Predicting Cardiovascular Disease from Patients’ Hospital Data

Abstract

Cardiovascular diseases continue to be one of the leading causes of mortality worldwide, necessitating accurate predictive models for early diagnosis and intervention. This paper presents a comprehensive literature review of the subject and attempts to build an accurate predictive model for cardiovascular diseases. Leveraging a hospital dataset, a rigorous data preprocessing and analysis are conducted to enhance data quality and feature representation. Subsequently, several predictive models including logistic regression, support vector machines, k-nearest neighbours and decision trees are employed to predict cardiovascular diseases with a primary objective of achieving high accuracy. Through systematic experimentation and analysis, the study manages to produce a highly accurate predictive model.

Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.8 million lives each year[1]. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. These conditions can lead to angina, strokes, heart attacks and even heart failure[2], especially if left untreated. Therefore it is vital that these diseases are diagnosed accurately, so that the patient can be provided treatment.

However, in many developing countries, the diagnosis of cardiovascular diseases is often hindered by significant challenges stemming from limited access to healthcare resources and trained medical professionals. That leads to underdiagnosis and inadequate management of these conditions[3]. As a result, a considerable portion of these populations remains at risk of undetected CVDs. It is therefore crucial to find tools that could facilitate CVD diagnosis in areas with few trained medical professionals.

In recent years, advancements in machine learning and predictive analytics have shown promise in improving the accuracy of CVD diagnosis and prognosis. This paper delves into a comprehensive literature review, exploring various models utilised in predicting cardiovascular diseases, with the aim of enhancing early detection and patient care. Based on the knowledge gathered from the literature review, an attempt will be made at developing an accurate predictive model.

Literature Review

There have been numerous attempts at predicting CVDs from hospital data utilising machine learning techniques. These efforts have shown promising results in enhancing early detection and risk assessment. The following paragraphs will describe and critically evaluate a few of the frequently cited papers about the matter.

In Mohan et. al. (2019)[4], the UCI dataset of 297 patients was used to achieve notable progress. The dataset contained 13 features of which 11 were deemed crucial. These vital features were: chest pain type, resting blood pressure, cholesterol, blood-sugar, resting electrocardiogram results, maximum heart rate, angina, oldpeak, slope of peak exercise, number of major vessels, status of heart. Many of those features are prevalent in all papers on the subject, indicating their predictive impact. Numerous machine learning techniques were applied to build an accurate predictive model, including naive bayes, decision trees, support vector machines, logistic regression and many others. In the end, the highest achieved accuracies were at 88.4% using hybrid random forest model and 86 for support vector machines. While the final accuracy is quite high, it leaves room for improvement. Although the appropriate features were selected and numerous machine techniques applied, it seems that the small size of the dataset hindered the result.

Similar problems were encountered in many other papers including Dinesh et. al. (2018)[5] and Shah et. al. (2020)[6], which also used the UCI dataset and tested several different models including the k-nearest neighbours, which yielded the highest accuracy of 90.8%. While they managed to achieve higher accuracy than the previous two papers[4][5], there is still some room for improvement as the size of the dataset appears to be a serious roadblock.

However in Hagan et. al. (2021)[7], which used the same dataset, they were able to achieve a 96% accuracy by using combinations of more sophisticated machine learning techniques with tuned parameters like the extra trees model with gini criterion.

Overall, the studies thus far applied machine learning techniques with various levels of sophistication to achieve high predictive accuracy from relatively small datasets. It seems to promise that an even higher accuracy can be achieved by applying the same machine learning techniques to a larger and high quality dataset.

Dataset

To find the appropriate dataset for the task, a number of factors were taken into account. Training an accurate model requires a considerable amount of high quality data. Therefore it is vital that the dataset selected contains a large number of entries and relevant columns to provide sufficient information. It is also crucial that the dataset is not plagued by multiple missing values, duplicate rows or lack of the most relevant features described in the literature review section.

After the long process of selection, a dataset published by the Lincoln University College[8] (available at https://data.mendeley.com/datasets/dzz48mvjht/1) was chosen as the most suitable for the task. The data comes from a hospital in India and has 1000 entries and 14 columns, which are described in table 1.

