TEESSIDE
UNIVERSITY

COMPARATIVE EVALUATION OF MACHINE LEARNING ALGORITHM ON CORONARY HEART DISEASE PREDICTION

Part 2 of ICA Project Report

TSEGHA MADONNA (Student)
B1215326

Abstract

This study aimed to study and predict the risk of coronary heart disease through the application of machine learning algorithms. Two primary algorithms, namely Support Vector Machine (SVM) and K-Nearest Neighbours (K-NN), were employed for the building of a predictive model. Additionally, results obtained from the Random Forest classifier were compared with those from the SVM and KNN algorithms to provide a comprehensive analysis and compare their accuracy levels, to select the best-performed model.

The dataset used for this study was sourced from Kaggle, a recognized platform for hosting datasets relevant to machine learning and general data science projects. Leveraging the R programming language, the study implemented machine learning prediction models on the dataset.

The highlight of the project focused on the performance comparison of the K-NN model with that of the support vector machine. Notably, the KNN had the highest accuracy, achieving a 74.16% accuracy over SVM which gave a 69.33%. This suggests a high level of precision in predicting the risk of coronary heart disease using the KNN algorithm.

The study further investigated the performance on Random Forest model on the dataset, when compared to the KNN, it performed better by 8%. The accuracy for the Random Forest model was 78.39%.

Furthermore, the temporal dimension of the study was emphasized by utilizing the dataset to predict patients' risk of coronary heart disease over ten years. This extended timeframe allowed for a robust evaluation of the algorithms' effectiveness in forecasting the development of coronary heart disease over a substantial duration.

In conclusion, the study underscores the efficacy of machine learning algorithms, particularly the K-Nearest Neighbours, in predicting a patient's risk of developing coronary heart disease.

The findings contribute valuable insights into the potential application of these algorithms for long-term risk assessment in cardiovascular health, with implications for clinical and preventive healthcare practices.

Problem Specification and Business Questions

Problem Specification

Cardiovascular diseases (CVDs) as defined by NHS is a general term for conditions that affect the heart or blood vessels. It is majorly influenced by the build-up of fatty deposits in the arteries, known as atherosclerosis, and an increased risk of blood clots (NHS, 2022).

Taking insights from a publication by (WHO), CVDs are the most prevalent cause of death globally. The estimated number of lives taken each year by the disease is said to be around 17.9 million lives. CVDs result from a disorder in the heart and blood vessels. Also, by the NHS, CVDs are one of the main causes of death and disability in the UK, however, they can be prevented by living a healthy lifestyle (WHO,2023), (NHS, 2022).

Although the exact cause of CVDs has not been defined, there are lots of factors that can increase its risk. These are considered as “Risk Factors“. The data set chosen for this project highlights these risk factors and predicts the risk of getting the disease due to the dependent variables. These are not limited to: High Blood Pressure, Smoking, Diabetes, Body Mass Index (BMI), etc.

However, early detection and treatment of patients who have a high risk of developing cardiovascular disease can help them maintain a healthier lifestyle and prevent the difficulties associated with it. (Zriqat, Altamimi, and Azzeh 2017).

Business Questions

Some of the business questions asked in the study were not limited to:

What is the correlation between lifestyle choices i.e the dependent variables and the risk of cardiovascular disease over ten years?

What model can best predict the outcome of having Coronary Heart Disease after ten years given the risk factors associated to the disease?

How can the model be improved to better perform on a similar dataset?

Contents

Abstract.....

Problem Specification and Business Questions.....

Problem Specification .....

Business Questions.....

INTRODUCTION AND BACKGROUND 1

Aim.....1

The goal of the study using the chosen dataset is to estimate a patient's 10-year risk of coronary heart disease (CHD) using machine learning. The study will create a model that can forecast a patient's risk based on a number of variables. 1

Objectives.....1

RELATED STUDIES.....1

Two classification algorithms (KNN and SVM) will be used in this study to determine a patient's 10-year risk of coronary heart disease. 2

METHODOLOGY 2

Dataset Description 2

Data Pre-processing.....3

Handling missing values 3

Exploratory Data Analysis 4

Distribution Varibales in the dataset 5

Corellation and Data filtering 5

Data Normalization.....6

Balancing the Dataset.....6

Training and Testing 6

Partitioning the data frame.....6

Machine Learning Models and Algorithms 6

Model Application 7

Confusion matrix 7

Comparing the Models.....8

Testing Random Forest 8

RESULT AND CONCLUSION.....9

RECOMMENDATIONS AND FURTHER STUDIES.....9

REFERENCE 9
APPENDIXES 10

INTRODUCTION AND BACKGROUND

Coronary heart disease (CHD) is a lethal condition, particularly in the elderly and middle-aged population. In 2013, 17 million people died due to CHD, accounting for half of all mortality in the United States and the most prevalent cause of death globally. 3 million fatalities from CHD occur before the age of sixty, according to the WHO.

However, If healthy eating habits and active living had been promoted earlier, 90% of these deaths may have been prevented.

In traditional healthcare, doctors identify ailments based on symptoms, leading to unintentional mistakes and higher healthcare costs. Machine learning can simulate human decision-making abilities to answer queries like heart disease age, gender, treatment history, and heart attack risk estimation. This technology can help improve healthcare quality and reduce errors, ultimately improving patient care and overall patient outcomes. (Daniel J., 2015).

Data mining is an interdisciplinary field that integrates machine learning, neural networks, database technology, information science, and statistical techniques. However, it is not commonly utilised on medical data, but clear presentation can increase its utility.

Machine learning (ML) is a computational method used in AI systems to gather, convert, and update data. It helps people learn by providing data or illustrations. When computational solutions are not available, structural patterns are weak, or when the topic is unknown, machine learning can be useful. The purpose of this study is to identify pertinent heart disease risk variables over a ten-year period and use the ML risk base model to predict overall risk.

