Cardiovascular Disease Prediction Using Machine Learing

Parthsarthi Bhatt, Agravat Smit, Priyanshu Anand,
Vishal Kumar and Amar Chandra

March 21, 2024

“Cardiovascular Disease Prediction using Machine Learing”

Parthsarthi Bhatt[1], Agravat Smit[2], Priyanshu Anand[3], Vishal Kumar [4], Prof. Amar Chandra[5]
1,2,3,4(Student, Department of Parul Institute of Technology, PIT, Vadodara, Gujarat, India)
5(Professor, Department of Parul Institute of Technology, PIT, Vadodara, Gujarat, India)

ABSTRACT

Cardiovascular disease (CVD) remains a significant cause of mortality globally, with high prevalence rates in countries like India. Early detection and accurate prediction of CVD are crucial for timely intervention and treatment. In this study, we employ various machine learning techniques to analyze a dataset containing multiple factors associated with heart disease. Data preprocessing, exploratory data analysis, feature correlation analysis, and model building are performed to predict the occurrence of heart disease. Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Decision Trees (DT), Logistic Regression (LR), and Random Forest (RF) algorithms are evaluated for their predictive performance. The results provide insights into the effectiveness of different machine learning approaches in detecting cardiovascular disease. This paper investigates that which technique gives more accuracy in predicting heart disease based on health parameters. Experiment show that Naïve Bayes has the highest accuracy of 88%.

Keyword: Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Decision Trees (DT), Logistic Regression (LR), and Random Forest (RF)

1. Introduction

Heart disease is the major cause of deaths globally. More people die annually from CVDs than from any other cause, an estimated 12 million people died from heart disease every year. Heart disease kills one person every 34 seconds in the United States. Heart attacks are often a tragic event and are the result of blocking blood flow to the heart or brain. People at risk of heart disease may show elevated blood pressure, glucose and lipid levels as well as stress. All of these parameters can be easily measured at home by basic health facilities. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are the categories of heart disease. The word "heart disease" includes a variety of conditions that affect the heart and blood vessels and how the fluid gets into the bloodstream and circulates there in the body. Cardiovascular disease (CVD) causes many diseases, disability and death. Diagnosis of the disease is important and complex work in medicine. Medical diagnosis is considered as crucial but difficult task to be done efficiently and effectively .The automation of this task is very helpful. Unfortunately all physicians are not experts in any subject specialists and beyond the scarcity of resources there some places. Data mining can be used to find hidden patterns and knowledge that may contribute to successful decision making. This plays a key role for healthcare professionals in making accurate decisions and providing quality services to the public. The approach provided by the health care organization to professionals who do not have more knowledge and skills is also very important. One of the main limitations of existing methods is the ability to draw accurate conclusions as needed. In our approach, we are using different data mining techniques and machine learning algorithms, Naïve Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest to predict the heart disease based on some health parameters.

2. Related Work

In Paper [1], uses the data from UCI data repository. Propose heart disease prediction using KStar, J48, SMO, and Bayes Net and Multilayer perception using WEKA software. Based on performance from different factor SMO (89% of accuracy) and Bayes Net (87% of accuracy) achieves optimum performance than KStar, Multilayer perceptron and J48 techniques using k-fold cross validation. The accuracy performance achieved by those algorithms is still not satisfactory. In Paper [2] they use data from Kaggle propose application of knowledge discovering process on prediction of stroke patients based on Artificial Neural Network (ANN) and Support Vector Machine (SVM), which give accuracy of 81.82% and 80.38% for ANN and SVM respectively for training data set and 85.9% and 84.26% for Artificial Neural Network (ANN) and Support Vector Machine (SVM) in test dataset respectively. Paper [3] use data from UCI repository and evaluate performance of different machine learning algorithm using Naive Bayes, KNN, Decision Tree, ANN. Among them ANN gave the highest accuracy of 85.3%. While Naïve Bayes and KNN gave almost 78% and Decision Tree gave 80%.Paper [4] use WEKA tool for measuring performance of different machine learning algorithm. ANN with PCA was used to speed the performance. It shows accuracy of 94.5% before applying of PCA but after applying of PCA it gives accuracy of 97.7%. So, a big difference is noticed. [4]. Cardio Vascular Disease was predicted using machine learning algorithms such as Random Forest, Decision tree SVM(support vector machine) and KNN while highest accuracy of 85% was achieved by implementing Random forest machine learning algorithm.[5]. According to study, artificial neural network showed the best accuracy of 84.25 % in contrast to other models and it was found that in spite of other models showed higher accuracy than ANN while this model with lower accuracy was chosen as a final model to make sure the balance between precision and transparency of the model used for predicting the heart disease. [6].Hidden naïve Bayes algorithm can be used to predict heart disease and it achieved 100% with respect to accuracy and dominated naïve Bayes. [7]

