Cardiovascular Disease Prediction

MGSC-5126-10: Data Mining

Shannon School of Business, Cape Breton University

Dr. Ebrahim Sharifi

December 8, 2023

List of Figures

Figure 1. Percentage Count for Cardiovascular Disease Categories after Balancing...................... 6
Figure 2. Detecting and Removing Outliers 7
Figure 3. Exploratory Data Analysis 8
Figure 4. Correlation Matrix 10
Figure 5. Feature Importance – Decision Tree Model 10
Figure 6. Feature Selection – Chi – Square Statistic..11
Figure 7. Confusion Matrix –Logistic Regression 13
Figure 8. Evaluation Matrix for Logistic Regression 13
Figure 9. ROC Curve – Logistic Regression 13
Figure 10. Decision Tree Diagram with all features... 14
Figure 11. Confusion Matrix for Decision Tree Classifier...14
Figure 12. Evaluation Matrix for Decision Tree Classifier.. 15
Figure 13. ROC Curve – Decision Tree Classifier 15
Figure 14. Confusion Matrix – SVM 16
Figure 15. Evaluation Matrix for SVM...16
Figure 16. ROC Curve – SVM..16
Figure 17. Accuracy Trend for Different k Values.. 17
Figure 18. Confusion Matrix – KNN... 17
Figure 19. Evaluation Matrix for KNN...18
Figure 20. ROC Curve – KNN 18
Figure 21. Comparison of four Models...18
Figure 22. Accuracy and Precision of four Models 19
Figure 23. Actual Vs Predicted Values.. 19

INDEX

1. Introduction 4
   1.1. Overview of Dataset 4
   1.2. Metadata 4
2. Data Preparation 5
   2.1. Balancing the dataset 6
   2.2. Finding Outliers 6
   2.3. Handling Outliers 7
3. Exploratory Data Analysis 7
4. Feature Selection 8
   4.1. Correlation Matrix 9
   4.2. Decision Tree 10
   4.3. Chi – Square Statistic 10
   4.4. Data Understanding 11
5. Method Application 12
   5.1. Logistic Regression 12
   5.2. Decision Tree Classifier 14
   5.3. Support Vector Machine (SVM) Model 16
   5.4. K Neighbors Classifier 17
6. Model Evaluation 19
   6.1. Prediction for Unseen Data 20
7. Conclusion 20
8. References 22

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths (WHO, 2021). This project aims to utilize machine learning models to accurately predict the possibility of a person experiencing a cardiovascular disease in their lifetime. It seeks to identify major contributing factors that determine if a person would ever face the risk of experiencing the disease.

Given that the primary aim of the project involves only 2 possible outcomes, the project focuses on binary classification to identify disease risk in a person. It focuses on the following machine learning models: Logistic Regression, Decision Tree, Support Vector Machine and K Nearest Neighbor models.

1.1. Overview of Dataset

The dataset was sourced from Kaggle (link: https://tinyurl.com/4hchpuut), but originally was compiled during the time of medical examination. It contains 70,000 patient records, with 12 features. The binary target variable gives output 1 if the patient has the risk of experiencing a cardiovascular disease in the lifetime and 0 if otherwise.

1.2. Metadata

1. id: is a unique numeric identifier for each record on the dataset.
2. age: is the age of the individual in days and is a valuable metric in identifying the risk of disease.
3. gender: Indicates the gender of individuals either male-1 or female -2.
4. height: Height of individuals in centimeters, providing a quantitative measure of their vertical stature.
5. weight: Weight of individuals in kilograms, serving as a quantitative measure of their mass.
6. ap_hi: Systolic blood pressure, representing the pressure in the arteries during

the contraction of the heart muscle.

7. ap_lo: Diastolic blood pressure, representing the pressure in the arteries when the heart is at rest between beats.

8. cholesterol: Categorizes individuals into three levels: 1 for normal, 2 for above normal, and 3 for well above normal cholesterol levels.

9. gluc: Categorizes individuals into three levels: 1 for normal, 2 for above normal, and 3 for well above normal glucose levels.

10. smoke: Binary variable indicating whether an individual is a smoker (1) or a non-smoker (0).

11. alco: Binary variable indicating whether an individual consumes alcohol (1) or not (0).

12. active: Binary variable indicating whether an individual is physically active (1) or not (0)

13. cardio: Binary variable representing the presence (1) or absence (0) of cardiovascular disease in individuals. It serves as the target variable for binary classification.

2. Data Preparation

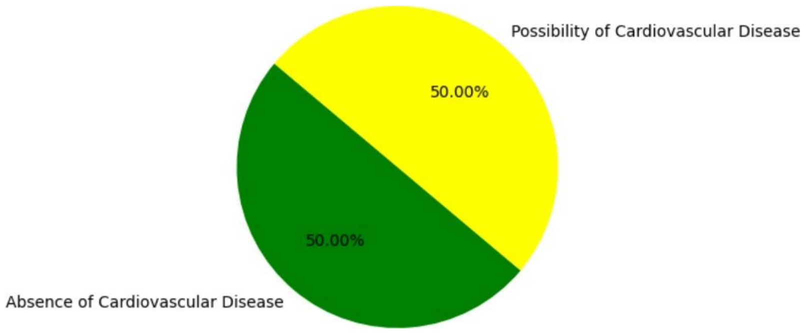
Data pre-processing includes data cleaning, normalization, transformation, feature extraction and selection, etc. (Kotsiantis et al., 2006). The process starts by importing necessary libraries from Python such as NumPy, pandas, matplotlib, seaborn, scipy, sklearn, joblib etc.

Data cleaning was initiated by dropping the feature “ID” as it is unique to each record and hence could not be a contributing factor and all the columns were renamed. Duplicate records were deleted and negative values in Systolic and Diastolic variables were removed. The resulting dataset had 66801 records and 12 variables. Feature extraction was performed on the variable “AGE” to extract the age in years instead of in days.