Aim

The goal of the study using the chosen dataset is to estimate a patient's 10-year risk of coronary heart disease (CHD) using machine learning. The study will create a model that can forecast a patient's risk based on several variables.

Objectives

a. Explore data processing methods in coronary heart disease prognosis.
b. Build machine learning models (K-Nearest Neighbours and Support Vector Machine Algorithms) in predicting the probability of developing CHD.
c. Compare the accuracy and sensitivity of the models.

RELATED STUDIES

Numerous research have employed data mining techniques to examine Heart disease data and its survival rate. For instance, numerous trials have been conducted on the diagnosis of coronary diseases, using approaches like clustering, association rules methodology, and classification algorithms. These studies have shown the usefulness of these techniques in predicting and understanding various diseases. (Jyoti Sony et al 2011).

By classifying patients into cohesive groups according to general distribution patterns and connections between data attributes, clustering techniques help data scientists comprehend naturally occurring data groupings and spot cardiac diseases. [Karegowda et al 2012].

Although clustering is a useful tool for other classification algorithms due to its affordability, it is also frequently employed as a pre-processing step in large data sets to identify hidden patterns in cardiac patients and clarify data distribution. (2009) Numerous studies have utilized

association rules mining to identify patterns in large patient data sets for predicting heart attacks, thereby enhancing the effectiveness of heart attack prediction and detecting heart disease.

[Jyoti Sony et al. 2011] and [P. Chandra et al. 2012] In order to find links between patient features, the study created a model, which found interesting factors in the dataset. It also established an integrated technique of association and classification to identify rules in the database and build an effective classifier. [J. Sony 2011]. Association algorithms are effective in health prediction, but they have been found to have poor efficiency due to a significant number of criteria being uninteresting. [Moreno 2005].

Classification is a data mining technique used to construct models for identifying heart disease. It involves assigning data to specific groups or categories, aiming to accurately identify the intended class for each occurrence. Classification starts with a predetermined assignment in which projected test results are compared to known target values. To create and testing models, historical data is frequently divided into two groups, with probability and class tasks being used to identify each model.

Nadiah A. Baghdadi et al. 2023 illustrated results on a similar dataset on cardiovascular diseases using various classifiers such as the SVM (Support Vector Machine) to measure accuracy and precision. Other top-performing classifiers included; RandomForest, LogisticRegression, and KNN which gave similar accuracy and precision scores.

The authors [Chen, J., and Greiner, 1999] Chen and Greiner (1999) applied classification algorithms to heart disease datasets, observing various findings on examples using SVM, neural artificial network, and decision tree.

For the purpose of accurately predicting coronary heart disease, [Z. Khan, D. K. Mishra, V. Sharma, and A. Sharma 2020] studied machine learning techniques such as random forest support vector machine, Gauss-naive bayesian, Decision Trees, K-nearest neighbour, and logistic regression classifiers.

Finally, Xin Qian et al. 2022 used SVM in the prediction of cardiovascular disease since SVM is one of the most commonly used ML algorithms that can effectively classify small samples and nonlinear data. However, the prediction outputted by the SVM model was by default, hence the probability of CVD was not directly predicted. The Platt scaling method was then used to predict the probabilities output using four other models for accurate prediction of CVD risk and the identification of high-risk groups.

Two classification algorithms (KNN and SVM) will be used in this study to determine a patient's 10-year risk of coronary heart disease.

METHODOLOGY

Dataset Description

The dataset, obtained from Kaggle contained 3,390 rows and 17 columns, each representing a potential future risk around the patients'; demographic, behavioral, and medical histories. The target column, containing the outcome of Coronary Heart Disease (CHD) over ten years, will be predicted using a Support Vector Machine model (SVM) and K-Nearest Neighbours (KNN).

FEATURES	DATA TYPE	DESCRIPTION
ID	Numeric	Data Serial Number
Sex	Categorical variable (binary)	1 (Male) or 0 (Female)
Education	Categorical variable (ordinal)	1 = Junior School, 2 = High School, 3 = College, 4 = College
Smoking	Categorical variable (binary)	Yes (1) or No (0) to smoking
Age	Integer	The age range in the dataset was between 32 - 70
BPMeds	Categorical variable (binary)	Yes (1) or No (0) to Blood Pressure Medication
prevalentStroke	Categorical variable (binary)	Previous history of Stroke
prevalentHyp	Categorical variable (binary)	Previous history of hypertension
Diabetes	Categorical variable (binary)	Yes (1) or No (0) to Diabetes
cisPerDay	Integer	Average number of Cigarettes Per Day
totCol	Numeric	Cholesterol Level
sysBP	Numeric	Systolic Blood Pressure
diaBP	Numeric	Diastolic Blood Pressure
BMI	Numeric	Body Mass Index
heartRate	Numeric	Patient's heart rate
Glucose	Numeric	Patient's glucose level
Target variable	Numeric	Risk of patient having Coronary Heart Disease (CHD) - (1 = Yes, 0 = No)

Table 1: Attributes of dataset.

Data Pre-processing

Handling missing values

Using the packages installed in R programming, the dataset was examined and visualised to determine the link between the variables. As shown in the diagrams, missing values, and "Null" rows were removed from the data. (Fig 1a) show that 510 rows (14%) of the sample data, were missing values. Of these missing values, 8.97% constitute of the projected glucose column. The median value was used to replace the missing values for the BMI and glucose columns. The remaining missing values were eliminated since they were less than 5%.

