**COS7045-B ADVANCED MACHINE LEARNING**

**Table of Contents**

[Task A 3](#_Toc140683551)

[Name and description data set 3](#_Toc140683552)

[Literature review 4](#_Toc140683553)

[Description of data mining questions and expected outcomes 6](#_Toc140683554)

[Task B 7](#_Toc140683555)

[Statistical analysis and visualisation, correlation analysis and normalisation (if appropriate) 7](#_Toc140683556)

[Analysis and discussion 19](#_Toc140683557)

[Task C 19](#_Toc140683558)

[Description and justification of algorithm, tool and process choice 19](#_Toc140683559)

[Description of practical application (e.g. algorithms for preprocessing, mining and/or modelling, evaluation) 23](#_Toc140683560)

[Results and discussion 23](#_Toc140683561)

[Task D 24](#_Toc140683562)

[Legal, Social, Ethical, Professional Issues 24](#_Toc140683563)

[Conclusions and discussion 24](#_Toc140683564)

[Comparison with relevant academic publications and results 25](#_Toc140683565)

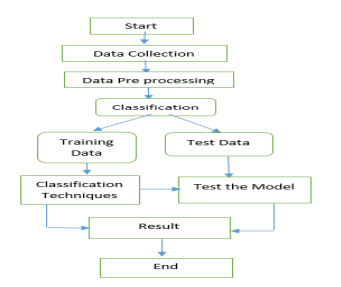
[Reference list 27](#_Toc140683566)

# Task A

## Name and description data set

The dataset known as “heart” has been selected from the following link: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset that describes the diagnosis of human health using different tests, observations and components. The average age in years for the participants in the dataset is around 54.36 with a variance and standard deviation of 82.98790698 and 9.08210099 respectively. The dataset consists of 14 rows and 304 columns with the following attributes:

|  |
| --- |
| “Age” = shows the ages of the participants with an average pof 54 years  “Sex” = shows the gender of the participants were 0 is for female and 1 is for male  “chest pain type (4 values)-cp” = describes the chest pains of the participants from 0 to 3  “resting blood pressure-trtbps” = describes the blood pressure of the participants at rest from 94 to 200.  “serum cholesterol in mg/dl-chol” = records the cholesterol levels of the participant starting from 126 to 564.  “fasting blood sugar > 120 mg/dl-fbs” = records the blood sugar level of the participanst after fasting from 0 to 1  “resting electrocardiographic results (values 0,1,2)-rectecg” = provides the electrocardiographic results during rest from 0 to 2  “maximum heart rate achieved-thalachh” = provides the heart rate of the participant for maximum values from 71 to 202  “exercise induced angina-exng” = values from 0 to 1  “oldpeak = ST depression induced by exercise relative to rest” = values from 0 to 6.2  “the slope of the peak exercise ST segment-slp” = values from 0 to 2  “number of major vessels (0-3) colored by flourosopy-caa” = values from 0 to 4  “thal: 0 = normal; 1 = fixed defect; 2 = reversable defect-thall” = values from 0 to 3  “The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.-output” = values from 0 to 1 |



**Figure 1: Model used for predicting heart disease**

(Source: Hasan *et al.* 2022)

The data is collected from four databases which are Hungary, Switzerland, Long Beach V and Cleveland where the “target” field refers to the presence of “heart disease" among the participants in which the “0” stands for no disease and “1” stands for heart disease. All of the variable types are in numeric form.

## Literature review

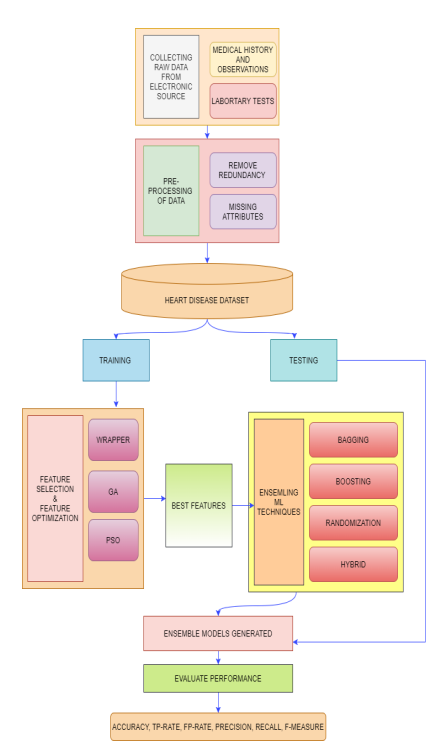
According to Hasan *et al.* (2022), after data collection the pre-processing step is vital for identifying the variable type, number of rows and columns, missing values, and spell errors. It is important to understand that if the data is not evaluated properly before analysis then the prediction power of the model would deteriorate with lower accuracy models. It is evident from previously done research studies that the heart disease prediction model is a problem related to classification and clustering which would be reduced to small attributes for better accuracy. Decision trees are often used in such instances since they are able to mimic human thinking processes which makes the results easy to understand. Even the entire structure is like a tree where each of the characteristics is analysed and used for the calculation. The formula followed by iot is as follow:

“Information gain= Entropy(S) – [(Weighted avg)\*Entropy (each feature)]”

In this equation entropy calculates the impurity present within the selected attribute. It can be further measured as:

“Entropy(S) = -P(yes)log 2 P(yes) – P(no)log 2 p(no)”

In the above equation S denotes the “Total no of samples”, P(yes) denotes the “Probability of yes” and P(no) denotes the “Probability of no”.



**Figure 2: Proposed method**

(Source: Jasjit and Singh, 2022)

In the research done by Jasjit and Singh(2022), it can be seen that they reviewed the existing models to identify the key attributes involved based on which a novel model would be made. The feature collection makes it imminent for the researchers to be applied to the attributes. Different types of Ml techniques they applied for heart disease prediction that would help in providing new insights and correlations among the variables. The data mining tool is very crucial while working with datasets for applying machine learning. They used a similar dataset that showed different features using PSO (“Particle Swarm Optimization”).

Logistic regression models (1/1+exp(-1)) defines the parameters creating a model for predicting the output in binary which makes it suitable for disease predictions. According to a research done for predicting tumour for determining breast cancer logistic regression was applied on the dataset to predict tumour values where the weighted sum was modified to “logistic function” (Talha *et al.* 2020). SVM and KNN were applied to illustrate the receiver operating characteristics (ROC) for justifying multi-class “classifier performance”. This would also help in visualising the data to determine the participants class to be with disease or no disease. The accuracy results would show the efficiency of the model for predicting the values that would help in determining the type of algorithm that was suitable for justifying the research model with highest accuracy like K-NN confusion matrix (Talha *et al.* 2020).

