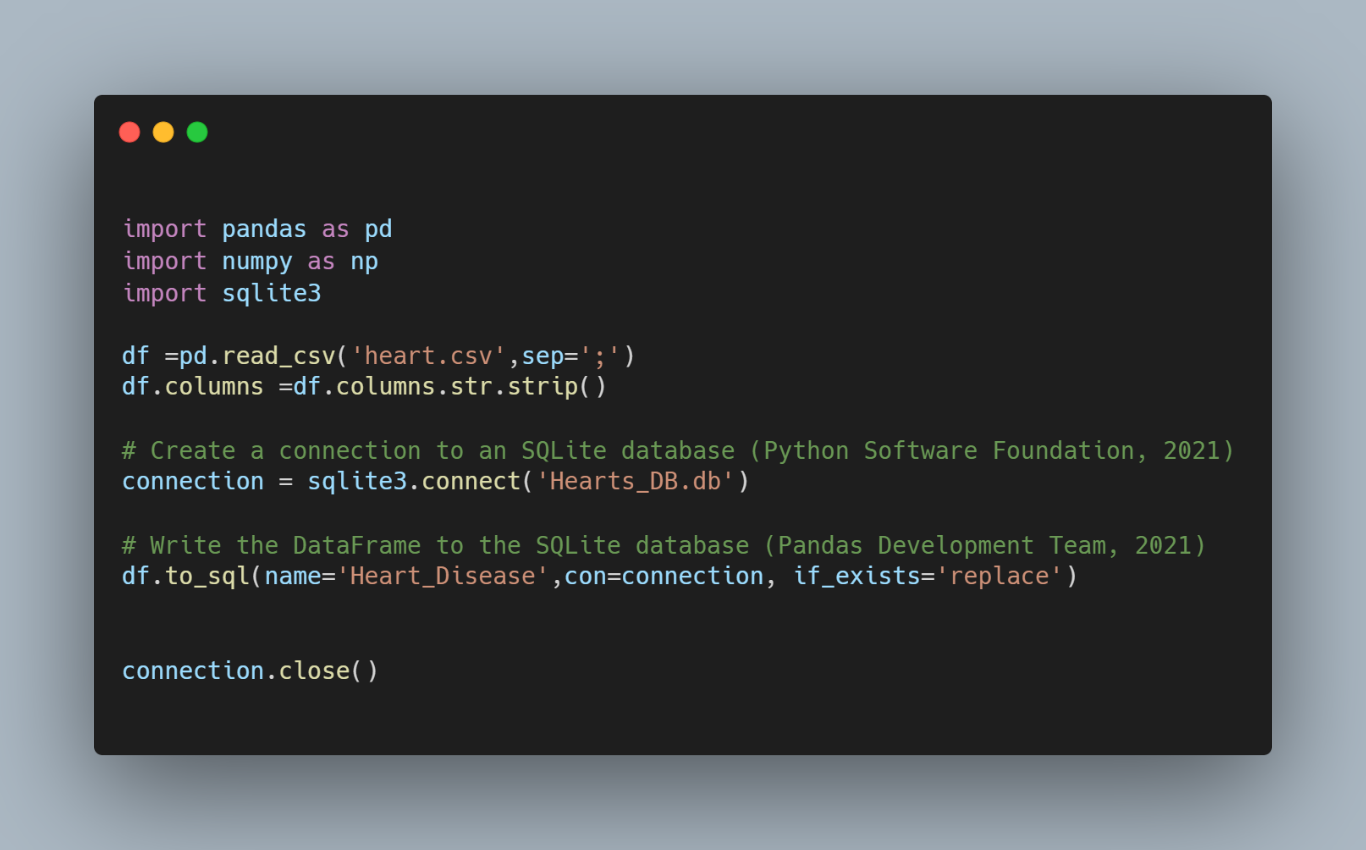
[Question 4 32](#_Toc170229567)

[References 32](#_Toc170229568)

# Question 1

I start by reading the “heart.csv” into a data frame called “df”. The column names of the “df” are cleaned by stripping whitespace using the “str.strip()” method. A connection to an SQLite database named “Hearts\_DB.db” is then created using sqlite3.connect(). The “df” is written to this database with the name “Heart\_Disease” using the to\_sql() method. If a table with this name already exists in the database, it will be replaced. Finally, the connection to the database is closed using ‘connection.close() ‘.