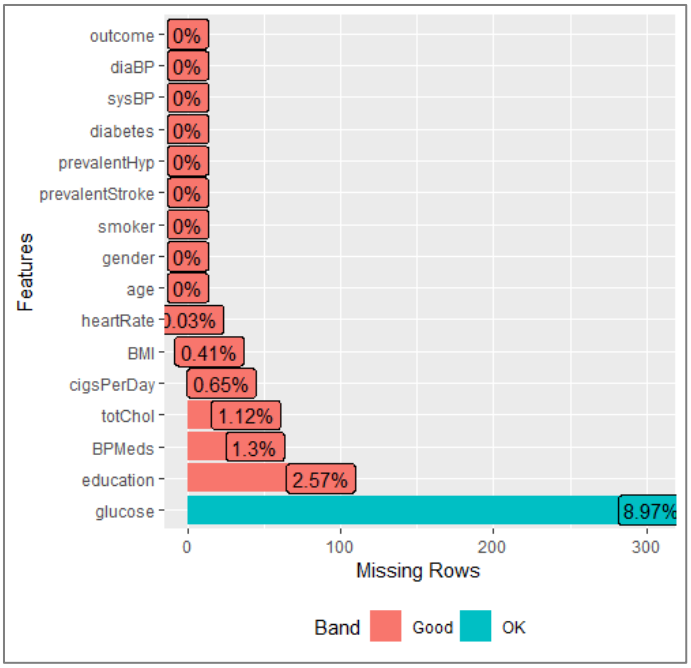


Fig 1a: Shows the variables with missing values with the glucose column having 8.97% missing values.

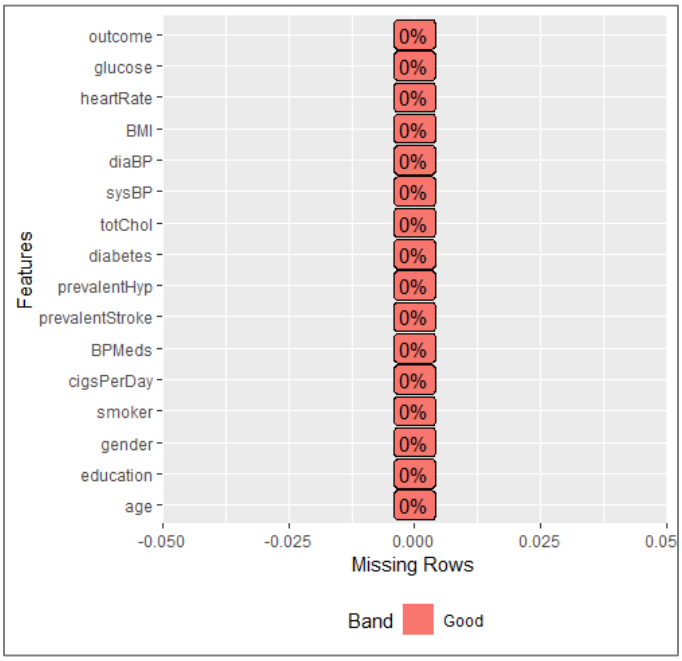
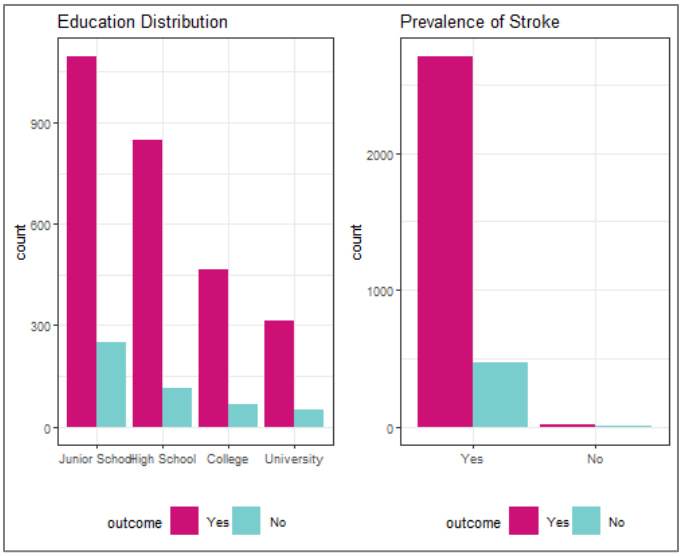
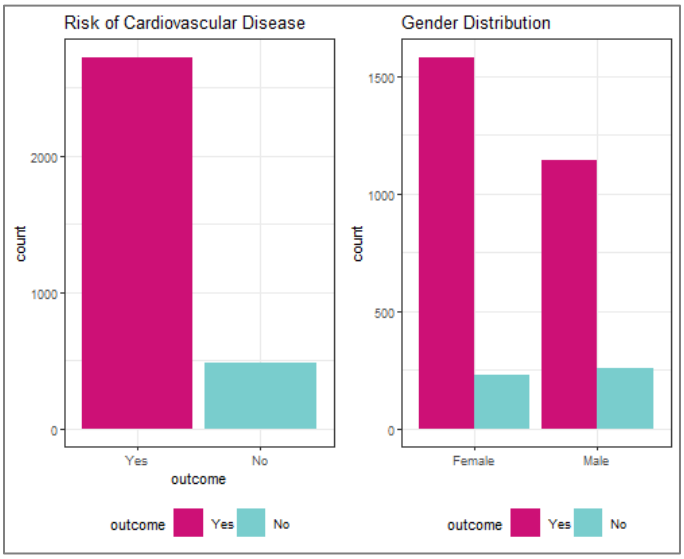