## Description of data mining questions and expected outcomes

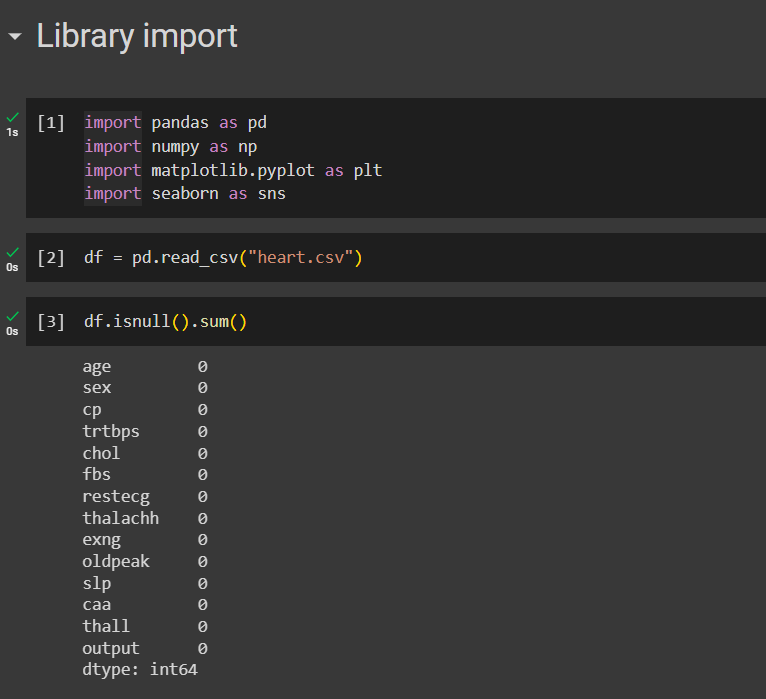
The goal of this study is to create a predictive model for heart disease using data from Kaggle's "heart" dataset. Each row in the dataset represents a participant, and the 304 columns in it contain various parameters including age, gender, the type of chest discomfort experienced, blood pressure, cholesterol levels, and other heart health-related information.

1. How effective are decision tree-based models at predicting heart disease when compared to other machine learning methods (such logistic regression, SVM, and KNN)?
2. What is the created heart disease predictive model's accuracy and effectiveness?
3. How can the suggested model help with early detection and comprehension of the variables affecting heart health?
4. Is it possible to accurately assess the impurity within attributes and identify their significance in predicting heart disease using the information gain calculation method?
5. What new information and correlations between variables may be gleaned from the "heart" dataset by using machine learning techniques?

The intended result of this study is the development of a precise and trustworthy predictive model that can identify the presence or absence of heart disease based on the provided characteristics. The model seeks to uncover important characteristics and relationships among factors that contribute to the prediction of heart disease by examining the dataset and utilizing machine learning techniques (Sami *et al.* 2021). Decision trees are useful for these kinds of prediction tasks, according to earlier research, since they can imitate human thought processes and produce outcomes that are easy to understand. In order to assess the impurity within attributes and ascertain their significance in the prediction of heart disease, the research will investigate the usage of decision trees and the computation of information gain. The study will also examine the performance of various machine learning methods like logistic regression, SVM, and KNN in predicting heart disease (Talha *et al.* 2020). The effectiveness and accuracy of various models will be assessed, and the best method will be chosen depending on the level of accuracy attained.

# Task B

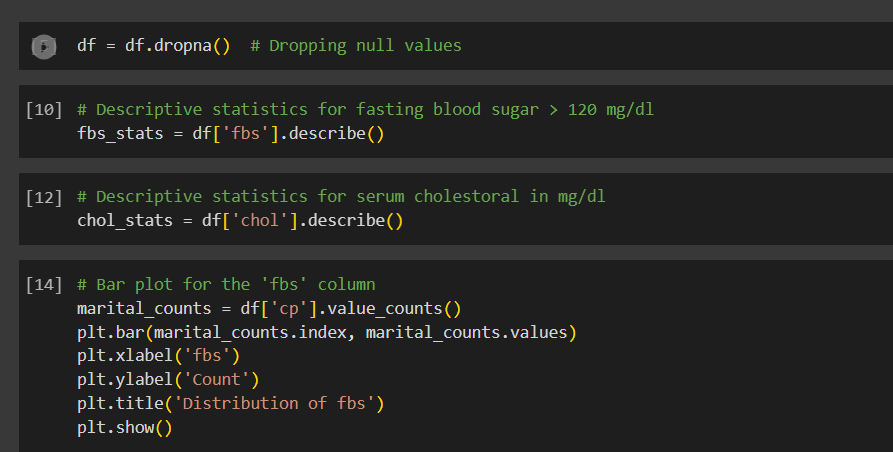
## Statistical analysis and visualisation, correlation analysis and normalisation (if appropriate)

****

**Figure 3: Importing and loading the dataset and checking the null value**

(Source: Created In Python)

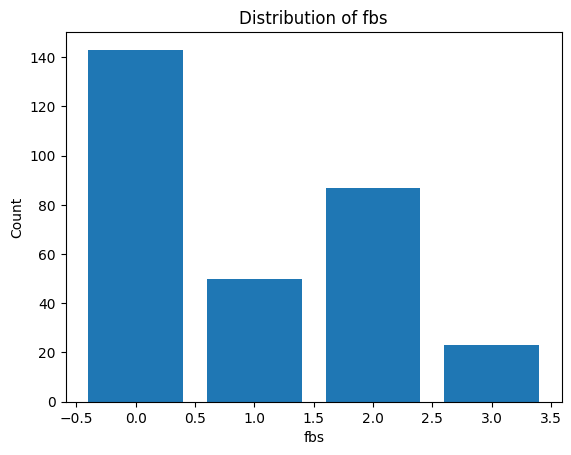
In the above figure, the initial process of the data analysis has been the importing of the essential libraries. Each library that is imported has its own functions and applications that are used for the different types of data analysis and modifications or data mining procedures. The pandas library that is used in this Python programming is for the data manipulation and data extraction by which the dataset uploaded in the python framework can be evaluated. Numpy library is used for the mathematical calculations and storing the number in an array form for performing different types of mathematical calculations.The matplot library is used to plot the diagrams and to present the data in the pictorial form so that the data calculations and numeric variables can be well understood by the visualisation of data. Seaborn library is used for the advanced mathematical calculations that figure out the standard deviation , mean and max values. The second coding part explores the understanding of the loading of a dataset and producing the data variables in a new dataframe. The third part of the coding relates to the understanding of any null values present in the data and checking the sum value of the null values stored in the columns and that is resulted as 0.