The code in Create\_sqlite3 DB.py follows:



Read the data from the sqlite3 Hearts\_DB database to be used for the next questions:



The Output to view the queried data base:

A screenshot of a computer program

Description automatically generated

# Question 2

1. Imputing missing values in datasets is crucial as it preserves statistical power and prevents bias, enhancing the reliability of models (Little & Rubin, 2002; Schafer & Graham, 2002). The choice of imputation method, such as mean or regression imputation, should be carefully selected based on the data’s characteristics (Rubin, 1996; Sterne et al., 2009).

## The following has been done to determine if there is any Null/None values

|  |  |
| --- | --- |
| Code: | Output: |
|  |  |

As displayed, there are no empty values to impute. I furthermore explored outliers using the following code:

|  |  |
| --- | --- |
| Code:  A screenshot of a computer program  Description automatically generated | Output: |

Here is the Box plots for the numerical data, to highlight outliers:

A screenshot of a graph

Description automatically generated

Code:

A screen shot of a computer

Description automatically generated

All these outliers are replaced via mean imputation. Mean imputation, also known as mean substitution, is a statistical method used to handle missing data in a dataset. In this method, missing values of a certain variable are replaced by the mean of the non-missing cases of that variable (Statistics Globe, n.d.).

Code:

A screen shot of a computer

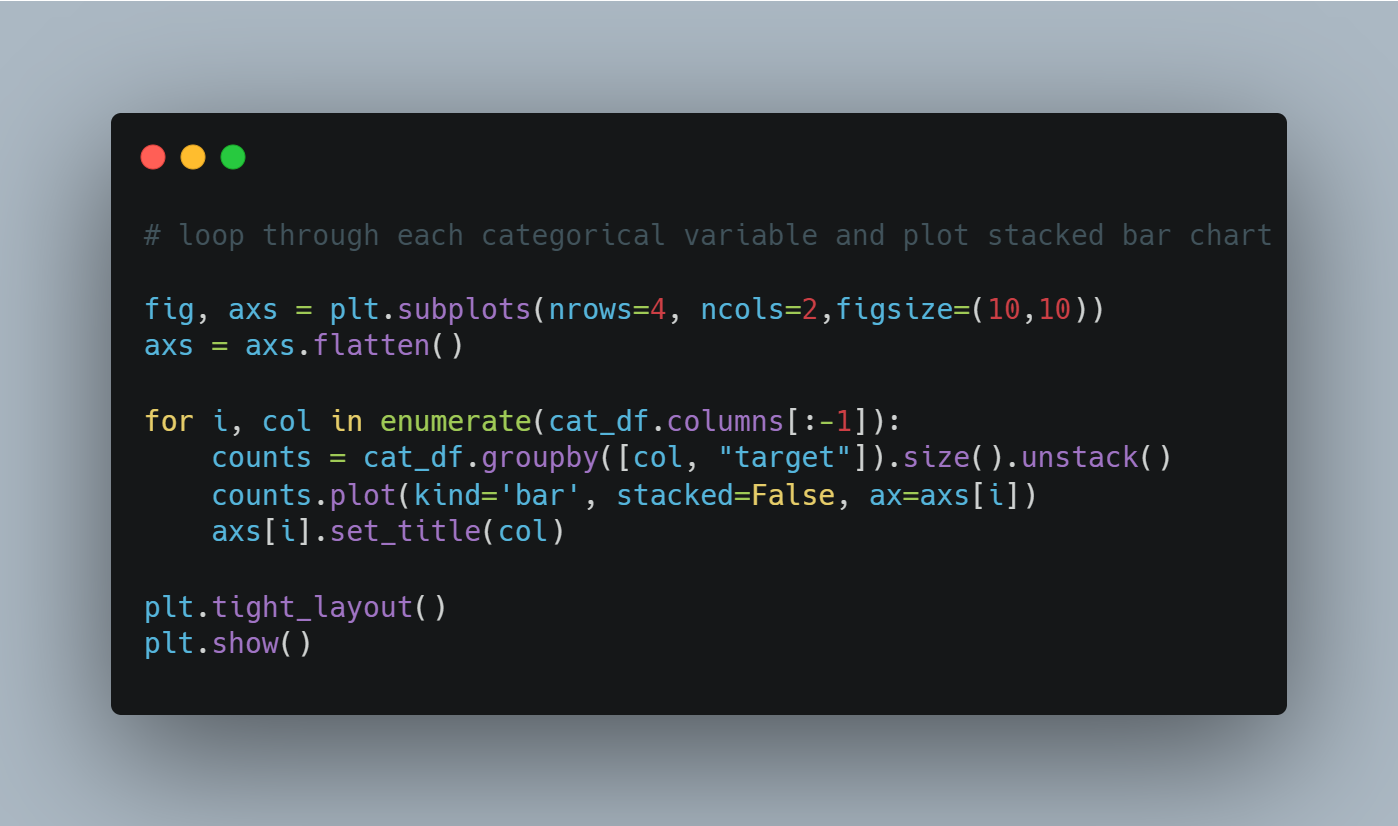
Description automatically generated

*Categorical Variables Distribution Charts per Target Variable Class*

A group of blue and orange bars

Description automatically generated

The following code was used to generate the bar charts of the distribution of the categorical variables with respect to the target variables classes.

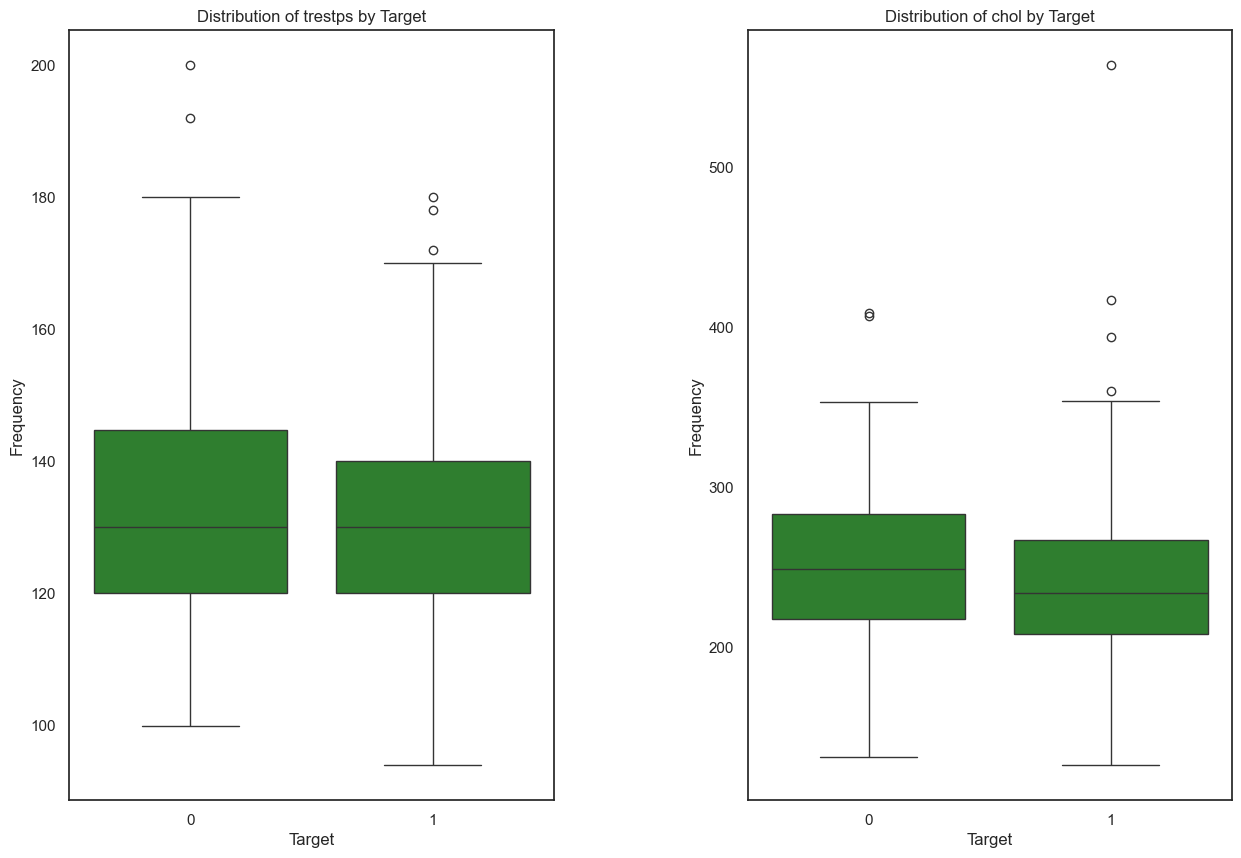


From the charts we can make the following observations:

1. The distribution suggests that class 1 of the target variable is more prevalent in females compared to males. Inciting that Females are more likely to have heart disease in ratio to men.
2. Categories 1,2 and are clearly more associated with class 1, heart disease, of the target variable.
3. The ‘fbs’ variable seems to have a similar distribution for both classes of the target variable.
4. Individuals with abnormal ‘restecg’ results appear to be more associated with class 1 of the target variable, this could indicate people with abnormal resting electrocardiographic results are somewhat correlated with heart disease.
5. The distribution suggests a clear difference between the categories within ‘exang’ in relation to the target variable classes. Patients who do not have exercise induced angina are strongly correlated with heart disease and vice versa for individuals who are considered with regards to their hearts.
6. Patients with mild number of major vessels coloured appear to be much more associated with class 1(heart disease) of the target variable. There is a very strong correlation between heart disease and mild number of major vessels coloured by fluoroscopy.
7. The ‘slope’ variable shows a distinct distribution for the two classes of the target variable. Downwards slope is highly correlated to having a heart disease and
8. Certain categories within ‘thal’ appear to be more associated with class 1 of the target variable.



*Trestps & Chol Distribution Charts per Target Variable Class*



1. Regarding box plot of ‘trestbps’:
2. The median is less sensitive to skewed data and outliers than the mean (Johnson, 2015). Extreme values pull the mean away from the center of the distribution (Brown, 2007). For both targets, the median is around the same value slightly above 120.
3. The Interquartile Range (IQR) measures the spread of the middle 50% of your data (Brown, 2007). A larger IQR indicates that the central portion of your data is spread out further (Smith, 2010). In other words, the data points are more widely dispersed from the mean when the IQR is high. Conversely, smaller values of the IQR show that the middle values cluster more tightly (Brown, 2007). The interquartile range (IQR) for target 0 appears slightly larger than for target 1 (heart disease), with more outliers above the upper whisker.
4. Regarding the box plot of ‘chol’:
5. The median value for target 0 is higher compared to target 1, with the median value just below 250 for target 0 and just above or at about 240 for target 1.
6. The interquartile range (IQR) is larger for target 1 than it is for target 0.
7. There are several outliers indicated in both targets; however, there are more outliers above the upper whisker in target 1 (heart disease).

*Thalach, Oldpeak & Age Distribution Charts per Target Variable Class*

A group of green boxes

Description automatically generated

1. Regarding the box plot of ‘thalach’:
2. The box plot for target 0 shows a median around 140, with the interquartile range (IQR) approximately between 120 and 160.
3. There are outliers indicated below the lower whisker.
4. For target 1, the median is higher, around 160, with an IQR roughly between 150 and 170.
5. Regarding the box plot of ‘oldpeak’:
6. The IQR (body) for target class 0 (healthy) is noticeably greater than class 1.
7. Likewise, the maximum is much greater for target 0.
8. There are very few outliers above the upper whisker for the target class 0.
9. For target 1, the median is close/at zero with an IQR that are smaller than preceding target class
10. There are also outliers present.
11. Regarding the box plot of ‘age’:
12. The IQR for target 1 is much greater than target 0’s IQR. This indicates 50% of individuals with a heart disease is spread out between the ages of 45 and 60.
13. The higher median for target 1 indicates, that 50% of the individuals with a health heart are older than 57.

The following code was used to generate the bar charts of the distribution of the categorical variables with respect to the target variables classes:



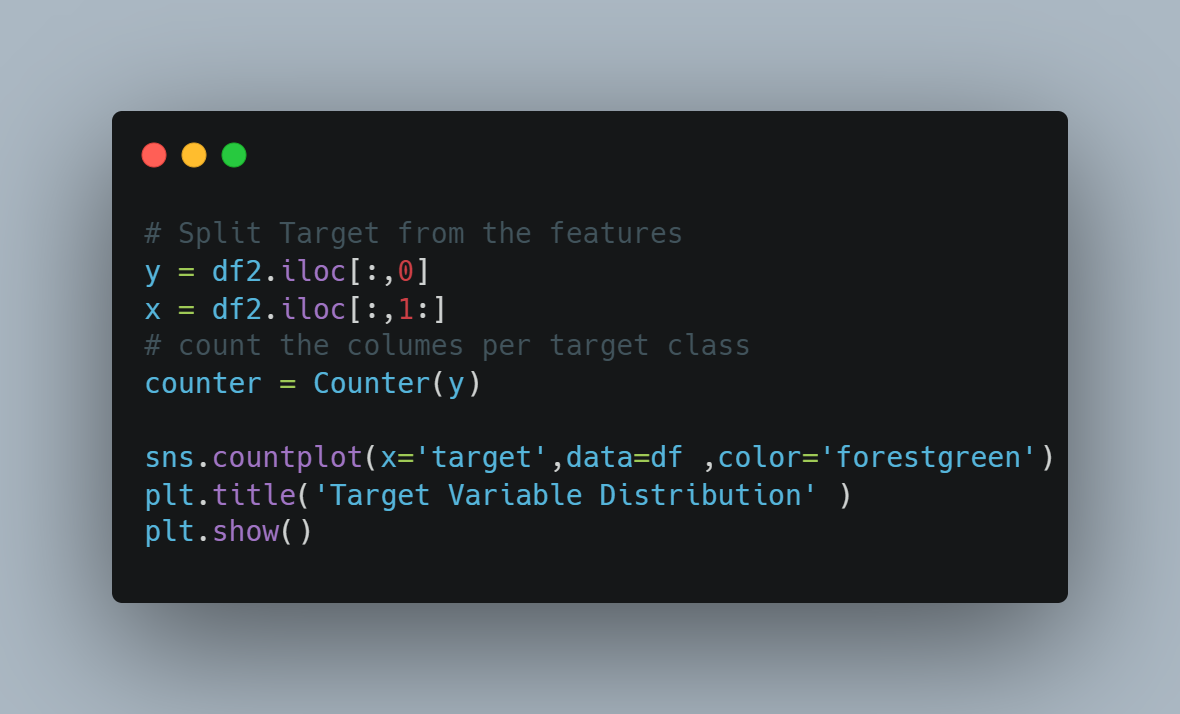
# Question 3

I start by standardizing all my numerical variables. Standardization is a statistical technique often used in data analysis and machine learning. It involves rescaling the values of a numerical feature to have a mean of 0 and a standard deviation of 1. This process is crucial as it brings different variables to the same scale, allowing for fair comparison and accurate model predictions (Jain, 2019).

Code used:

I then follow up to see the target variable distribution:

Code used:

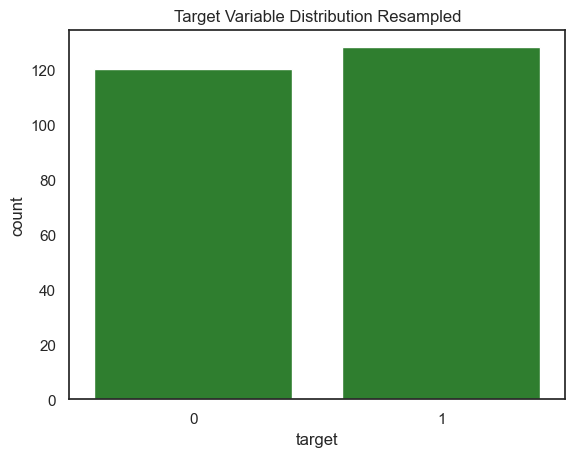


A green bar graph with white text

Description automatically generated

There is a slight imbalance, I do need to employ a sampling method. I decided to use a sampling method called SMOTETomek. SMOTETomek is a hybrid resampling technique that combines the Synthetic Minority Oversampling Technique (SMOTE) and Tomek Links methods. SMOTE works by generating synthetic examples from the minority class to balance the dataset, while Tomek Links remove the instances of the majority class that are close to the minority class, thus making the decision boundary more distinct (Sharma & Gosain, 2023). The combination of these two techniques in SMOTETomek can lead to a more robust model. The oversampling part (SMOTE) adds more information to the minority class, helping the model to better understand its characteristics. On the other hand, the under sampling part (Tomek Links) removes noisy instances from the majority class, reducing the risk of overfitting and making the decision boundary clearer (Wongvorachan et al., 2023). Therefore, SMOTETomek can be a good sampling method when dealing with imbalanced datasets, as it not only balances the classes but also improves the quality of the data used for training the model.