2.1. Balancing the dataset

The resulting dataset had a slightly larger number of records of patients who would face the risk of cardiovascular disease. Standard classifiers such as logistic regression, Support Vector Machine (SVM) and decision trees are suitable for balanced training sets. When facing imbalanced scenarios, these models often provide suboptimal classification results, i.e. a good coverage of the majority examples, whereas the minority examples are distorted (Lane et al., 2012). Hence the data was balanced using Random Under Sampling, where records from the majority class are randomly deleted until the number of records in both classes are equal.

Figure 1. *Percentage Count for Cardiovascular Disease Categories after Balancing*



2.2. Finding Outliers

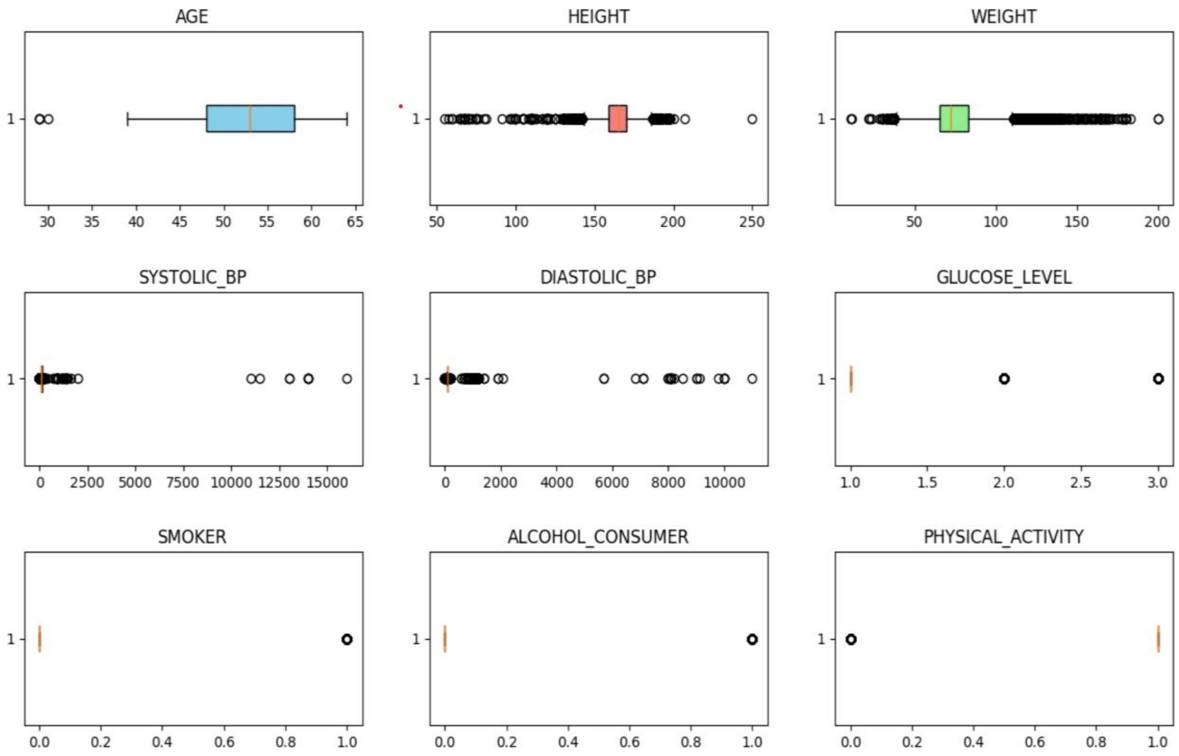
An outlier is a data point outside the accepted range of values. Outliers can pose a significant threat to model performance, particularly in algorithms, such as linear regression and k-nearest neighbors. Majority of models are based on Euclidean distances, and outliers can mislead the process, skewing the regression line or decision boundary and hence provide inaccurate predictions.

The range of accepted values is determined using the formulas:

Lower Quartile = Q1 - 1.5 * IQR

Upper Quartile = Q3 + 1.5 * IQR

Figure 2. *Detecting and Removing Outliers*



2.3. Handling Outliers

The variables GLUCOSE_LEVEL, SMOKER, ALCOHOL_CONSUMER and PHYSICAL_ACTIVITY had no outliers.

Age: Only the data for ages between 40 and 65 is considered.

Height: Only the data with a height between 4.5ft to 6.5ft is considered.

Weight: Only the data with a weight between 40 kg and 180 kg is considered.

Systolic_BP: Only the values within the range 90 and 170 are considered.

Diastolic_BP: Only the values within the range 65 and 105 are considered.

Once all the outliers are dropped, the dataset now has 63828 records with 12 columns.