****

**Figure 4 : Code for Dropping null values and plotting fasting blood sugar bar chart**

(Source: Created In Python)

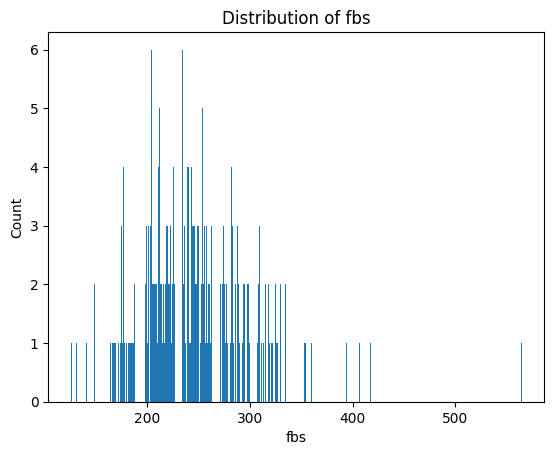
In the above figure, the code that has been written in the initial lines is for dropping the NA values from the dataset. The NA values in the dataset can lead to the different errors and execution and may figure out the wrong predictions and visualisations of data, so this code has been written to remove those null values from the dataset. In the second part of the line the code has been written to describe the statistics for the fasting blood sugar.In the third line of code the descriptive statistics for serum cholesterol is written that can produce a list of statistical values. The 14th line of code is written for generating the barplot for the ‘fbs’ column which recipients and explores the information of the fasting blood sugar for the marital people

****

**Figure 5 : Fasting Blood sugar bar chart visualisation**

(Source: Created In Python)

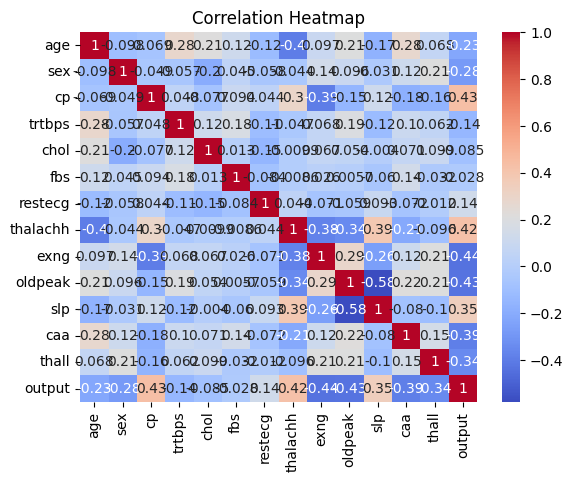
In the above figure, the bar chart visualisation has been done for the fasting blood sugar which shows the distribution of the chest pain type in different intervals. The highest and the lower range bar chart values are visualised and the counts in the different intervals are highlighted with blue colour bar.

****

**Figure 6 : Bar Chart Plotting for cholesterol counts detected with fasting blood sugar**

(Source: Created In Python)

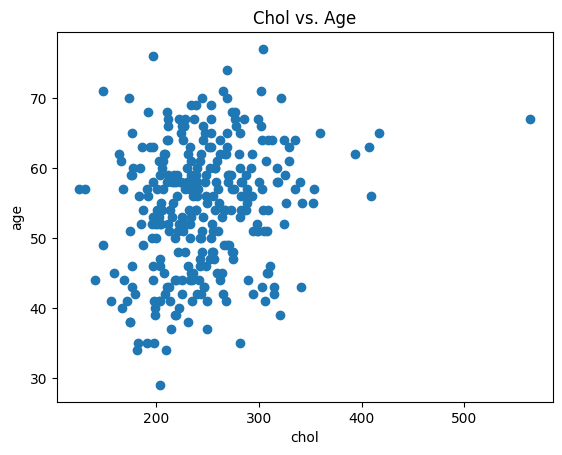
In the above figure, the bar chart plotting has been done to evaluate the distribution of fbs for the cholesterol counts. The bar chart is plotted into x y graph which shows the cholesterol level in x axis and the counts of the cholesterol in the y axis.

****

**Figure 7 : Plotting Correlation Heatmap**

(Source: Created In Python)

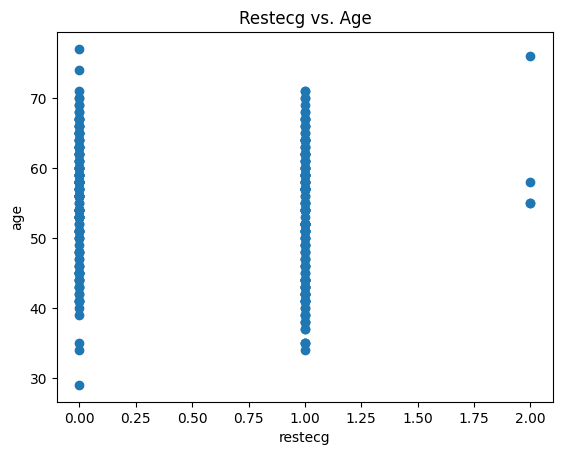
In the above figure, the correlation heatmap is generated for the heart.csv datasets and its respective column variables. This correlation map feature is to evaluate and visualise the connection and mutual relationship among the variables and to evaluate the strength of the different column variables relationship. The outliers can also be detected using the correlation heatmap along with linear and nonlinear connections among the variables can also be figured out. It generates the relations of data in a matrix form so that the data can be summarised in an effective manner by which program can solve the mathematical calauctions efficient and understand the relationship.

****

**Figure 8 : Scattered Plotting of Cholesterol and Age Column**

(Source: Created In Python)

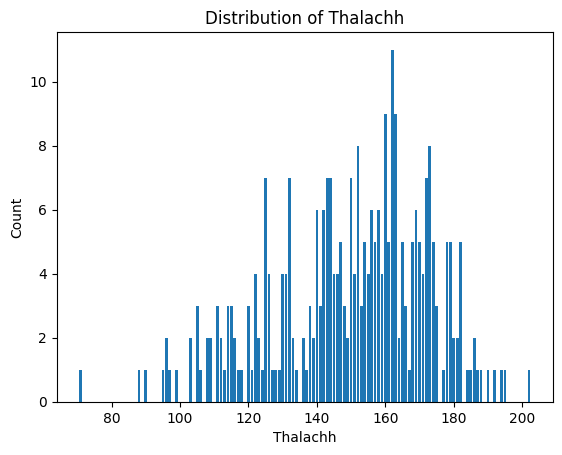
In the above figure , the scattered plot is done for the ‘chol’ and age column which shows the relationship among the variables between the two columns. The scattered dots that are closer indicates there is a strong relationship between the variables and the scattered dots that are outside means some data are there between the columns that are not relatable .