Results after SMOTETomek is applied:

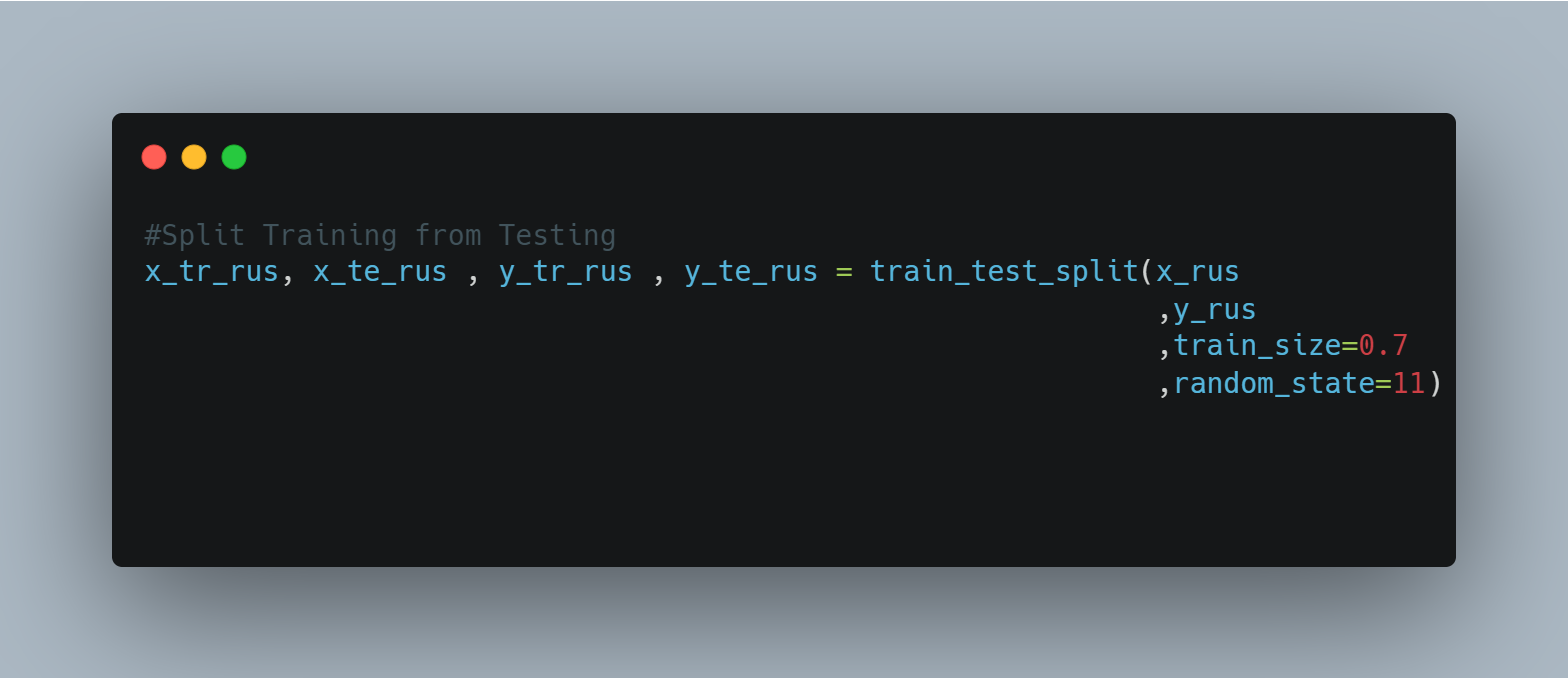


Code (I start by splitting data up in a combination set & Validation set, which after I apply the SMOTETomek sampling to the Combination set):

A screenshot of a computer program

Description automatically generated

Final preprocessing is splitting the sampled data in *training* & *testing*.

Code: 

3.2.

The models I decided to fit & train & validate is:

1. Gradient Boosting Classifier (GBC)
2. Logistic Regression
3. Stochastic Gradient Descent Classifier (SGD)

Introductions, advantages & disadvantages for each model follows:

1. Gradient Boosting Classifier (GBC)

Gradient Boosting is a powerful ensemble machine learning algorithm that builds a model from an ensemble of weak learners (typically decision trees) in a stage-wise fashion. It optimizes a differentiable loss function by adding weak learners to minimize the residual errors (Friedman, 2001).

Advantages:

1. It often provides predictive scores that are far better than other algorithms (Natekin & Knoll, 2013).
2. It can handle missing data, so imputation is not required (Elith et al., 2008).