3. Exploratory Data Analysis

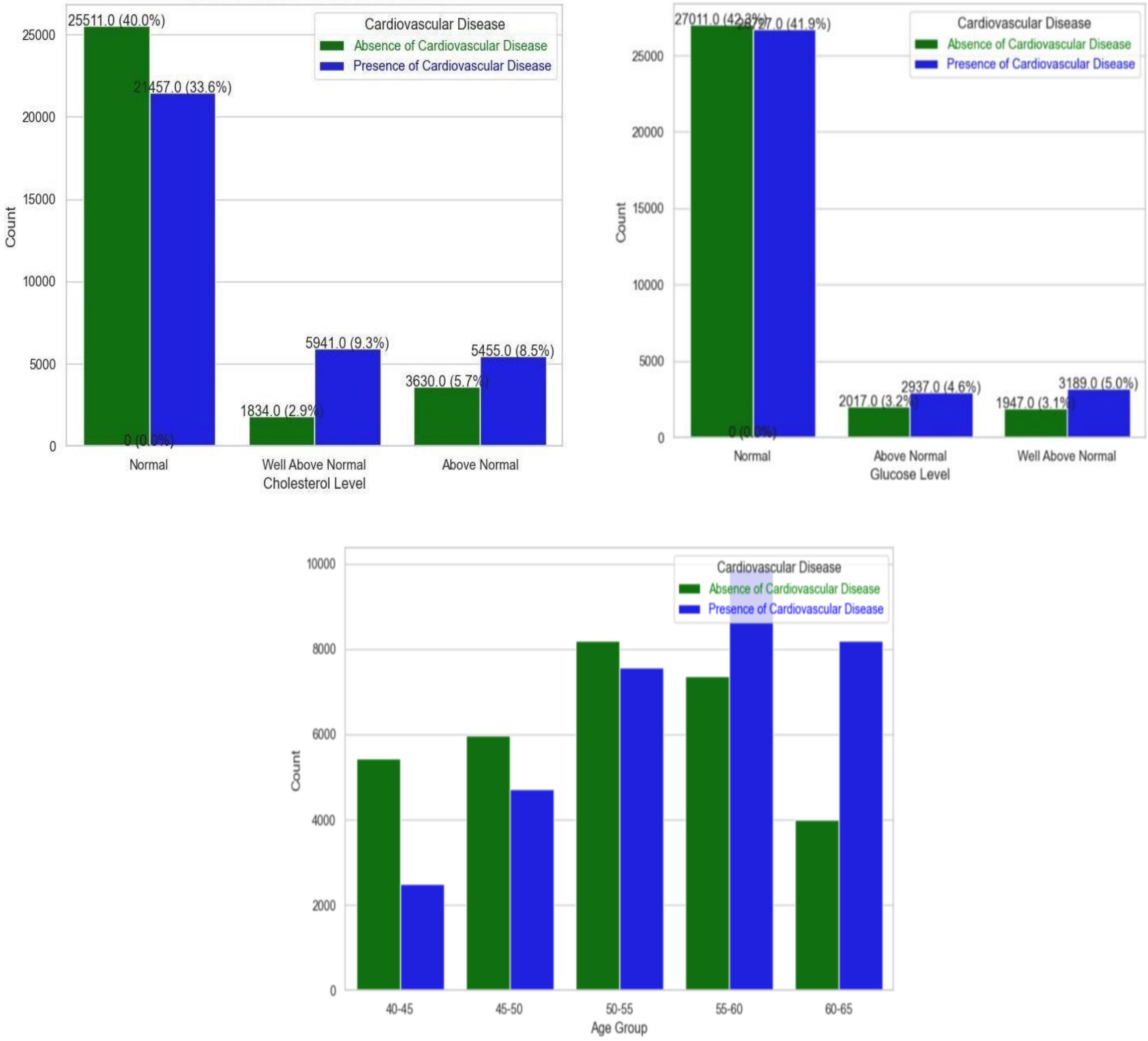
The data implies that the number of records with a risk of cardiovascular disease is high. Thorough data exploration is performed to draw more insights and patterns from this data.

Age: Patients in the age groups of 55 to 60 and 60 to 65 face a higher risk of cardiovascular diseases.

Glucose Level: Patients with normal glucose levels had less probability of facing cardiovascular disease.

Cholesterol Level: Patients with normal cholesterol levels had less probability of facing the disease.

Figure 3. *Exploratory Data Analysis*



4. Feature Selection

Feature selection is performed on a dataset to reduce the number of input variables by ranking the features that have the most influence on the target variable. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model (Brownlee, 2020).

As a part of this process, three following techniques were used:

a. Correlation Matrix

b. Decision Tree

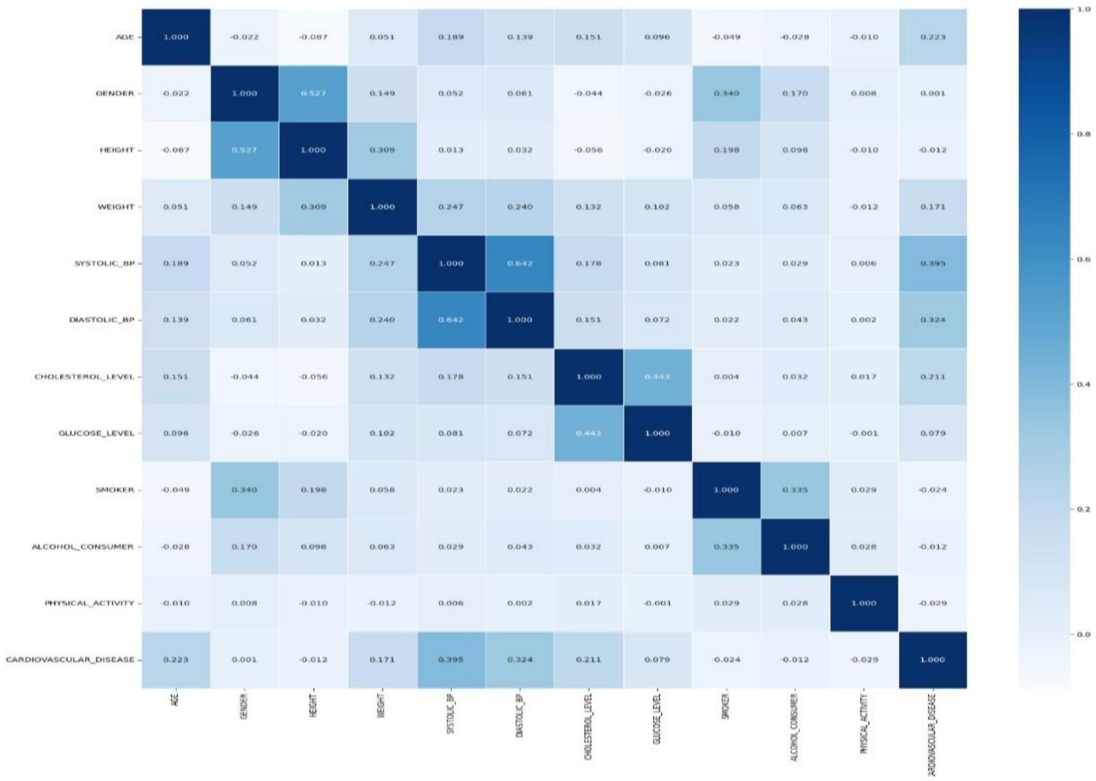
c. Chi-Square Statistic

4.1. Correlation Matrix

The correlation matrix is a square and symmetrical matrix with dimensions (K × K), where the entry at position (i, j) represents the correlation between columns i and j of the variable X (Pérez et al., 2009). This statistical technique is used to evaluate the relationship between variables in a data set. The matrix is a table where each cell has a correlation coefficient, with 1 denoting a positive relationship with the target variable, 0 denoting a neutral relationship, and -1 a negative relationship between the variables. Values closer to 1 imply strong relationship and values closer to 0 imply weak relationships.