Table 1: Cardiovascular Disease Dataset

Attribute	Range and unit	Data type
Patient ID	Unique number	Integer
Age	20-80 years	Integer
Gender	0=Female, 1=Male	Category
Chest Pain Type	0=typical angina, 1=atypical angina, 2=non-anginal pain, 3=asymptomatic	Category
Resting Blood Pressure	94-200 (in millimetres of mercury)	Integer
Serum cholesterol	126-564 (in milligrams per 100 millilitres)	Integer
Fasting blood sugar	0=no more then 120g/100ml, 1=more than 120 mg/100ml	Category
Resting electrocardiogram results	0=normal, 1=having ST-T wave abnormality, 2=showing probable left ventricular hypertrophy	Category
Maximum heart rate achieved	71-202 in beats per second	Integer
Exercise induced angina	0=no, 1=yes	Category
Oldpeak =ST	0-6.2 in millimetres	Float
Slope of the peak exercise ST segment	1=upsloping, 2=flat, 3=downsloping	Category
Number of major vessels	0-3	Category
Classification Target	0=no heart disease, 1=heart disease	Category

Preprocessing

Correctly preprocessing the data before training the models is crucial for achieving high accuracy. The initial preprocessing ensures that there are no empty fields, no duplicate rows in the dataset, that the categorical data is expressed in numbers, that the dataset is balanced in its classification target and that the continuous features are more-less normally distributed. Fortunately, in the case of this dataset, no such issues were present.

The subsequent feature selection process immediately excludes the patient_id features as it is known to have no impact on the classification. Determining which other features to drop will depend on several factors including their correlation with the target and other features, the literature review and the strength of their association with the target. Figures 1 and 2 show the bivariate analysis of categorical and continuous variables respectfully and help understand the predictive potential of variables:

Figure 1: Bivariate analysis for categorical features:

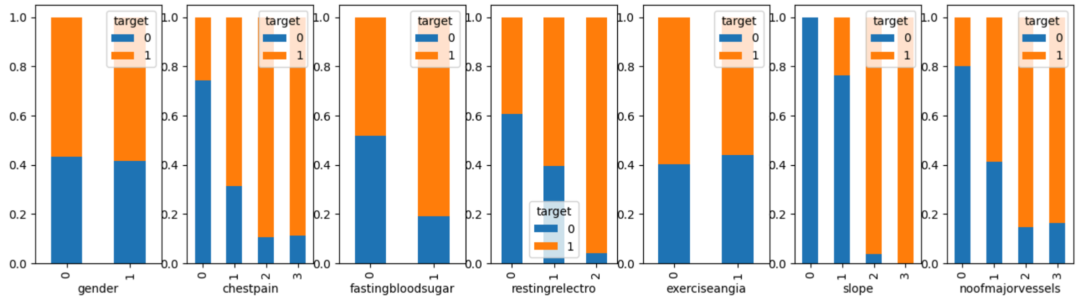


Figure 2: Bivariate analysis for continuous features:

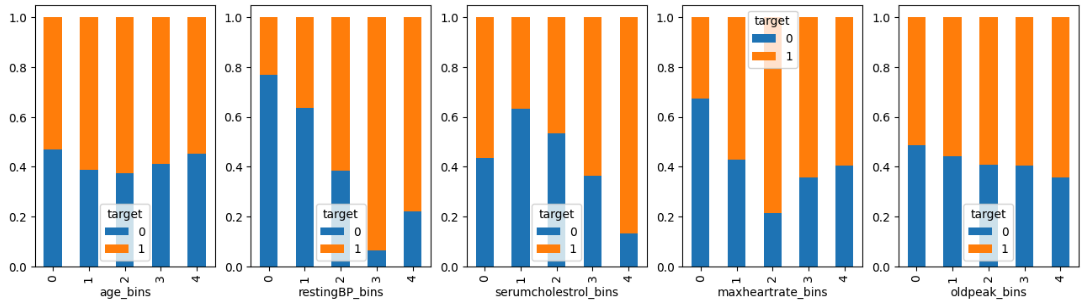


Figure 3 shows the correlation matrix between all the features and figure 4 shows how strongly each variable is associated with the target as a by-product of learning a predictive model (Extra Trees Classifier in this case).