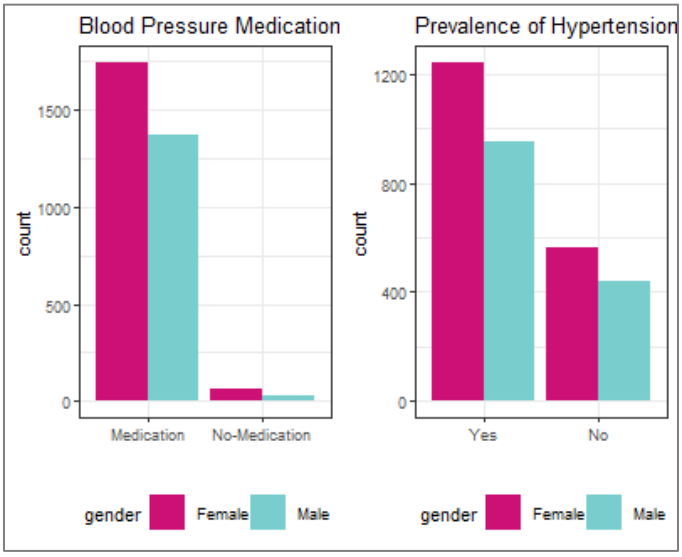
Fig 1b: Shows the variables after the glucose and the BMI variables were replaced with the median values and the rows with missing values removed.

Exploratory Data Analysis

To gain a deeper understanding of the dataset, some visualisations of the variables were created. Although those with lower levels of education appeared to be more susceptible to coronary heart disease, it would be advisable to investigate the potential causes in greater detail. It's critical to keep in mind that the dataset pertains to Americans, whose access to the healthcare system is restricted. Individuals with lower educational attainment are more likely to have restricted access to healthcare due to financial restrictions.







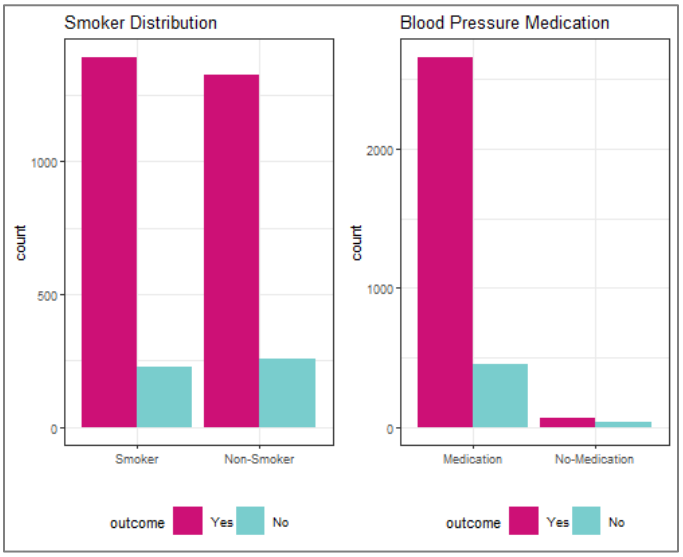


Fig 2: The above visuals show the distribution of some of the variables of the dataset.

50% (1,687 out of 3,390) of the patients were smokers. 54% (911) of the patient's that smoke are male while the women make up the remaining 46%(776).

The Patients were group into 5 categories by Age range, being that the lowest age group was 30 and the highest 70 years.

Patients between 40 – 49 (1,316) years of age were the most (39%) category in the dataset and were the second highest ranked based on Average cholesterol level.

The age range of 60-69 years had the highest average cholesterol level of 250. 47. These findings show that the elderly patients had higher cholesterol levels compared to other age groups.

Distribution of Varibales in the dataset

To density charts show the distribution of the various variables in the dataset.

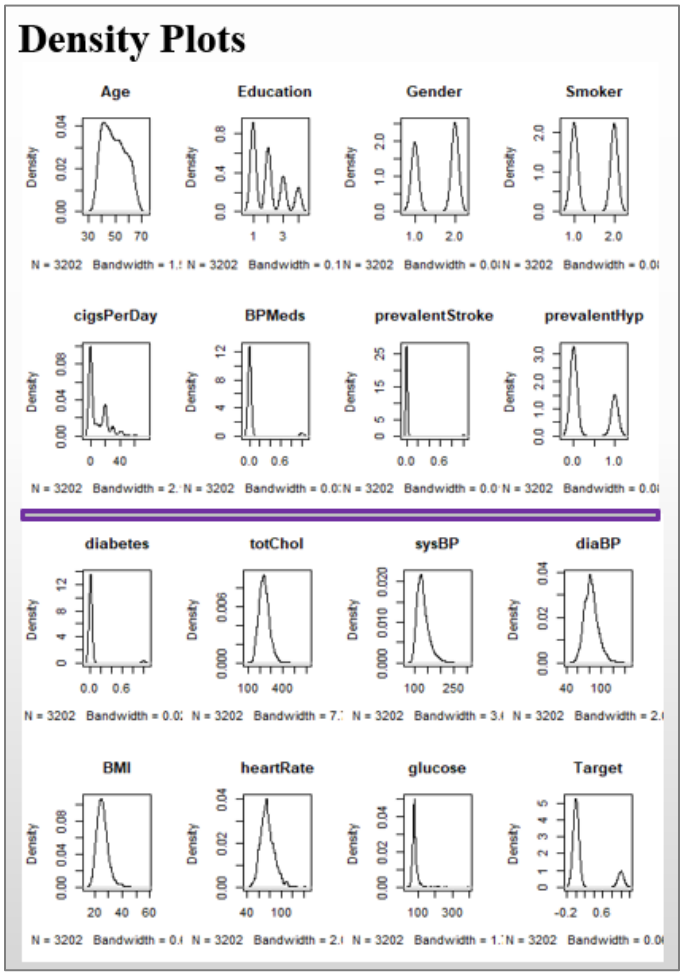


Fig 3: Density Plots showing the distribution of the dataset. The youngest age group in the dataset is 30 years, while the oldest age group is 70 years. The average age of patients in the research is 50 years.