Upon utilizing this method, it can be inferred that Cholesterol level, blood pressure (both systolic and diastolic), Age, Weight and Glucose Level have a strong relationship with cardiovascular diseases.

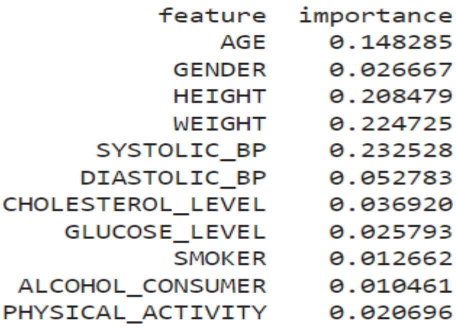
Figure 4. *Correlation Matrix*



4.2. Decision Tree

A Tree-based model is used for feature importance by determining how much each feature contributes to reducing the uncertainty in the target variable. The features Systolic_BP, Weight followed by Height appear to have the strongest influence on the target variable.

Figure 5. *Feature Importance – Decision Tree Model*



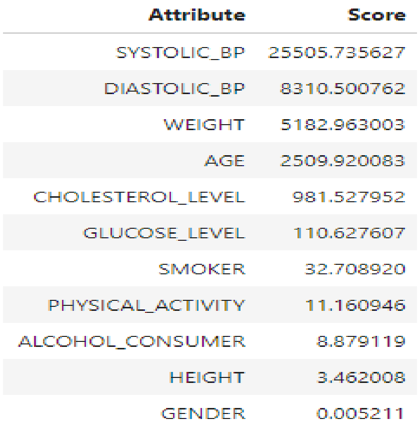
```
           feature  importance
               AGE    0.148285
            GENDER    0.026667
            HEIGHT    0.208479
            WEIGHT    0.224725
       SYSTOLIC_BP    0.232528
      DIASTOLIC_BP    0.052783
 CHOLESTEROL_LEVEL    0.036920
     GLUCOSE_LEVEL    0.025793
            SMOKER    0.012662
  ALCOHOL_CONSUMER    0.010461
 PHYSICAL_ACTIVITY    0.020696
```

4.3. Chi – Square Statistic

The chi-square test is a statistical method employed to assess disparities among categorical variables within a randomly selected sample, aiming to evaluate the adequacy of the fit (Hayes et al., 2023). Chi-square is used in feature selection to assess the independence between each feature and the target variable. By calculating the chi-square statistic for each feature, it measures the significance of the association. Features with higher chi-square values are likely to have a stronger influence on the target variable.

It can be inferred that Systolic_BP has the highest influence, followed by Diastolic_BP, Weight and then Age, while Height and Gender seem to have the least influence on the target variable.

Figure 6. *Feature Selection – Chi – Square Statistic*



Attribute	Score
SYSTOLIC_BP	25505.735627
DIASTOLIC_BP	8310.500762
WEIGHT	5182.963003
AGE	2509.920083
CHOLESTEROL_LEVEL	981.527952
GLUCOSE_LEVEL	110.627607
SMOKER	32.708920
PHYSICAL_ACTIVITY	11.160946
ALCOHOL_CONSUMER	8.879119
HEIGHT	3.462008
GENDER	0.005211

Based on the results from Feature Selection, the variables Gender, Height, and Alcohol Consumer dropped as they had the least impact on the target variable.

4.4. Data Understanding

Skewness of the dataset: Skewness is a measure of the symmetry of a distribution. The variables Age and Physical activity are negatively skewed, while all other variables are positively skewed. A negative skew implies that there are fewer extremely low values in the variable.

Kurtosis of the dataset: Kurtosis measures how the data is distributed. When the variables are plotted, it can be observed that they either have Platykurtic or Leptokurtic distribution. Thus, the data must be normalized.

Standard Scaler method is used to normalize the data. It is a method used in data preprocessing to scale numerical features to a standard range. It standardizes a feature by subtracting the mean and then scaling to unit variance. This changes the mean of the data to 0 with a standard deviation of 1. It is done to ensure that all the features are on a similar scale and no feature can create a bias in the machine-learning model.

5. Method Application

Upon completion of the preprocessing steps and understanding the data, the project proceeds to training the dataset on machine learning models suitable for binary classification prediction.

The use of Confusion Matrix in each of the models brings out the performance or the accuracy of the algorithm on the dataset and the Receiver Operating Characteristic (ROC) curve shows the performance of a classification model at all classification thresholds. This curve plots two parameters as True Positive Rate and False Positive Rate. The more the curve is away from the random classifier, the better the data is. Different models were applied, and the inference of the results is discussed here in detail.