****

**Figure 9 : Scatter plotting of Rest Ecg vs Age**

(Source: Created In Python)

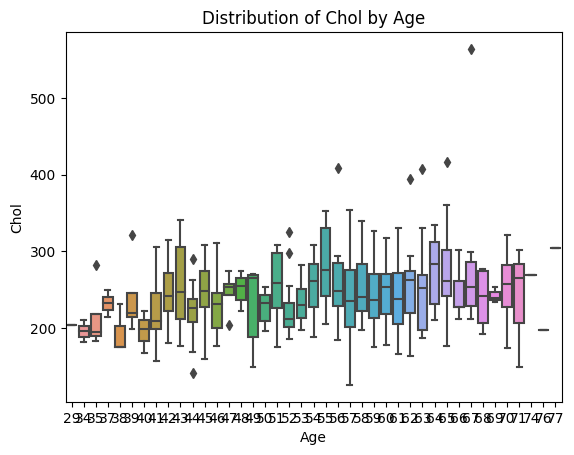
In the above figure, a different visualisation has been done by the process of scatter plotting for the two column that is ‘rest ecg’ and ‘Age’ and determines the relation between the values of the two columns that is mentioned in the x and y axis scatter plot graph .

****

**Figure 10: Bar Plotting of the thalachh column exploring maximum heart rate**

(Source: Created In Python)

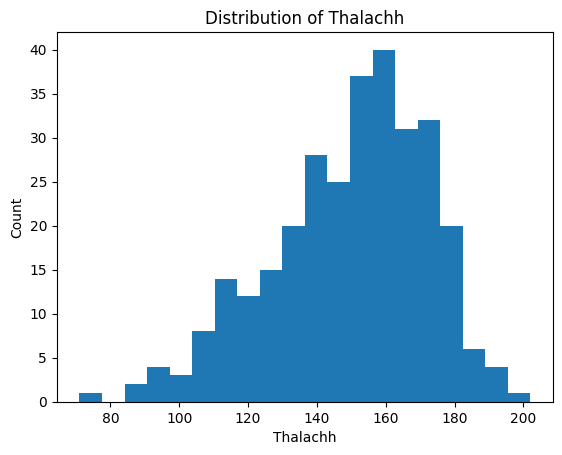
In the above figure, the bar plotting graph is illustrated and visualised which depicts the maximum heart rate counts for the thalachh column that is selected for addressing the maximum hearts rate diagnosis of the people mentioned in the different column of the respective dataset. The bar plotting is done in the x and y graph where the x axis evaluates the highest or maximum heart rate and the y axis depicts the number of heart rates depicted.

****

**Figure 11 : Box Plotting of Age By Cholesterol**

(Source: Created In Python)

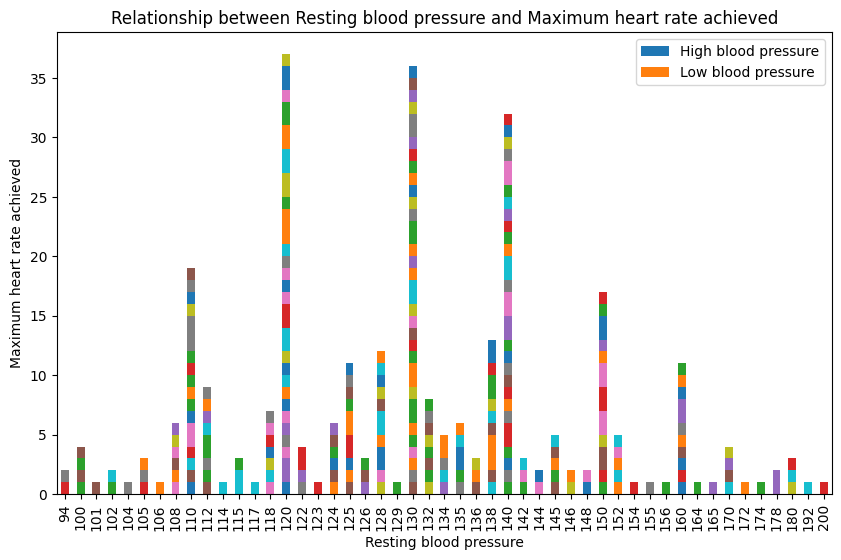
In the above figure, the mixed colour box plotting is done which shows a different form of visual representation of data. The 10 colour combination box plotting enhances the productivity of the data and the programmer can easily distribute the values of the selected columns and differentiate the counts of the selected data.

****

**Figure 12 : Histogram of maximum heart rate achieved**

(Source: Created In Python)

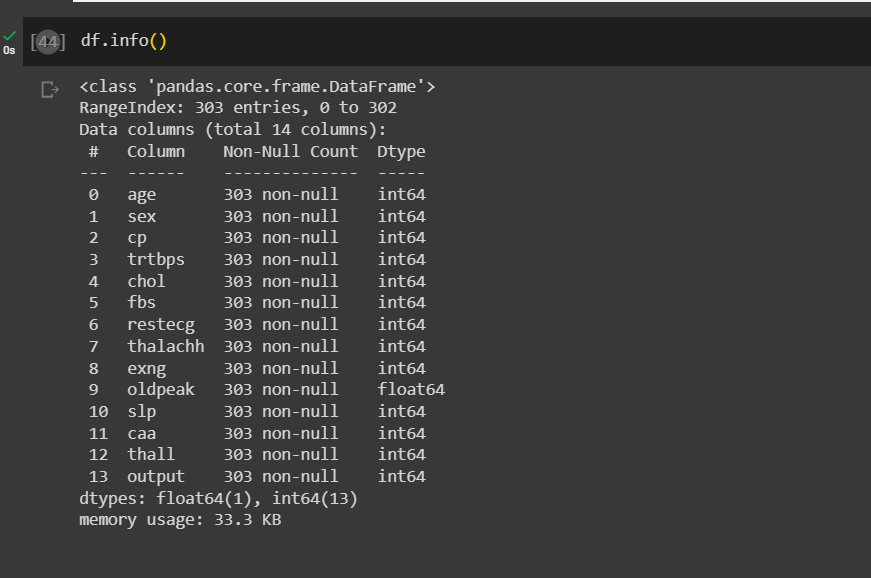
In the above figure, the histogram data is evaluated which visualises the statistical data into a visual form for the selected thalachh column that is related to the maximum heart rate received for the selected persons mentioned in the other column variables.