Patients smoked an average of 9 cigerates per day. The highest number of cigarattes somked by a patient per day was 70. The males smoked an average of 14 cigarrettes a day, while Females smoked only 6 cigarrates on averager per day.

The average BMI of the 3,390 patients was observed to be 25.79 while the average glucose level was 82.

Females had the most average heart rate of 77.22, while the Males had an average of 74.35.

Corellation and Data filtering

Variables that have a strong association or corrrelation with the target class are filtered out using a correlation matrix (cm), which may cause bias in the model. The dark colour shades on the grid indicate a positive corelation, moving towards 1. Their colour intensity is reflected in the correlation coefficients. For instance, there is a high correlation between smoker and cigerattes per day, but not with gender, or symbolic blood pressure. Every variable correlated with the target class was examined. All of the characteristics will be kept and used to train the models because there is no strong correlation between them and the target class. Below, in Fig. 4, is the correlation matrix of the variables.

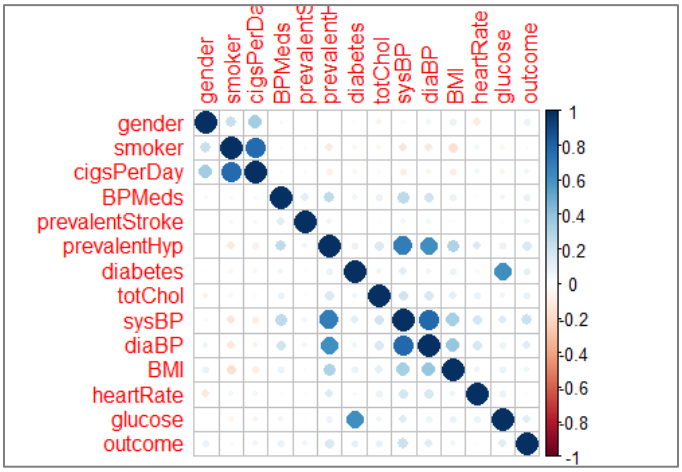


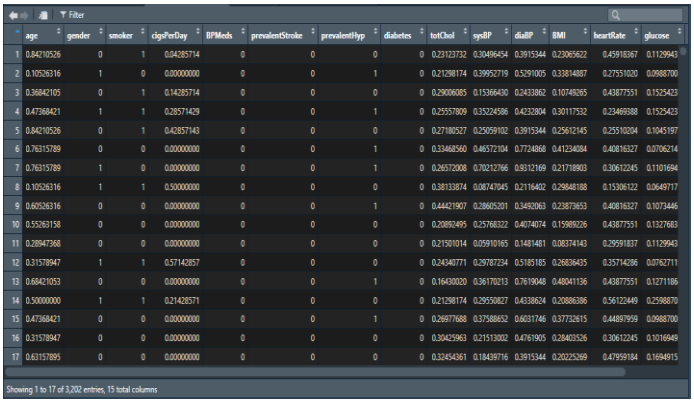
Fig 4: Plot of the correlation matrix that displays the color-based correlation intensity.

There isn't a single variable that strongly correlates with the outcome, or target variable. This suggests that when determining a patient's risk of developing coronary heart disease, all variables are important.

Data Normalization

This is a procedure employed during the data pre-processing phase, which gets the data ready for building machine learning using neural networks or models like Support Vector Machine, etc. (Muhammad Ali and Faraj, 2014).

To guarantee that the dataset fit into the models correctly, it was rescaled between 0 and 1.



	age	gender	smoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
1	0.84210526	0	1	0.04285714	0	0	0	0	0.23123732	0.30496454	0.3915344	0.23065622	0.45918367	0.1129943
2	0.10526316	1	0	0.00000000	0	0	1	0	0.21298174	0.39952719	0.5291005	0.33814887	0.27551020	0.0988700
3	0.36842105	0	1	0.14285714	0	0	0	0	0.29006085	0.15366430	0.2433862	0.10749265	0.43877551	0.1525423
4	0.47368421	1	1	0.28571429	0	0	1	0	0.25557809	0.35224586	0.4232804	0.30117532	0.23469388	0.1525423
5	0.84210526	0	1	0.42857143	0	0	0	0	0.27180527	0.25059102	0.3915344	0.25612145	0.25510204	0.1045197
6	0.76315789	0	0	0.00000000	0	0	1	0	0.33468560	0.46572104	0.7724868	0.41234084	0.40816327	0.0706214
7	0.76315789	1	0	0.00000000	0	0	1	0	0.26572008	0.70212766	0.9312169	0.21718903	0.30612245	0.1101694
8	0.10526316	1	1	0.50000000	0	0	0	0	0.38133874	0.08747045	0.2116402	0.29848188	0.15306122	0.0649717
9	0.60526316	0	0	0.00000000	0	0	1	0	0.44421907	0.28605201	0.3492063	0.23873653	0.40816327	0.1073446
10	0.55263158	0	0	0.00000000	0	0	0	0	0.20892495	0.25768322	0.4074074	0.15989226	0.43877551	0.1327683
11	0.28947368	0	0	0.00000000	0	0	0	0	0.21501014	0.05910165	0.1481481	0.08374143	0.29591837	0.1129943
12	0.31578947	1	1	0.57142857	0	0	0	0	0.24340771	0.29787234	0.5185185	0.26836435	0.35714286	0.0762711
13	0.68421053	0	0	0.00000000	0	0	1	0	0.16430020	0.36170213	0.7619048	0.48041136	0.43877551	0.1271186
14	0.50000000	1	1	0.21428571	0	0	0	0	0.21298174	0.29550827	0.4338624	0.20886286	0.56122449	0.2598870
15	0.47368421	0	0	0.00000000	0	0	1	0	0.26977688	0.37588652	0.6031746	0.37732615	0.44897959	0.0988700
16	0.31578947	0	0	0.00000000	0	0	0	0	0.30425963	0.21513002	0.4761905	0.28403526	0.30612245	0.1016949
17	0.63157895	0	0	0.00000000	0	0	0	0	0.32454361	0.18439716	0.3915344	0.20225269	0.47959184	0.1694915

Showing 1 to 17 of 3,202 entries, 15 total columns

Fig 5: The normalized data frame

Balancing the Dataset

Because the Target variable had a distribution of 85% (No) and 15% (Yes). The minority group had to be upsampled using the SMOTE method, which brought the observations down to 2,410 and the distribution were 60% (1,446 - No) and 40% (964 - Yes). This had to be done for better performance of the model as shown in Fig 6.