5.1. Logistic Regression

Logistic regression estimates the probability of an event occurring, based on a given dataset of independent variables (IBM, n.d.). Since the outcome is a probability, its value lies between 0 and 1. Logistic regression is a supervised machine learning algorithm that makes use of logistic functions to predict the probability of a binary outcome. It is plotted between the target variable and the eight input variables. The evaluation metrics and Confusion Matrix are as below, followed by the ROC curve:

Figure 7. *Confusion Matrix – Logistic Regression*

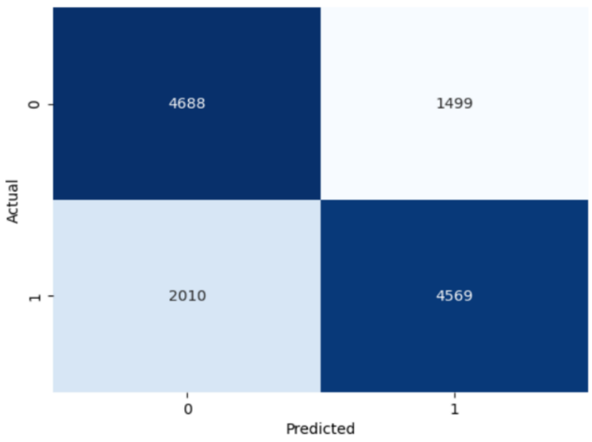
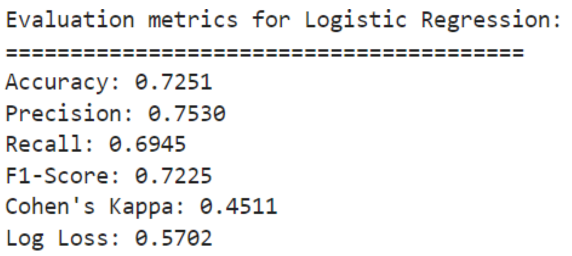
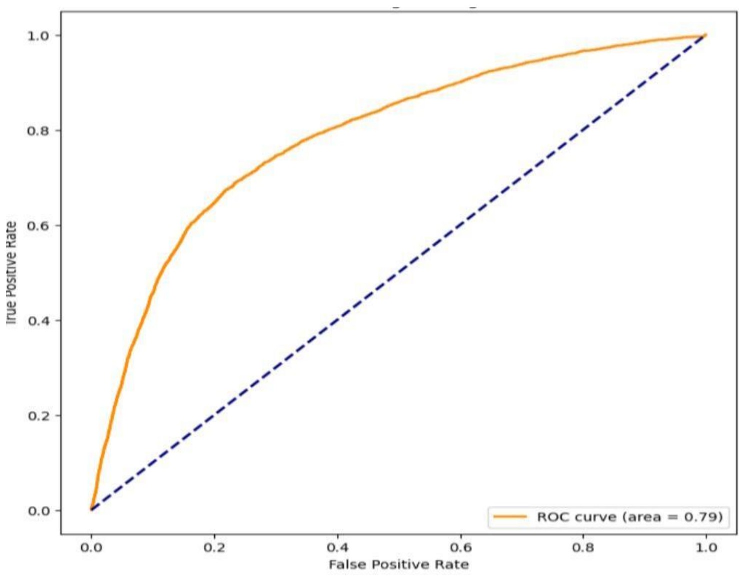


Figure 8. *Evaluation Matrix for Logistic Regression*



```
Evaluation metrics for Logistic Regression:
========================================
Accuracy: 0.7251
Precision: 0.7530
Recall: 0.6945
F1-Score: 0.7225
Cohen's Kappa: 0.4511
Log Loss: 0.5702
```

Figure 9. *ROC Curve – Logistic Regression*



The Logistic Regression model could predict the outcome with 72% accuracy and with a precision of 75%. The ROC curve has an Area Under Curve of 0.79, suggesting that the model is moderately good at predicting correctly, but there is more room for improvement.

5.2. Decision Tree Classifier

A decision tree represents a non-parametric supervised learning technique employed for classification and regression purposes. It features a hierarchical tree structure encompassing a root node, branches, internal nodes, and leaf nodes (IBM, n.d.). It has a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes.