****

**Figure 13 : Stacked Bar plotting of ‘trtbps’ and ‘thalachh’**

(Source: Created In Python)

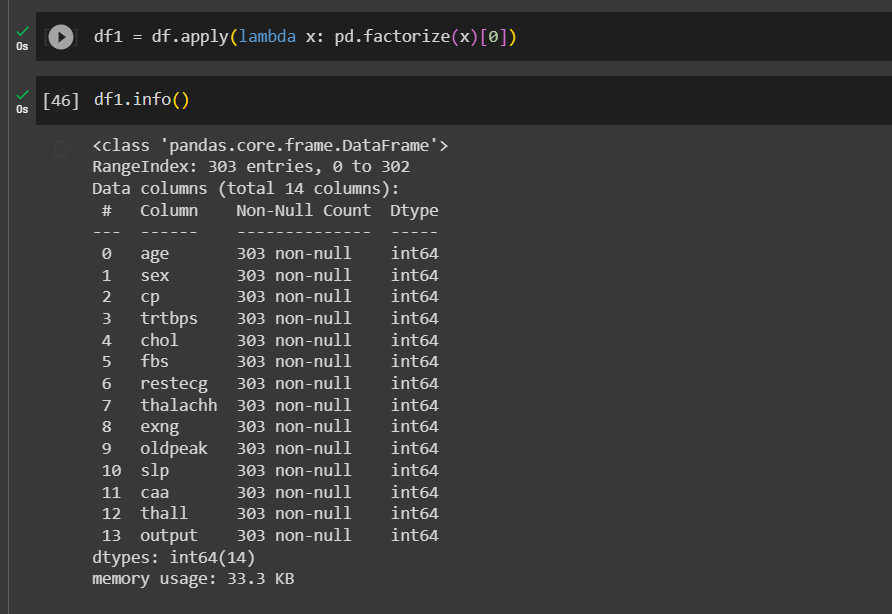
In the above figure , the stacked bar plotting is done for the two columns that are ‘trtbps’ and ‘thalachh’ which depicts the understanding of the visual representation of the data varying at each interval. This stacked bar chart evaluates the relation between the resting blood pressure and the maximum heart rate achieved. The stacked bar chart is distributed in the x and y axis. The x axis represents the blood pressure variation in respect to the maximum heart rates that is changing at different intervals. The y axis represents the number counts of the resting blood pressure which is starting from 94 and ending up to 200.

****

**Figure 14 : Dataframe Info**

(Source: Created In Python)

In the above figure, the dataframe info is displayed. This dataframe info is displayed and generated after loading the datasets into a new dataframe where the information of the dat column and its respected variable values are displayed for the data mining process. The dataframe info is an important step for the data mining and data preprocessing step that explore the information about the non null count and datatypes of the column variables. The above figure determines the maximum number of integer data types for each of the respective columns except the ‘oldpeak’ column that is a float data type. The data type is very important for the further data analysis and the application of models. The number of integer data type is 13 and the float data type is 1.

****

**Figure 15 : Converting categorical data into integer data**

(Source: Created In Python)

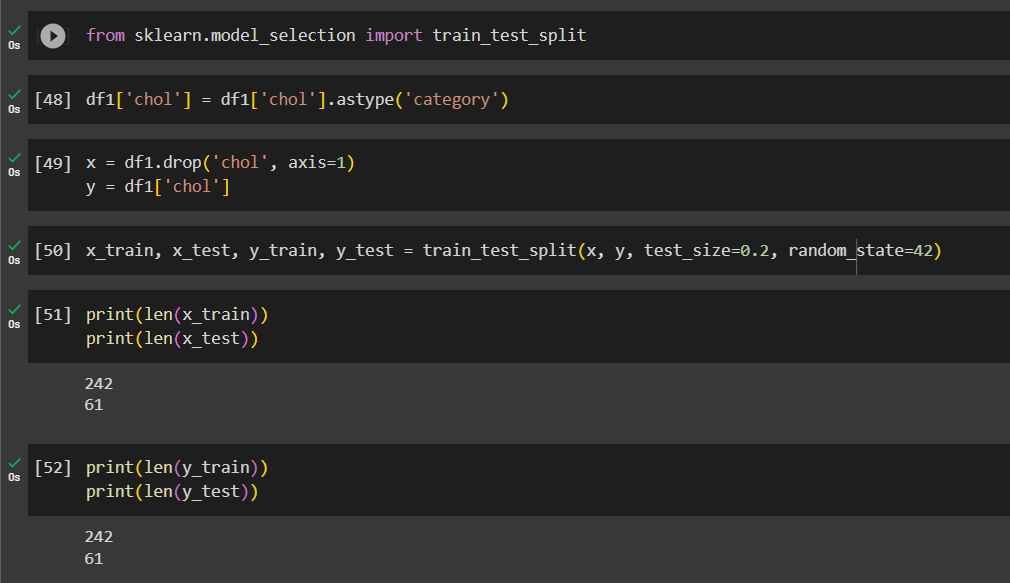
In the above figure, the lambda vectorizer is used and a new dataframe is created where the categorical columns of data are changed into integer data type so that the application of models can be fitted well without any error or complexities in the data execution. This converting of data can be a major reason for the predicting the accuracy of the model correctly without any errors

## Analysis and discussion

In this portion of the project report, deep learning methods have been used which includes the loading of the dataease, data mining , data preprocessing and data visualisation. The loading of the datasets includes the important process in the data mining process , the data preprocessing involves the advanced mathematical calculation that is used to figure out the important data calculation by which the authenticity of the dataset can be monitored and executed. The data visualisation is one of the major processes by which the user can get attracted and interested in the proceeds of data analysis as it clears the understanding by presenting a visual based information of data with charts and figures.

# Task C

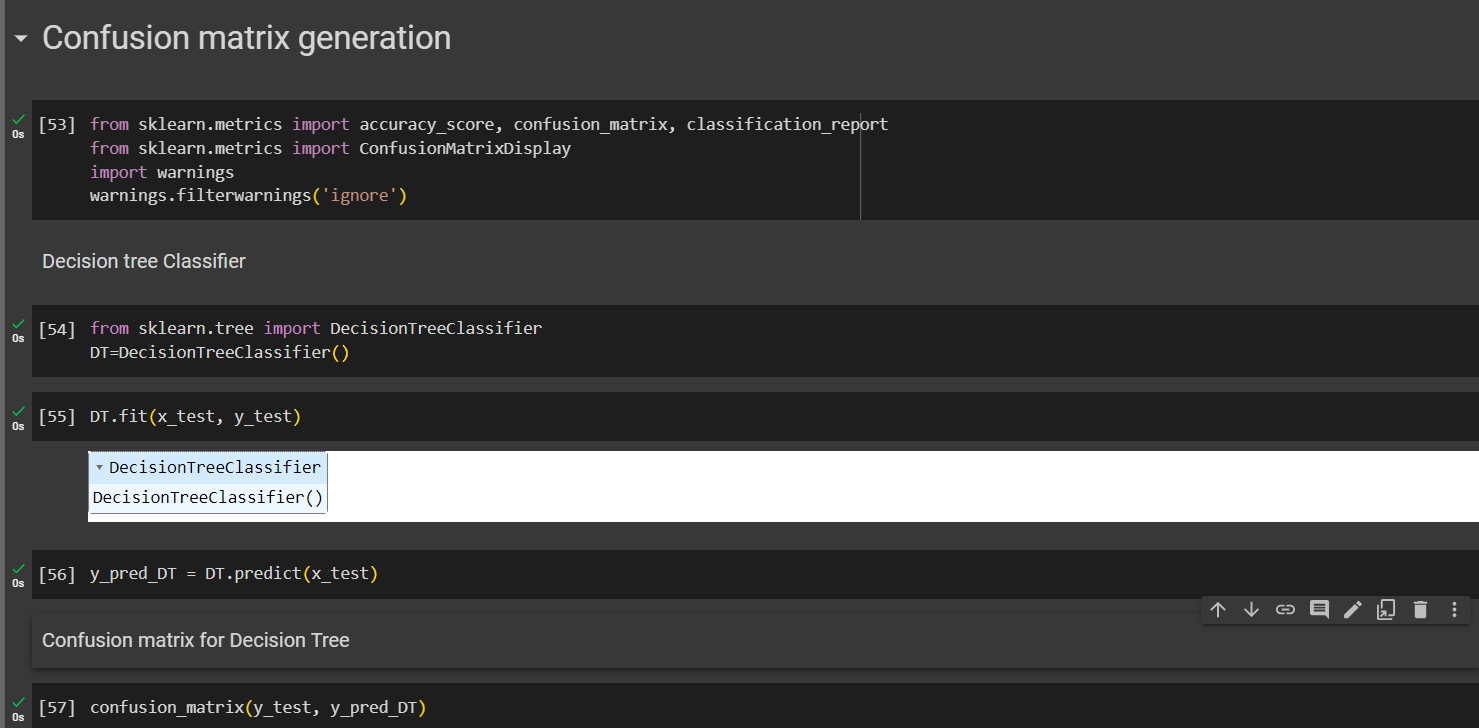
## Description and justification of algorithm, tool and process choice