Figure 3: Correlation matrix between all features

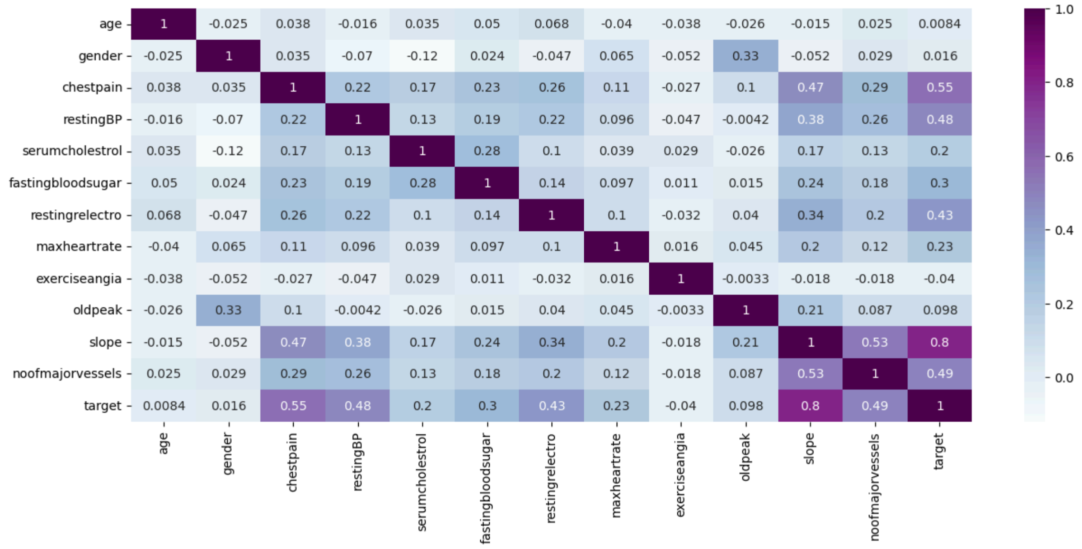
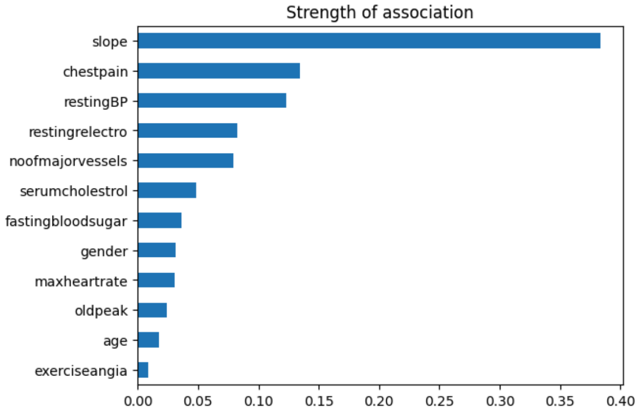


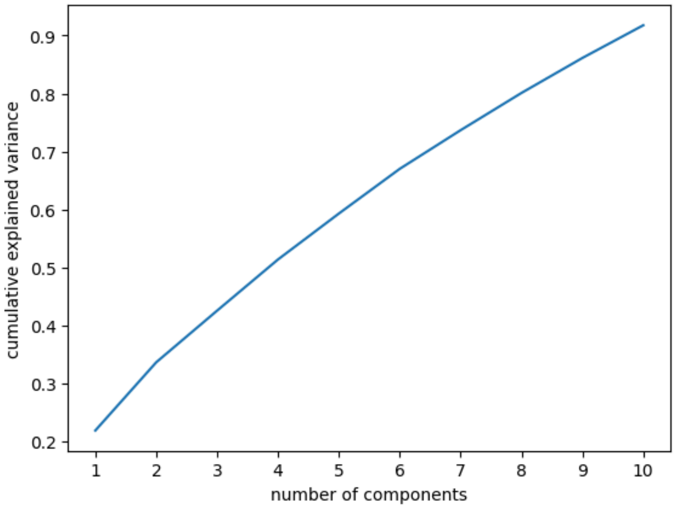
Figure 4: Impact of variables on extra trees predictive model



Across the measures presented in figures 1-4 as well as the literature review seem to indicate that exercise angina, age and gender are the least impactful features in the dataset. Therefore the model will be trained on combinations of features with those 3 features included or removed.

An attempt had also been made at introducing dimensionality reduction, however upon observation of the effect, it can clearly be seen that cumulative explained variance is not hugely affected by the number of components as figure 5 shows. Therefore the dimensionality reduction was not introduced to further proceedings.

Figure 5: Principal component analysis - almost linear cumulative explained variance



Models

The two models that were selected for the purpose of this task were support vector machines and logistic regression. These models have been used in all literature reviewed and consistently yielded high accuracy, both in the referenced papers and this paper.

Support Vector Machine (SVM) classification involves finding the hyperplane that best separates different classes in the feature vector space with the maximum margin. SVM identifies support vectors—data points closest to the decision boundary—and optimises weights and bias to maximise the margin while minimising classification error. SVM is versatile, memory-efficient, and widely used. Therefore it is one of the models used.