Figure 10. *Decision Tree Diagram with all features*

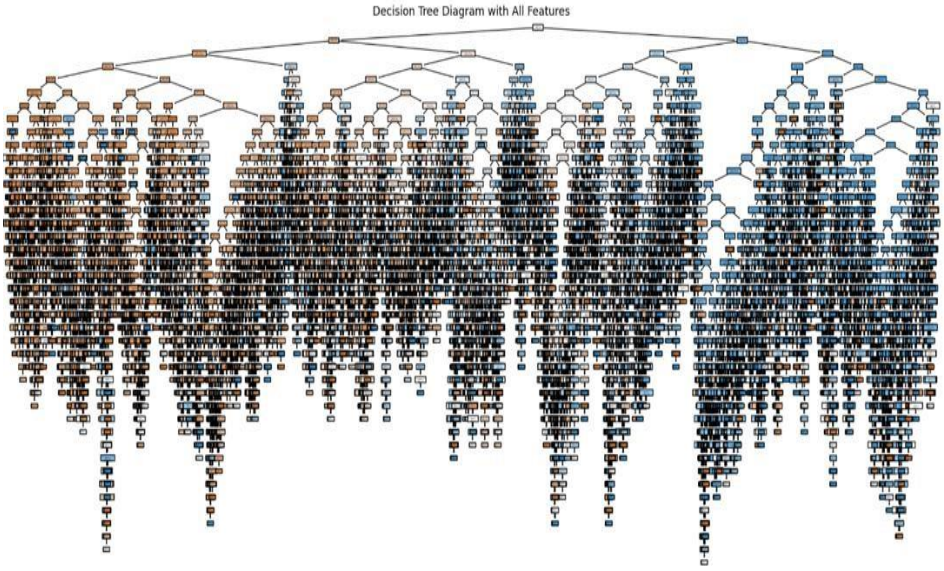


Figure 11. *Confusion Matrix for Decision Tree Classifier*

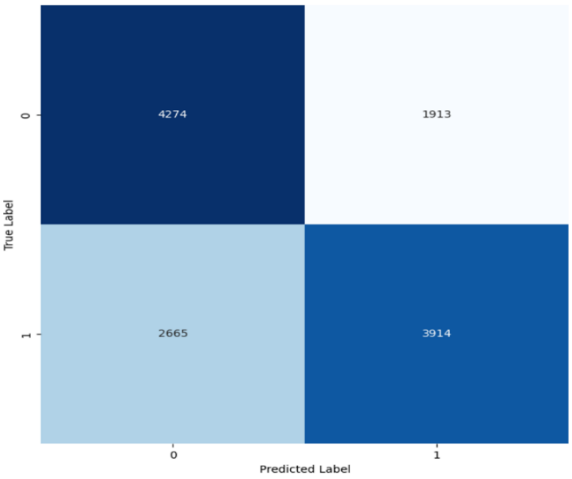
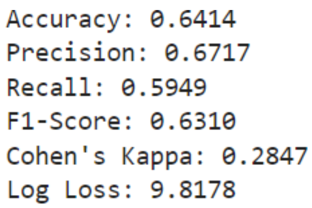
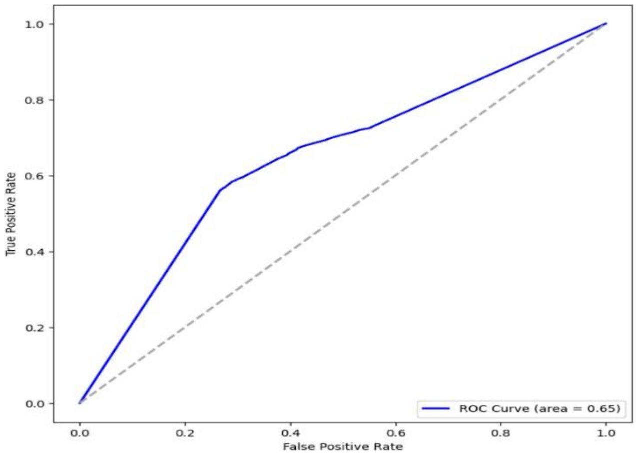


Figure 12. *Evaluation Matrix for Decision Tree Classifier*



```
Accuracy: 0.6414
Precision: 0.6717
Recall: 0.5949
F1-Score: 0.6310
Cohen's Kappa: 0.2847
Log Loss: 9.8178
```

Figure 13. *ROC Curve – Decision Tree Classifier*



From this, it can be inferred that the model has an accuracy of 64% and precision of 67%. The ROC AUC is 0.65 and the ROC curve is flatter than the Logistic Regression model. When compared to the Logistic Regression model, the Decision Tree classifier demonstrates slightly lower performance.

5.3. Support Vector Machine (SVM) Model

A support vector machine (SVM) uses supervised learning models to solve complex classification, regression, and outlier detection problems by performing optimal data transformations that determine boundaries between data points based on predefined classes, labels, or outputs. The evaluation metrics and Confusion Matrix are as below, followed by the ROC curve:

Figure 14. *Confusion Matrix – SVM*

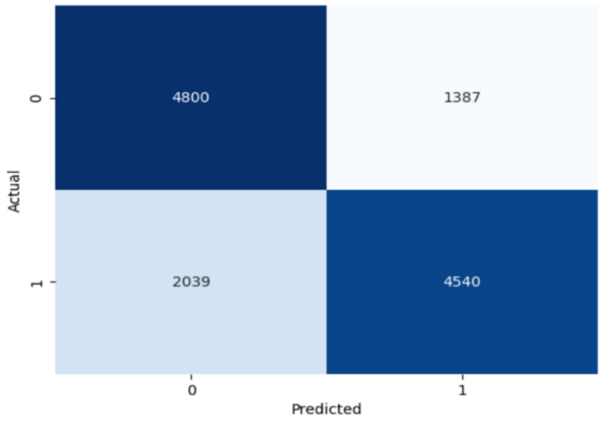
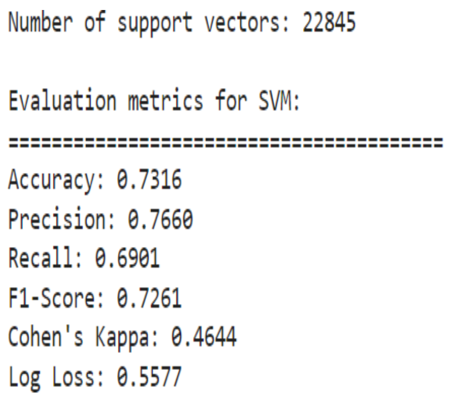


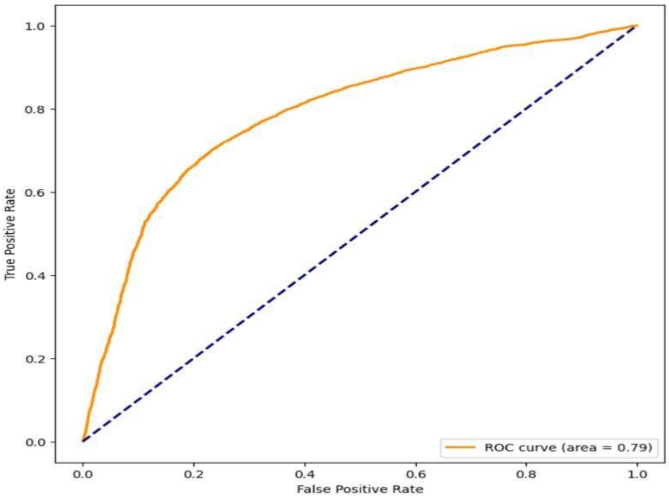
Figure 15. *Evaluation Matrix for SVM*



```
Number of support vectors: 22845

Evaluation metrics for SVM:
========================================
Accuracy: 0.7316
Precision: 0.7660
Recall: 0.6901
F1-Score: 0.7261
Cohen's Kappa: 0.4644
Log Loss: 0.5577
```

Figure 16. *ROC Curve – SVM*

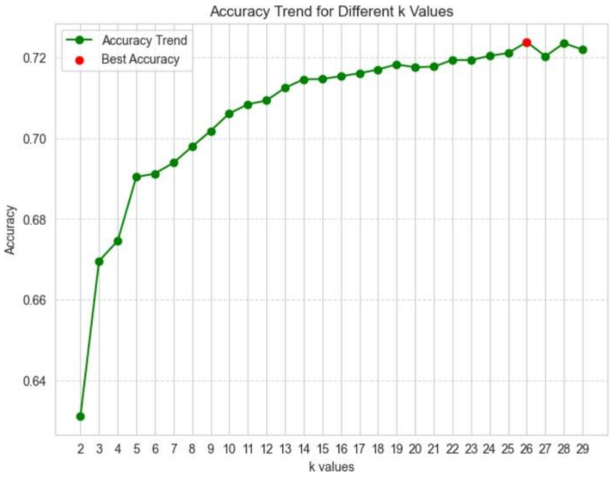


The model has an accuracy of 73% and precision of 76%. The ROC AUC is 0.79 which indicates that the model is relatively good at predicting the target variable. The SVM model showcases better performance when compared to the 2 previous models.