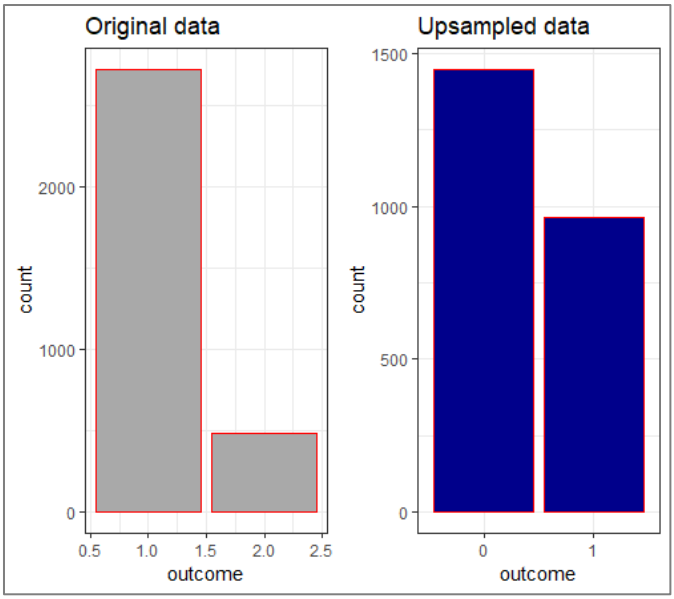


Fig 6: Plot showing the original and upsample dataset

Training and Testing

Partitioning the data frame

For data partitioning, the upssampled dataset was firstly shuffled and then split in half: 30% (722) was put aside for testing and 70% (1,688) for training the model. The data were also separated using the x and y axes, where the dependent variable (target class) was represented by the y axis and the remaining attributes as independent variables. The idea of the test set is to ensure that the model built is able to predict accurately the outcome given the dependent variables in the dataset.

Machine Learning Models and Algorithms

The risk of Coronary Heart Disease will be predicted compariengthe accuracy from two models; Support Vector Machine (SVM) and K-Nearest Neighbours (KNN). Random Forest machine learning was also tested on the dataset, to see how it will also perform with the dataset.

Support Vector Machine (SVM) are supervised machine learning systems that examine data for classification and regression.

KNN on the other hand is an instance-based, non-parametric machine learning algorithm. It works based on a data point's classification according to the feature space's k-nearest neighbors' majority class. The 'class' package in R offers a productive KNN implementation, making it simple to include this technique in projects involving data analysis and classification.

The two main advantages they offer over contemporary algorithms like as neural networks are as follows: When there are few samples (in the thousands), they work fast and effectively.

The dataset had 3,390 rows, which the SMOTE approach balanced to decrease to 2,960 observations. When there is a distinct margin of distinction between classes, the approach works well. (Bruno 2017).

Model Application

There are various approaches for establishing a machine learning model's architectural framework. The best model architecture for a given model isn't always known, so it's important to try out a lot of different options. Giving the software total command over every research and model architecture choice allows for this. Hyperparameters are parameters that have an impact on the design of a model.

On the other hand, cross-validation is a sampling technique that is employed to examine machine learning algorithms and forecast their potential performance on an alternative test dataset. A range of measures of our modeling method can be collected by repeatedly running it on different subsets of the data.