Logistic regression estimates coefficients for each predictor variable, which are used to linearly combine the features and produce the log-odds of the outcome. These log-odds are then transformed into probabilities using the logistic function. The model is trained by minimising the logistic loss or cross-entropy loss function, which penalises deviations between predicted probabilities and actual class labels. Logistic regression is a simple yet effective algorithm, widely used in various domains. Therefore it was deemed suitable for the task.

For the purpose of effective learning the data in all predictive columns was normalised to have a mean 0 and a standard deviation of 1. Next the dataset had been split randomly into 80% training data and 20% test data. Subsequently the models were to learn from the training data and output the accuracy, f1, precision and recall on the test data. In addition for every set of features the best value for regularisation was found. The results of that are shown in table 2:

Table 2: Results of trained models on a testset: c: regularisation, a: accuracy, p: precision, r: recall.

	SVM	**Logistic Regression**
All features	c=2.5, a=0.980, p=0.991, f1=0.983, r=0.974	c=0.7, a=0.965, p=0.966, f1=0.970, r=0.974
Without age	c=1, a=0.97, p=0.974, f1=0.974, r=0.974	c=0.7, a=0.965, p=0.966, f1=0.970, r=0.974
Without gender	c=0.5, a=0.980, p=0.983, f1=0.983, r=0.983	c=0.2, a=0.965, p=0.974, f1=0.970, r=0.966
Without exercise angina	c=2.3, a=0.975, p=0.991, f1=0.978, r=0.966	c=0.3, a=0.975, p=0.983, f1=0.979, r=0.974
Without age and gender	c=0.9, a=0.975, p=0.967, f1=0.979, r=0.991	c=0.3, a=0.960, p=0.966, f1=0.966, r=0.966
Without age and exercise angina	c=0.7, a=0.980, p=0.983, f1=0.983, r=0.983	c=1.5, a=0.960, p=0.966, f1=0.966, r=0.966
Without gender and exercise angina	c=4, a=0.985, p=0.983, f1=0.987, r=0.991	c=0.5, a=0.965, p=0.974, f1=0.970, r=0.966
Without all 3	c=3, a=0.985, p=0.975, f1=0.987, r=1	c=0.3, a=0,960, p=0.966, f1=0.966, r=0.966

Conclusion

The results from the implemented models can definitely be considered a success. While the results were similar for all configurations, the highest accuracy, precision, recall and f1 were achieved by the support vector machine model when all 3 selected features were removed. The achieved accuracy is higher than most of the accuracies from the literature review, which seems to prove that using a larger dataset leads to even better results.

Although the results achieved in this study are satisfactory with accuracy of over 98%, it still could be possible to train better models. Despite the fantastic performance of the model, it is important to acknowledge the limitations of this study. The main limitation is that the dataset comes from a single hospital and might therefore not be sufficiently diverse or contain biases. Hence when considering possible attempts at improvement, one should consider using larger and more diverse datasets, including extra features (perhaps graphical) or experimenting with other machine learning or deep learning models.

Bibliography and Resources:

1. Kaptoge, S., Pennells, L., De Bacquer, D., Cooney, M.T., Kavousi, M., Stevens, G., Riley, L.M., Savin, S., Khan, T., Altay, S. and Amouyel, P., 2019. World Health Organization cardiovascular disease risk charts: revised models to estimate risk in 21 global regions. *The Lancet global health*, *7*(10), pp.e1332-e1345.
2. NHS, n.d. Cardiovascular disease, NHS website. Available at: https://www.nhs.uk/conditions/cardiovascular-disease/ [Accessed March 10th 2024]
3. Celermajer, D.S., Chow, C.K., Marijon, E., Anstey, N.M. and Woo, K.S., 2012. Cardiovascular disease in the developing world: prevalences, patterns, and the potential of early disease detection. *Journal of the American College of Cardiology*, *60*(14), pp.1207-1216.
4. Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, pp.81542-81554.
5. K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Coimbatore, India, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550857
6. Shah, D., Patel, S. and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*(6), p.345.
7. Hagan, R., Gillan, C.J. and Mallett, F., 2021. Comparison of machine learning methods for the classification of cardiovascular disease. *Informatics in Medicine Unlocked*, *24*, p.100606.a
8. Doppala, Bhanu Prakash; Bhattacharyya, Debnath (2021), “Cardiovascular_Disease_Dataset”, Lincoln University College, Mendeley Data, V1, Available at: https://data.mendeley.com/datasets/dzz48mvjht/1 [March 10th 2024]