5.4. K Neighbors Classifier

KNN or K Nearest Neighbor is a Machine Learning algorithm that uses the similarity between data to make classifications (supervised machine learning) or clustering (unsupervised machine learning). This model tests different k values, and the best accuracy is observed at one point.

Figure 17. *Accuracy Trend for Different k Values*



The value of k used in the algorithm is obtained by iteratively checking which value of k gives the highest accuracy, between the ranges of 2 and 30. The above scatterplot visualizes all the values obtained and maximum accuracy was obtained at k =26.

Figure 18. *Confusion Matrix - KNN*

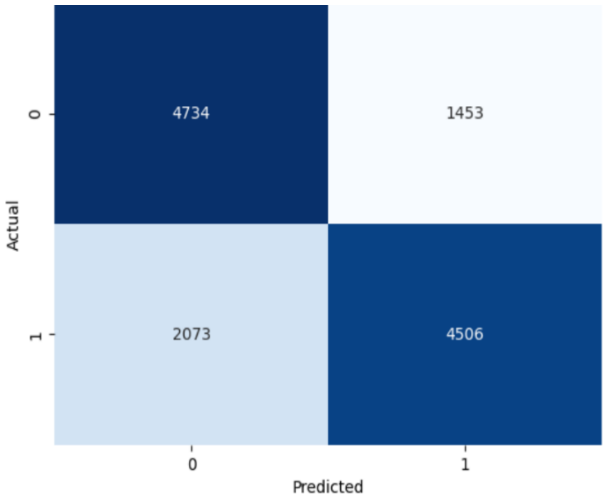
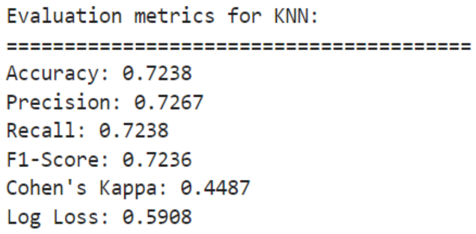
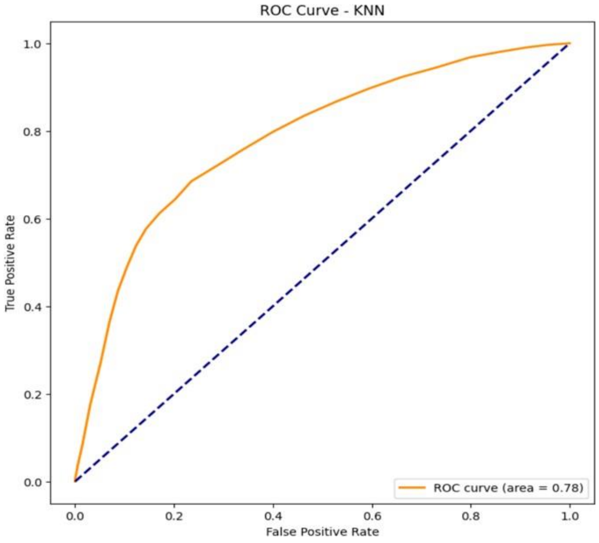


Figure 19. *Evaluation Matrix for KNN*



```
Evaluation metrics for KNN:
========================================
Accuracy: 0.7238
Precision: 0.7267
Recall: 0.7238
F1-Score: 0.7236
Cohen's Kappa: 0.4487
Log Loss: 0.5908
```

Figure 20. *ROC Curve - KNN*

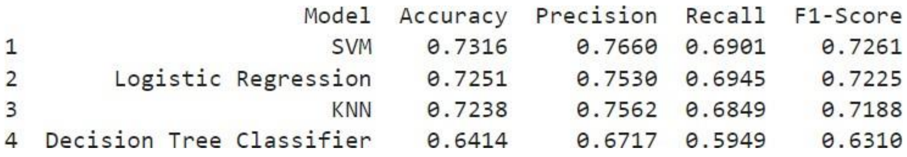


The model has an accuracy and precision of 72%. The ROC AUC is 0.78 which indicates that the model can predict the outcome with relative accuracy.

6. Model Evaluation

The results obtained from training the 4 models on the dataset are compared and the following insights were obtained.

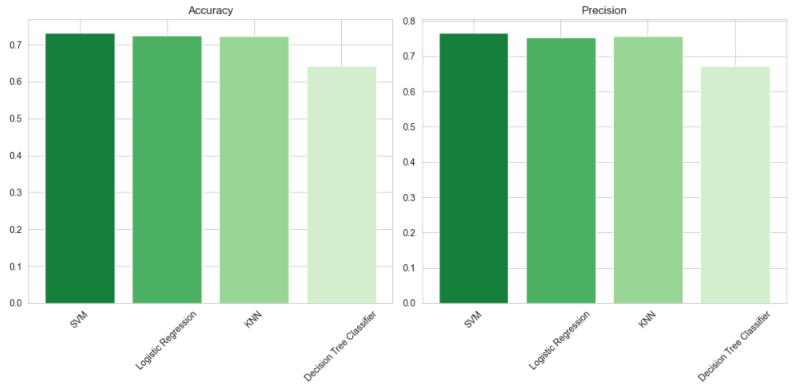
Figure 21. *Comparison of four Models*



```
                         Model  Accuracy  Precision  Recall  F1-Score
1                          SVM    0.7316     0.7660  0.6901    0.7261
2          Logistic Regression    0.7251     0.7530  0.6945    0.7225
3                          KNN    0.7238     0.7562  0.6849    0.7188
4     Decision Tree Classifier    0.6414     0.6717  0.5949    0.6310
```

The Support Vector Machine data mining model demonstrated the best performance in terms of all the metrics that were considered. The Logistic Regression model and KNN models also exhibited results nearly as precise as the SVM model.