Disadvantages:

1. It can be computationally expensive and slow to train, especially with large datasets (Natekin & Knoll, 2013).
2. It requires careful tuning of parameters and regularization to avoid overfitting (Natekin & Knoll, 2013).
3. Logistic Regression/Classifier\*\*

Logistic Regression is a statistical analysis model used for binary classification problems. It estimates the probability of an event occurring based on one or more predictor variables (Cramer, 2002).

Advantages:

1. It is easy to implement, interpret, and very efficient to train (Peng et al., 2002).
2. It can handle categorical features very well (Hosmer Jr et al., 2013).

Disadvantages:

1. It requires the independent variables to be linearly related to the log odds (Cramer, 2002).
2. It can only be used to predict discrete functions; hence the dependent variable is bound to the discrete number set (Cramer, 2002).
3. Stochastic Gradient Descent Classifier (SGD)

The Stochastic Gradient Descent (SGD) is an optimization algorithm used to find the values of parameters of a function that minimizes a cost function. It is best used when the datasets are large (Bottou, 2010).

Advantages:

1. It is efficient and easy to implement, providing a lot of opportunities for code tuning (Bottou, 2010).
2. It is suitable for large-scale problems due to its speed and memory efficiency (Zhang, 2004).

Disadvantages:

1. SGD requires several hyperparameters such as the regularization parameter and the number of iterations (Bottou, 2010).
2. SGD is sensitive to feature scaling (Bottou, 2010).

**Testing**

Testing Confusion Matrixes:

|  |  |
| --- | --- |
| Gradient Boosting Classifier:  A yellow and purple squares with numbers  Description automatically generated | Logistic Regression Classifier: |
| Stochastic Gradient Descent: | We can see the LR is slightly outperforming the other 2 models when inspecting the testing confusion matrices with 33 True positive predictions & the least number of false positives at 7 and very important the least number of false negatives at 3. |

ROC Testing Curves:

|  |  |
| --- | --- |
| GBC: | LR: |
| SGD: | Inspecting the ROC curve, it is quite clear that the GBC has the smoothest ROC and the highest AUC/ROC score at 0.925 |

The final testing metrics:

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | GBC | LR | SGD |
| Accuracy | 0.84 | 0.87 | 0.84 |
| Precision | 0.82 | 0.82 | 0.79 |
| Recall | 0.86 | 0.92 | 0.92 |
| F1-Score | 0.84 | 0.87 | 0.85 |
| AUC/ROC | 0.925 | 0.92 | 0.92 |

As seen in the testing metrics table, the Logistic Regression model is performing the best base on the testing data. But take note that the testing data is over & under sampled using SMOTETomek. So, to get a clear champion model, I will evaluate metrics on the pure validation data.

**Validation**:

Validation Confusion Matrices:

|  |  |
| --- | --- |
| Gradient Boosting Classifier: | Logistic Regression Classifier: |
| Stochastic Gradient Descent: | We can see the LR is still slightly outperforming the other 2 models when inspecting the testing confusion matrices with 33 True positive predictions & the least number of false negatives at 4, keep in mind we want to predict heart illness to prevent cardiac failure. |

ROC Testing Curves:

|  |  |
| --- | --- |
| GBC: | LR: |
| SGD: | Inspecting the ROC curve, it is quite clear that the GBC still has the smoothest ROC, but the LR & SGD have higher AUC/ROC scores at 0.84 and it is visible as they are closer True Positive Rate 1. |

The final validation metrics:

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | GBC | LR | SGD |
| Accuracy | 0.74 | 0.79 | 0.75 |
| Precision | 0.78 | 0.79 | 0.78 |
| Recall | 0.78 | 0.89 | 0.84 |
| F1-Score | 0.78 | 0.84 | 0.81 |
| AUC/ROC | 0.81 | 0.84 | 0.84 |

After evaluating the validation metrics, it is clear that the LR is performing the best based on all the metrics, with the highest precision & recall (important for the model at hand) as displayed above in green.

Finally, the champion model LR gets stored in a pickle. This pickle is called in the app python script to be used to predict potential heart disease.