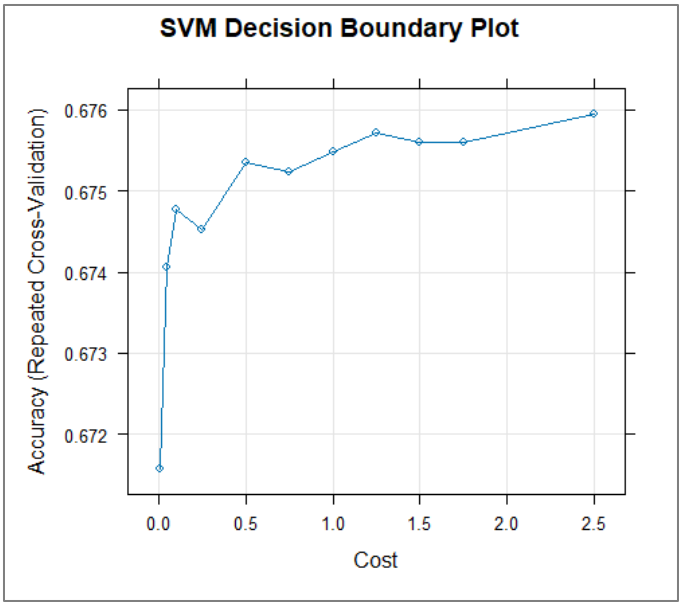


Fig 6: Plot SVM Model

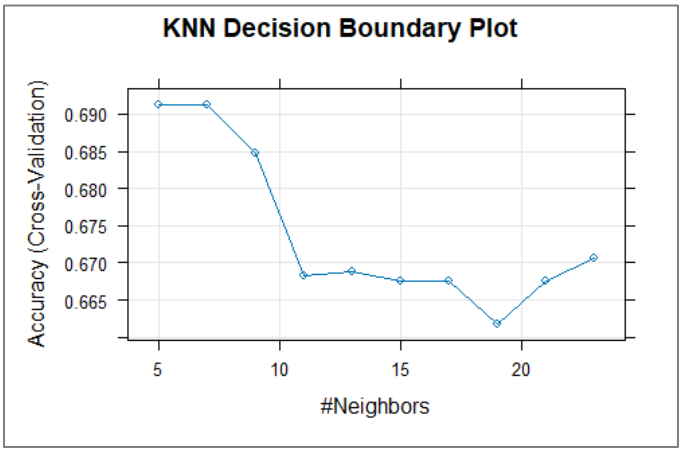
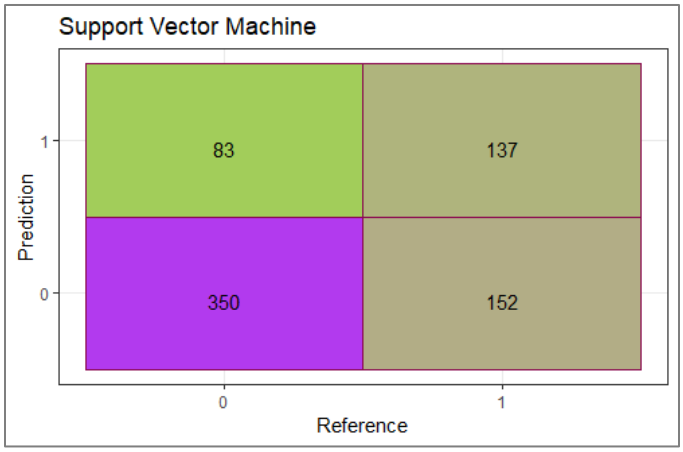


Fig 7: Plot showing KNN Model

Confusion matrix

A tabular representation of an algorithm's output visualisation is called a confusion matrix. Another term for it is an error matrix. The rows of the matrix show instances from real classes, and the columns show instances from predicted classes, or the other way around. The algorithms' confusion matrix is displayed below.



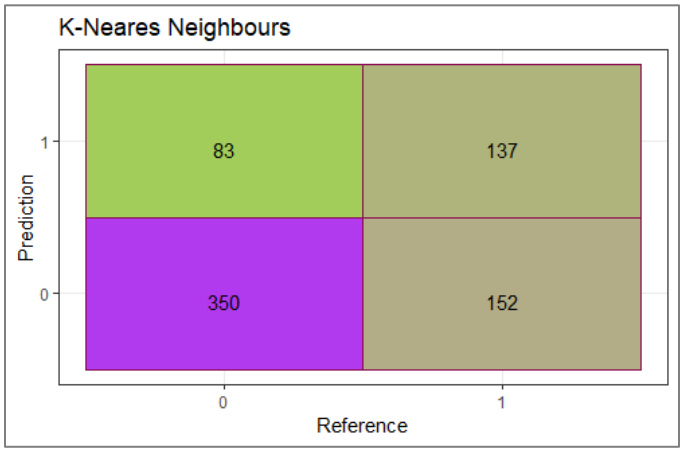


Fig 8: Shows the confusion matrix plot of both models

Comparing the Models

The two models, i.e SVM nad KNN were compared to measure the best performing Model. From the chart shown below, it is observed that the KNN model outperformed the SVM model. This was perfomed with a confidence level of 0.95.

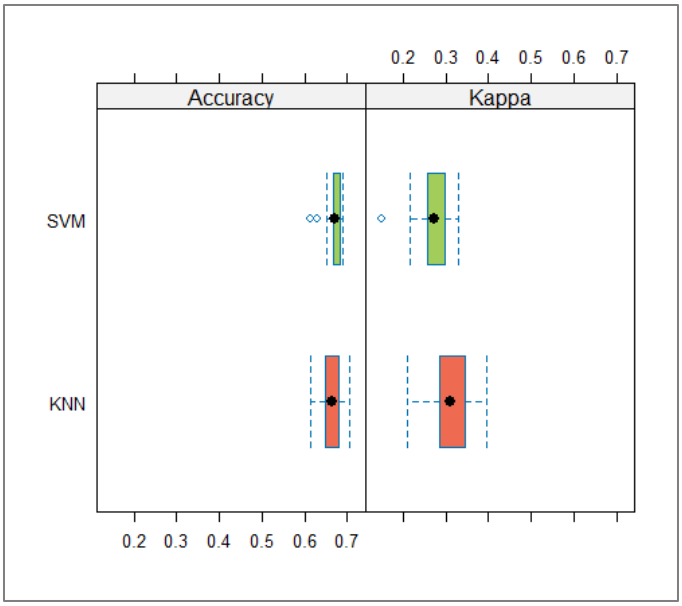


Fig 9: Compares the Accuracy and Kappa results for SVM and KNN

Testing Random Forest

This model was carried out so it can be compared against the best performed model between KNN and SVM. In which case the KNN performed best. So the model was compared with KNN. The results from the Random Forest Model outperformed that os KNN. Results are shown below.