Figure 22. *Accuracy and Precision of four Models*

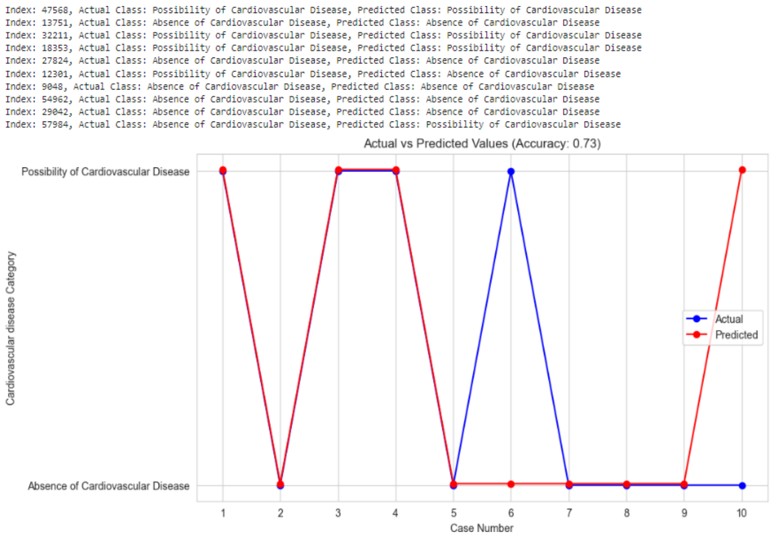


6.1. Prediction for Unseen Data

As can be observed, the Support Vector Machine model performed better than other models. Thus, the SVM model was used to predict whether a person would ever experience a cardiovascular disease in their life. The accuracy of the model is tested by giving it data it had never seen before.

As input, 10 cases were given to the model, out of which the model correctly predicted the outcomes for 8 cases, with an accuracy of 73%.

Figure 23. *Actual Vs Predicted Values*



```
Index: 47568, Actual Class: Possibility of Cardiovascular Disease, Predicted Class: Possibility of Cardiovascular Disease
Index: 13751, Actual Class: Absence of Cardiovascular Disease, Predicted Class: Absence of Cardiovascular Disease
Index: 32211, Actual Class: Possibility of Cardiovascular Disease, Predicted Class: Possibility of Cardiovascular Disease
Index: 18353, Actual Class: Possibility of Cardiovascular Disease, Predicted Class: Possibility of Cardiovascular Disease
Index: 27824, Actual Class: Absence of Cardiovascular Disease, Predicted Class: Absence of Cardiovascular Disease
Index: 12301, Actual Class: Possibility of Cardiovascular Disease, Predicted Class: Absence of Cardiovascular Disease
Index: 9048, Actual Class: Absence of Cardiovascular Disease, Predicted Class: Absence of Cardiovascular Disease
Index: 54962, Actual Class: Absence of Cardiovascular Disease, Predicted Class: Absence of Cardiovascular Disease
Index: 29042, Actual Class: Absence of Cardiovascular Disease, Predicted Class: Absence of Cardiovascular Disease
Index: 57984, Actual Class: Absence of Cardiovascular Disease, Predicted Class: Possibility of Cardiovascular Disease
```

Actual vs Predicted Values (Accuracy: 0.73)
Possibility of Cardiovascular Disease
Absence of Cardiovascular Disease
Cardiovascular disease Category
Actual
Predicted
1
2
3
4
5
6
7
8
9
10
Case Number

7. Conclusion

The Cardiovascular Disease Prediction project was carried out by using various machine learning models on data collected from patients during examination. The data preparation phase involved meticulous cleaning, feature engineering, and addressing outliers to enhance the robustness of the models.

After thoroughly exploring the data using the Logistic Regression, Decision Tree Classifier, Support Vector Machine Model and K Neighbours Classifier model, it was observed that the SVM model predicted the outcomes with the highest accuracy, closely followed by the Logistic Regression and K-Neighbors Classifier models.

The SVM model was selected to predict the outcome on unseen data and the model correctly predicted the outcome 8 out of 10 times with 73% accuracy.

While the results achieved were promising, the model metrics could be further enhanced, by improving the dataset to include a minimum of 1 million rows of trustable unbiased sources of patient data from different geographical locations, considering the impact on the health of every individual. By exploring additional relevant features such as genetic markers, lifestyle factors, or medical history the accuracy of the models could be further improved. This would enable the models to be more precise and potentially save lives.

References

Brownlee, J. (2020, August 20). *How to choose a feature selection method for machine learning*. MachineLearningMastery.com. https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Hayes, A., James, M. & Velasquez, V. (2023, May 22). Chi-Square (X2) statistic: *What it is, examples, how and when to use the test*. Investopedia. https://www.investopedia.com/terms/c/chi-square-statistic.asp

IBM. (n.d.). Decision Trees. IBM. https://www.ibm.com/topics/decision-trees

IBM. (n.d.). Logistic Regression. IBM. https://www.ibm.com/topics/logistic-regression

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International journal of computer science*, *1*(2), 111-117.

Lane, P. C., Clarke, D., & Hender, P. (2012). On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decision Support Systems*, *53*(4), 712-718.

Pérez, N. F., Ferré, J., & Boqué, R. (2009). Calculation of the reliability of classification in discriminant partial least-squares binary classification. *Chemometrics and Intelligent Laboratory Systems*, *9*5(2), 122-128.

World Health Organization. (2021, June 11). Cardiovascular diseases (CVDs). World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)