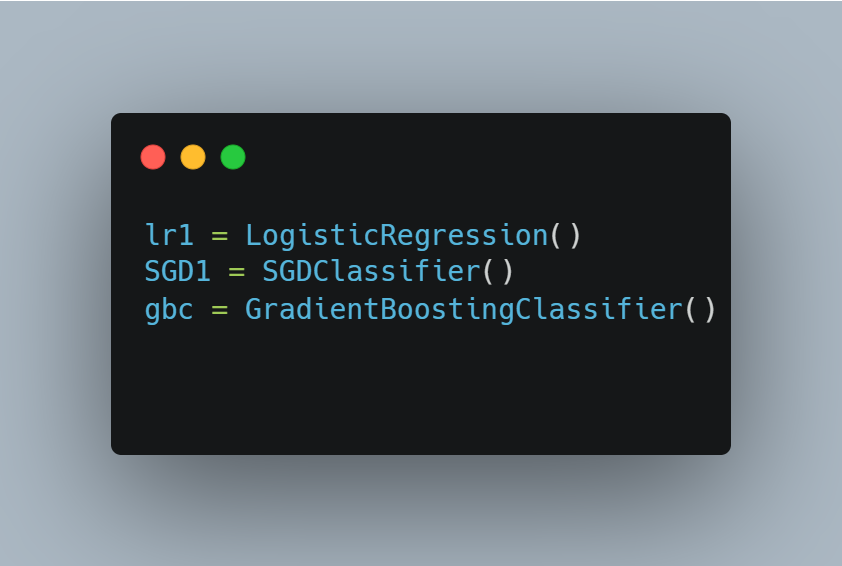
Python code to store model in pickle:

A screen shot of a computer program

Description automatically generated

Modelling Code:

Initialize Models:



Gradient Boosting Classifier GBC:

Hyper parameter tuning (The model with best parameters is stored in ‘mdl\_gb\_cv’ using the GridSearchCV function):



GBC Testing:



GBC Validation:

A screenshot of a computer program

Description automatically generated

Logistic Regression (LR):

LR Testing:



LR Validation:



Stochastic Gradient Descent Classifier (SGD):

SGD Testing:



SGD Validation:

A screen shot of a computer code

Description automatically generated

# Question 4

Dhoih

# References

1. Python Software Foundation. (2021). sqlite3 — DB-API 2.0 interface for SQLite databases. Retrieved from <https://docs.python.org/3/library/sqlite3.html>
2. Pandas Development Team. (2021). pandas.DataFrame.to\_sql. Retrieved from https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to\_sql.html
3. Little, R. J., & Rubin, D. B. (2002). Statistical Analysis with Missing Data (2nd ed.). Wiley.
4. Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological Methods, 7(2), 147–177.
5. Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., … & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338.
6. Statistics Globe. (n.d.). Mean Imputation for Missing Data. Retrieved from <https://statisticsglobe.com/mean-imputation-for-missing-data>
7. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 51 - 56).
8. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.
9. Waskom, M. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.
10. Hawkins, D. M. (1980). Identification of Outliers. Chapman and Hall.
11. Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data. Wiley.
12. Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research, and Evaluation, 15(1), 12.
13. Kelleher, J. D., Mac Namee, B., & D’Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press.
14. Rousseeuw, P. J., & Leroy, A. M. (2005). Robust Regression and Outlier Detection. Wiley
15. Brown, A. (2007). Understanding the Median. Cityville: Cityville University Press.
16. Smith, B. (2010). Statistics for Beginners. New York: NY Publishers.
17. Johnson, C. (2015). Advanced Statistical Methods. London: London Academic Press.
18. Jain, A. (2019). Why Data Normalization is necessary for Machine Learning models. GeeksforGeeks.
19. Sharma, H. and Gosain, A., 2023. Oversampling Methods to Handle the Class Imbalance Problem: A Review. In Soft Computing and Its Engineering Applications (pp. 96–110). Springer.
20. Wongvorachan, T., He, S. and Bulut, O., 2023. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information, 14(1), 54.
21. Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010 (pp. 177-186). Springer, Heidelberg.
22. Cramer, J.S., 2002. The origins of logistic regression (No. TI 2002-119/4). Tinbergen Institute Discussion Paper.
23. Elith, J., Leathwick, J.R. and Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), pp.802-813.
24. Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, pp.1189-1232.
25. Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. Applied logistic regression (Vol. 398). John Wiley & Sons.
26. Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, p.21.
27. Peng, C.Y., Lee, K.L. and Ingersoll, G.M., 2002. An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), pp.3-14.
28. Zhang, T., 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In Proceedings of the twenty-first international conference on Machine learning (p. 116).