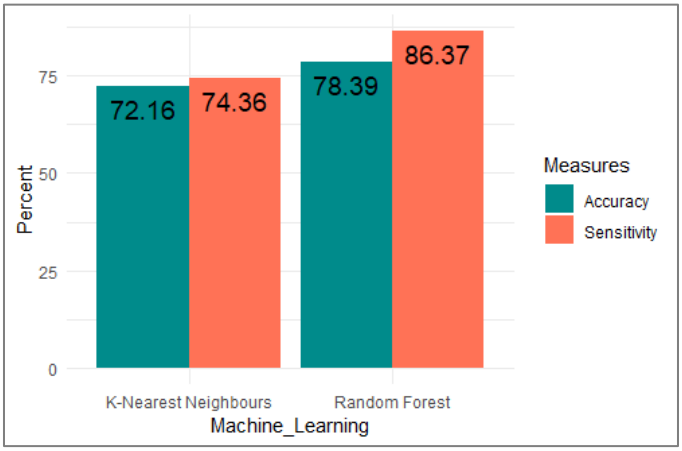


Fig 10: Shows the Accuracy and Sensitivity of KNN and Random Forest Models

RESULT AND CONCLUSION

With an accuracy of 74.16% as opposed to SVM's 70.0%, KNN was the most accurate. Nonetheless, the Random Forest model beat KNN by 8% when compared. The accuracy for the Random Forest model was 78.39%. This is consistent with similar research conducted by [Z. Khan, D. K. Mishra, V. Sharma, and A. Sharma 2020], which offered empirical research on a range of machine learning techniques for predicting coronary heart disease, with the random forest support vector machine, Gauss-naive bayesian, Decision Trees, and K-nearest neighbour.

RECOMMENDATIONS AND FURTHER STUDIES

The K-Nearest Neighbours model was the most accurate in this study for predicting the risk of Coronary Heart Disease over a ten-year period, with a performance rate of 74.16%.

The results must be validated, nevertheless, by comparing the models to additional datasets and investigating alternative models for supervised and unsupervised situations.

Given the low performance, additional research is necessary to determine how well the models work and what strategies will help them perform better. This could lead to the exploration of more sophisticated machine learning models in the future.

REFERENCE

Ahmad, P., S. Qamar, and S.Q.A. Rizvi, Techniques of data mining in healthcare: A review. International Journal of Computer Applications, 2015. 120(15).

Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbour nonparametric regression" The American Statistician. 46 (3): 175–185.

Chen, J., Greiner, R.: Comparing Bayesian Network Classifiers. In Proc. ofUAI-99, pp.101-I08, 1999.

Daniel J., Power, Ramesh Sharda, and Frada Burstein. Decision support systems. John Wiley & Sons, Ltd, 2015.

Fix, Evelyn; Hodges, Joseph L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF) (Report). USAF School of Aviation Medicine, Randolph Field, Texas.

Gupta, S., D. Kumar, and A. Sharma, Performance analysis of various data mining classification techniques on healthcare data. International journal of computer science & Information Technology (IJCSIT), 2011. 3(4).

J. Sony, U. Ansari, D. Sharma and S. Sony, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", (2011)

Jaskowiak, Pablo A.; Campello, Ricardo J. G. B. "Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data". *Brazilian Symposium on Bioinformatics (BSB 2011)*: 1–8. CiteSeerX 10.1.1.208.993

Jyoti Sony, Uma Ansari, Dinesh Sharma, Suita Sony "Predictive data mining for medical diagnosis: an overview of heart disease prediction" International Journal of Computer Science and Engineering, vol. 3, 2011

Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath. "Cascading k-means clustering and k-nearest neighbour classifier for categorization of diabetic patients." International Journal of Engineering and Advanced Technology 1.3 (2012): 147-151.

MUHAMMAD ALI, P. & FARAJ, R. 2014. Data Normalization and Standardization: A Technical Report.

Moreno, María N., Saddys Segrera, and Vivian F. López. "Association Rules: Problems, solutions and new applications." Actas del III Taller Nacional de Minería de Datos y Aprendizaje, Tamida (2005): 317-323

P. Chandra, M. Jabbar, and B. Deekshatulu, "Prediction of Risk Score for Heart Disease using associative Classification and Hybrid Feature Subset Selection," in 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 628–634

Rovina Dbritto, Anuradha Srinivasaraghavan, Vincy Joseph "Comparative Analysis of Accuracy on Heart Disease Prediction using Classification Methods". International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 11 – No. 2, July 2016

S. B. Patil and Y. S. Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction," International Journal of Computer Science and Network Security (IJCSNS), vol. 9, no. 2, pp. 228–235, 2009.

Sadh AS, Shukla N. Association rules Optimization: A survey. International Journal of Advanced Computer Research (IJACR). 2013; 3(9):111-5.

Tolles, Juliana; Meurer, William J (2016). "Logistic Regression Relating Patient Characteristics to Outcomes". *JAMA*. **316** (5): 533–4.

Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/

Y. Uzun and G. Tezel, “Rule learning with machine learning algorithms and artificial neural networks,” Journal of Seljuk University Natural and Applied Science, vol. 1, no. 2, 2012.

NAGAVELLI, U., SAMANTA, D. & CHAKRABORTY, P. 2022. Machine Learning Technology-Based Heart Disease Detection Models. J Healthc Eng, 2022, 7351061.

NHS. 2022. Cardiovascular disease [Online]. Available: https://www.nhs.uk/conditions/cardiovascular-disease/ [Accessed 22 November 2023].

PARTHIBAN, G. & SRIVATSA, S. 2012. Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal of Applied Information Systems, 3, 25-30.

PLATT, JC., 2000. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers. MIT Press, p. 61–74. DOI: 10.1016/j.enpol.2006. 07.010

QIAN, X., LI, Y., ZHANG, X., GUO, H., HE, J., WANG, X., YAN, Y., MA, J., MA, R. & GUO, S. 2022. A Cardiovascular Disease Prediction Model Based on Routine Physical Examination Indicators Using Machine Learning Methods: A Cohort Study. Front Cardiovasc Med, 9, 854287.

UMATHE, P. 2020. Data Science Life Cycle [Online]. Medium. Available: https://medium.com/@pumathe/data-science-life-cycle-4b0b6c4dfeef [Accessed 17 November 2023].

WHO. 2023. Cardiovascular disease [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1/ [Accessed 17 November 2023].

APPENDIXES

Codes written for the project will be shared in a separate document