****

**Figure 16 : Splitting the dataset by training and testing the data**

(Source: Created In Python)

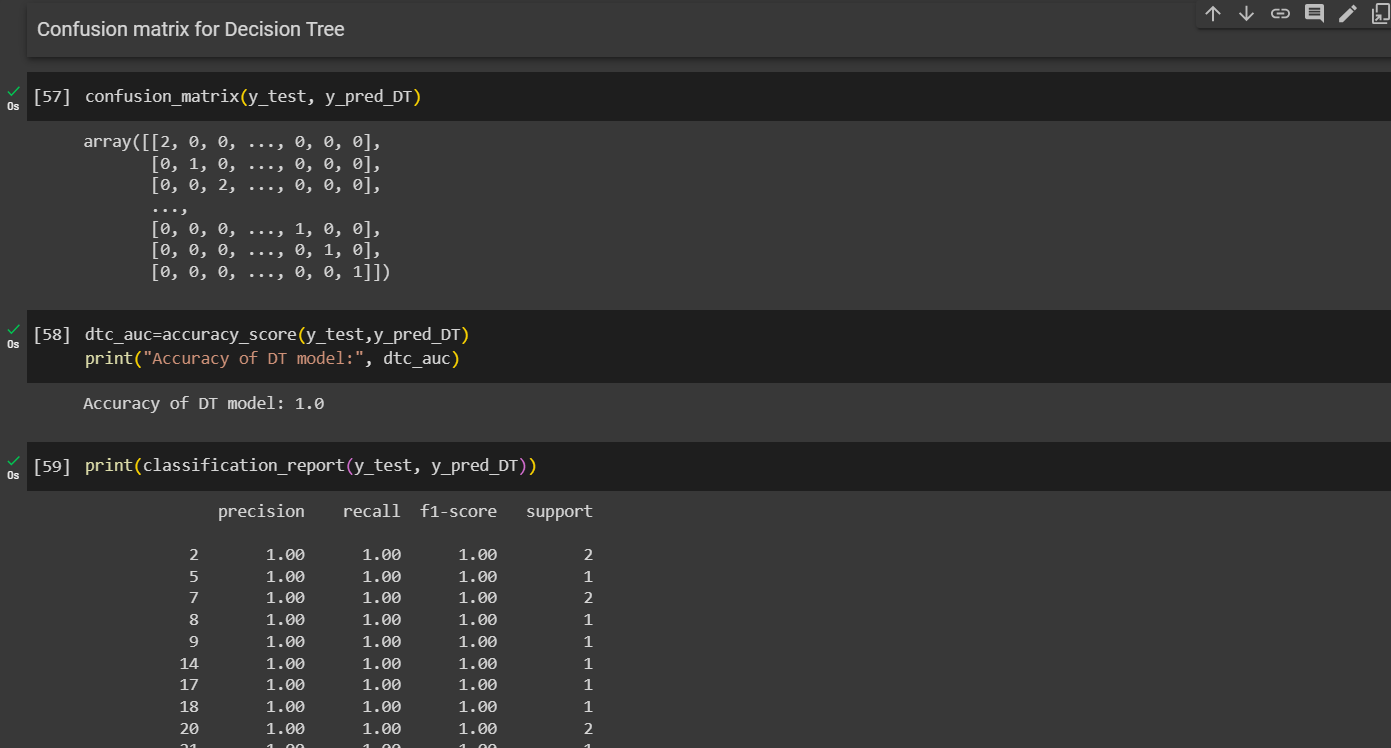
In the above figure, the splitting of the dataset has been done which includes one column named ‘chol’ from the dataset and is tested and trained which modifies the data and genetteeds a test and trained values that can be used for the application of models.

****

**Figure 17 : Application of decision tree classifier**

(Source: Created In Python)

In the above figure, the decision tree classifier model is used after importing the necessary libraries for the classification report and confusion matrix. At first the decision tree model is applied and fitted successfully.

****

**Figure 18 : Confusion matrix and prediction details of decision tree**

(Source: Created In Python)

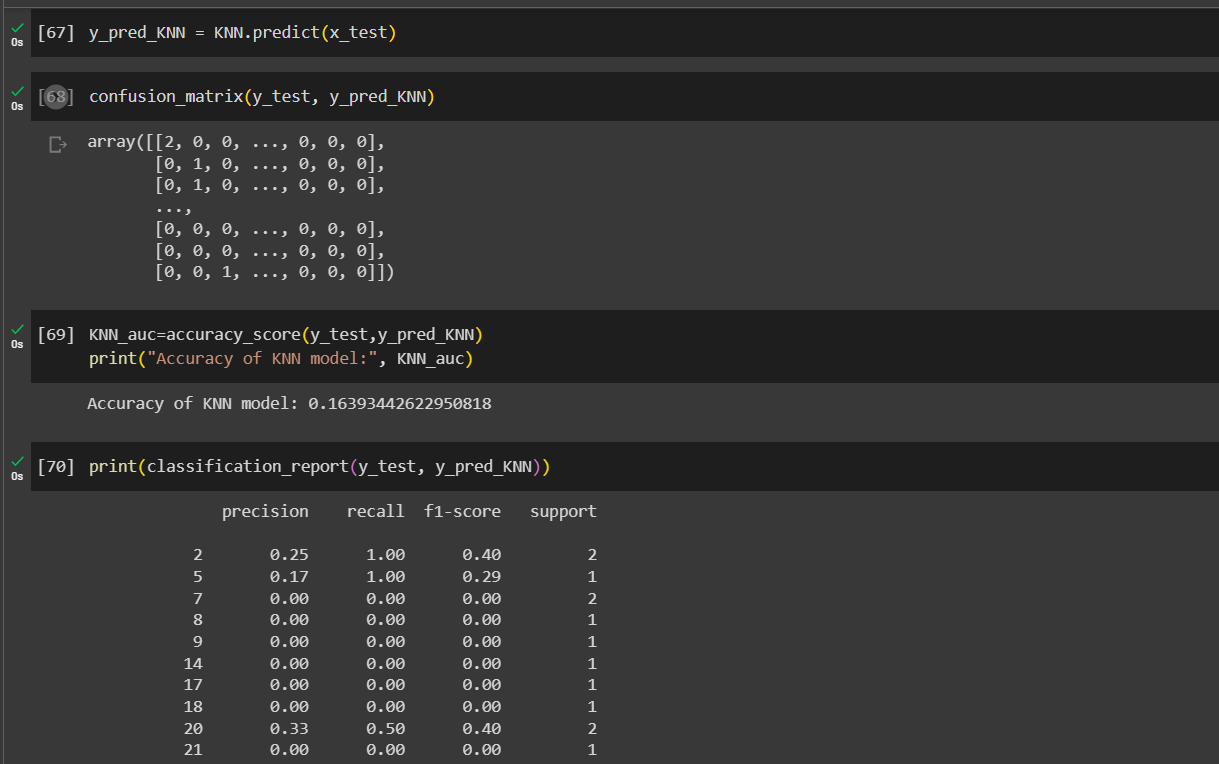
In the above figure, the confusion matrix and the prediction details of the decision tree classifier is evaluated which explores the understanding that the prediction is done in an efficient and sequential process. The confusion matrix is generated in an array format which stores a decrementing series, in the next part the decision tree model accuracy is evaluated at 100 percent. The classification report is also detailed which shows the different porediction that includes precision, recall , F1-score, and support.

****

**Figure 19 : Application and Prediction details of Linear regression model**

(Source: Created In Python)

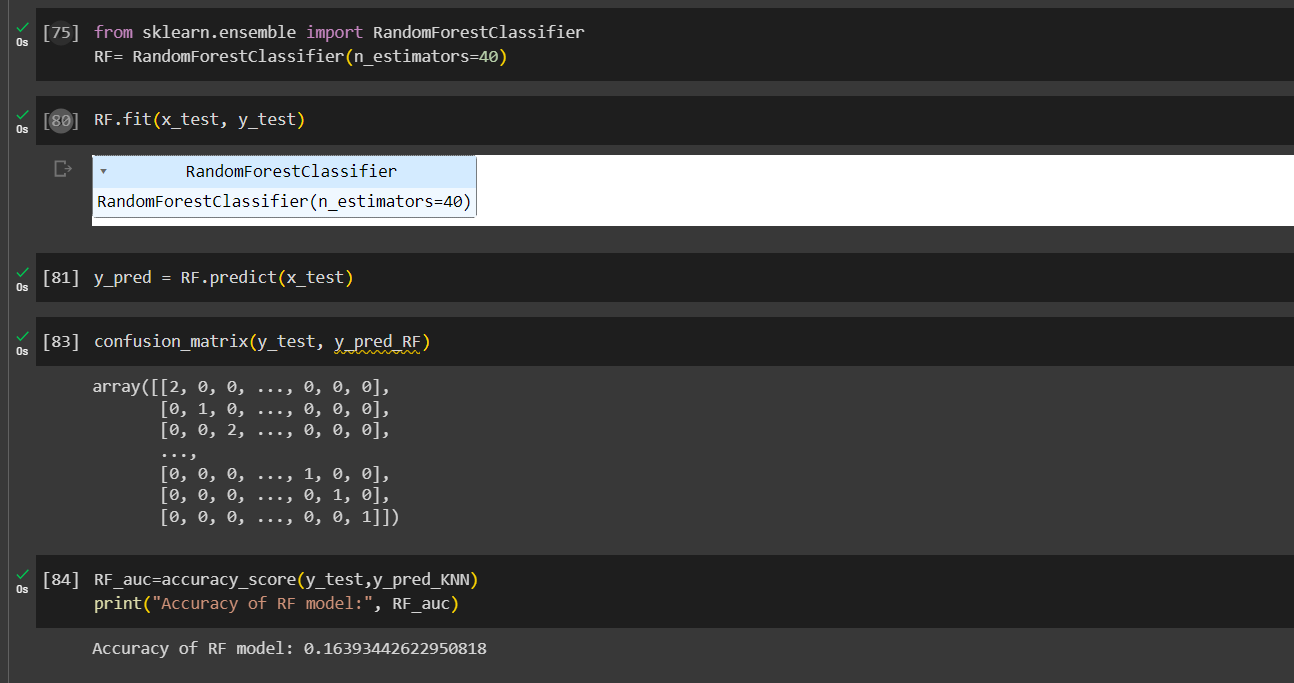
In the above figure, the application of linear regression model is being done after the importing of random forest classifier library which enables the linear regression model to be fitted successfully. The application of linear regression model is 26 percent which is comparably low than that of the decision tree classifier and that proves the decision tree classifier is efficient compared to the linear regression model that has been fitted.

****

**Figure 20 : Application and prediction details of KNN model algorithm**

(Source: Created In Python)

In the above figure, the application and prediction details of the KNN algorithm are obtained and evaluated. At first the KNN model has been fitted successfully , then the confusion matrix is generated accordingly. The accuracy of the KNN algorithm model that has been evaluated is 16 percent which is comparably lower than the decision tree classifier and linear regression model. The classification report of the KNN algorithm model is also predicted which shows the different values for the precision, recall, F1-score and support.

****

**Figure 21 : Application and prediction details of Random Forest Classifier**

(Source: Created In Python)

In the above figure, the application and the prediction details of the random forest classifier is done after importing the necessary random forest classifier. Then the model has been fitted successfully and explored and in the next process the confusion matrix is generated. The accuracy of the model that is evaluated is 16 percent.

## Description of practical application (e.g. algorithms for preprocessing, mining and/or modelling, evaluation)

The practical applications that are done are related to the steps that are done in the process of advanced machine learning which analyses the heart csv data that includes the multiple information of the data execution and data mining. The algorithms for preprocessing, mining and the modelling helps in the understanding of the dataset execution and finding the necessary calculations that can aid in the understanding of the dataset calculation and evaluation. The application of models is implemented and utilised to find out the correct predictions that can be used in the diagnosis of the people's heart condition at the different time intervals.

## Results and discussion

The decision tree models , KNN algorithm model, linear regression model and random forest classifier are the four models that are used in this project to check the utilities and figure out the accuracies of the model to check which model has the highest utility. The accuracy score of the model that is evaluated and figured out helps us to check which model can be utilised in the analysis of the heart related disease. The prediction classification based on the classification report can summarise and sort the niche measurements that can be done to evaluate the predictive accuracy in the terms of precision, recall, F1-score and support values.

# Task D

## Legal, Social, Ethical, Professional Issues

The legal social ethical and professional issues that a society accepts and utilises due to the integrated machine learning and artificial intelligence privacy and security, based on the discrimination and potential, can be a measured challenge in the judgement of people using the advanced machine learning process(Coskun et al. 2022). Integration of digital technology resources of inauguration and the data leak that have risen most probably in the field of analysing the data by the efficient tools. The security and databases that have reason higher in the recent days of Healthcare management also face deadly consequences that can be a major cause of the victim for the error in the healthcare reports(Majidi et al. 2022). The patients can face major problems if their reports are not evaluated and diagnosed properly. The healthcare management can face a trust issue because of the irregular re;orts and the improper treatment that is done on the wrong reports that is evaluated by the advanced machine learning based software which detect the root cause of the problems. The ECG machines algorithm should be properly coded and utilises that can measure the frequency of the heart rates and help the medical professionals to properly diagnose the heart as it is a matter of life and death (Balachandran et al. 2023). So the proper concern of health issues can strike the officials or healthcare management to utilise the tools or machine learning algorithm that evaluates the prediction with 100 percent accuracy or close to it. Likewise the decision tree classifier model that has been used in this project can be used in the healthcare departments because when the model is being predicted it shows the 100 percent accuracy and medical professionals can rely on the tools that gives 100 percent accuracy and proceed the diagnosis effectively without any complexities or fault. The healthcare facilities should provide a better facility for the patients as patients health can be affected or it can be worsened if in case of any vulnerability as different types of people staying in the hospitals have different problems so the transfer of disease to a low immune body patient can be fatal. At present, no clear guidelines exist to handle the legal and moral challenges that may develop as a result of the usage of machine learning in health care institutions(Sreemoyee et al. 2021). This evaluation seeks to address these essential concerns by emphasising the need of algorithmic openness, anonymity, and safeguarding for all individuals engaged, as well as counterintelligence of connected risks.

## Conclusions and discussion

The heart.csv that was provided has been analysed by the Python programming language is covered and analysed in this project report. The dataset has been analysed with several steps which includes the different algorithms, model applications and visualisations that can help the user to understand the dataset analysis utility and to extract the beneficial information that can be helpful in the healthcare management just like the heart data that has been analysed in which the possible charts and figure helps the user to understand the outcomes that can be helpful in the evaluation of the healthcare diagnosis possibility. Each model frequency is checked and evaluated so that the best possible model can be used after sorting the different model frequencies. The hyperparameters used for the algorithms can sort out the best performance. The classification report is the major procedure that can aid in the efficiency of the models and its applicability in the tools and machines that acquire the logarithms used in the machine learning process. The different logical algorithms can be used and checked to utilise the model efficiency which can be added and evaluated in the machines that use different precision accuracy which can be helpful in the diagnosis of health and other cardiovascular diseases. The data analysis that has been done is a major process for evaluating and checking the processes of machine learning that can help in the assessment of tools and technologies that are used in the healthcare management.

## Comparison with relevant academic publications and results

The project report also emphasises the importance of coordinated attempts to collect large datasets containing a variety of chronic illnesses and cardiovascular health indicators. The study advocates for the development of accessible machine learning techniques in order to better understand the basic mechanisms behind cardiovascular illness and stroke. The paper examines issues with the dataset and its context from legal, social, professional, and ethical perspectives. It emphasises the importance of data confidentiality and the need for privacy-protecting processes when working with personal health data. It also discusses potential biases in the dataset and how they may impact the fairness and generalizability of the models. Furthermore, the work emphasises the need of adhering to privacy laws and norms, as well as the ethical requirements that need to be focused on to overcome the issues that are faced in the healthcare sector and its management. The relevant papers and publications that are mentioned define the uses and utility of the softwares and tools like the advanced machine learning procedures in order to sustain the healthcare management and the diagnosis done by it. But the critical review part that is discussed focus on the challenges and drawbacks that should be monitored well in the healthcare management because the lives of the patients is a serious issue that should be focused on and the model in the machine learning process can be utilised that can predict the accuracy with 100 percent.

# Reference list

**Journals**

Hasan Sazzad Iqbal, S.M., Jahan, N., Afroja, S.M. and Khatun, M. 2022, "An Effective Analytics and Performance Measurement of different Machine Learning Algorithms for Predicting Heart Disease", International Journal of Advanced Computer Science and Applications, vol. 13, no. 2.

Jasjit, S.S. and Singh, D. 2022, "A Machine Learning Model for Predicting Heart Disease using Ensemble Methods", International Journal of Advanced Computer Science and Applications, vol. 13, no. 9.

Sami, O., Elsheikh, Y. and Almasalha, F. 2021, "The Role of Data Pre-processing Techniques in Improving Machine Learning Accuracy for Predicting Coronary Heart Disease", International Journal of Advanced Computer Science and Applications, vol. 12, no. 6.

Talha, A.K., Kadir, K.A., Nasim, S., Alam, M., Shahid, Z. and Mazliham, M.S. 2020, "Proficiency Assessment of Machine Learning Classifiers: An Implementation for the Prognosis of Breast Tumor and Heart Disease Classification", International Journal of Advanced Computer Science and Applications, vol. 11, no. 11.

Navarrete, J.P., Pinto, J., Figueroa, R.L., Lagos, M.E., Zeng, Q. and Taramasco, C. 2022, "Supervised Learning Algorithm for Predicting Mortality Risk in Older Adults Using Cardiovascular Health Study Dataset", Applied Sciences, vol. 12, no. 22, pp. 11536.

Ozaltin, O., Coskun, O., Yeniay, O. and Subasi, A. 2022, "A Deep Learning Approach for Detecting Stroke from Brain CT Images Using OzNet", Bioengineering, vol. 9, no. 12, pp. 783.

Peyvandi, A., Majidi, B., Peyvandi, S. and Patra, J.C. 2022, "Privacy-preserving federated learning for scalable and high data quality computational-intelligence-as-a-service in Society 5.0", Multimedia Tools and Applications, vol. 81, no. 18, pp. 25029-25050.

Sadi, K.A. and Balachandran, W. 2023, "Prediction Model of Type 2 Diabetes Mellitus for Oman Prediabetes Patients Using Artificial Neural Network and Six Machine Learning Classifiers", Applied Sciences, vol. 13, no. 4, pp. 2344.

Sreemoyee, B., Nilay, K., Pragati, A. and Priyank, J. 2021, "Machine learning concepts for correlated Big Data privacy", Journal of Big Data, vol. 8, no. 1.