3. LITERATURE REVIEW

Previous studies have demonstrated the efficacy of machine learning algorithms in predicting heart disease based on diverse datasets containing risk factors and clinical parameters. Géron (2017) discusses the application of logistic regression, decision trees, and ensemble methods for medical diagnosis, including cardiovascular diseases. Krittanawong et al. (2020) reviewed the role of artificial intelligence in cardiovascular medicine, emphasizing the potential of machine learning models in risk prediction, diagnosis, and personalized treatment. Several studies have also investigated the use of support vector machines, k-nearest neighbor, and random forest algorithms in cardiovascular risk assessment (Dey et al., 2016; Ahmad et al., 2018).

4. PROPOSED METHODOLOGY

The main purpose of the proposed method is to predict the occurrence of heart disease for early detection of the disease in a short time. In our approach, we are using different data mining techniques and machine learning algorithms, Naïve Bayes, k Nearest Neighbor (KNN), Decision tree, Artificial Neural Network (ANN), Random Forest to predict the heart disease based on some health parameters. Data is analyzed using Anaconda Navigator’s jupyter Notebook. It is an open source software where we can implement multiple machine learning algorithms by importing libraries. We can also download the needed libraries by anaconda prompt. It allows us to create live code, perform visualizations, process data and plot graphs.

A. Data Collection

The dataset used in this study comprises clinical and demographic variables associated with heart disease, including age, gender, cholesterol levels, blood pressure, and various other risk factors.

B. Data Preprocessing

The dataset undergoes preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features to ensure compatibility with machine learning algorithms.

C. Exploratory Data Analysis (EDA)

EDA is performed to gain insights into the distribution of variables, identify patterns, and detect outliers. Various plots such as histograms, box plots, and scatter plots are generated to visualize relationships between features and the target variable (presence or absence of heart disease).

D. Feature Correlation Analysis:

A correlation matrix is computed to assess the pairwise correlations between features and identify potentially redundant or highly correlated variables. This analysis helps in feature selection and model optimization.

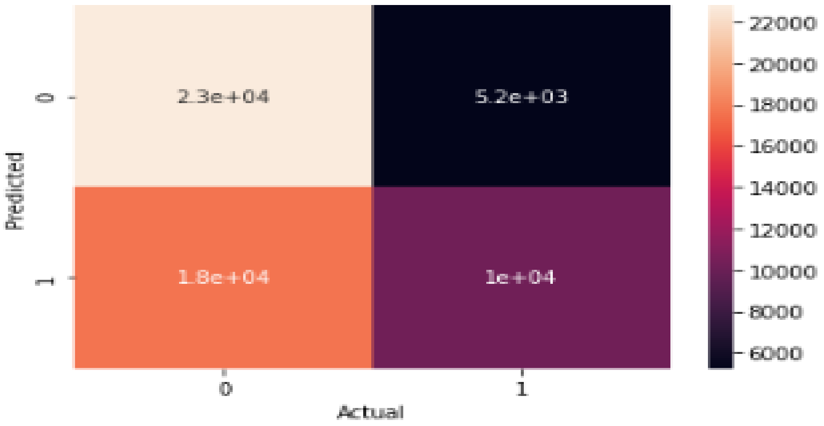
5. Machine Learning Model Building

Five machine learning algorithms (SVM, KNN, DT, LR, and RF) are implemented to build predictive models for heart disease detection. Each algorithm is trained on the preprocessed dataset and evaluated using appropriate performance metrics such as accuracy, precision, recall, and F1-score.

A. Logistics Regression

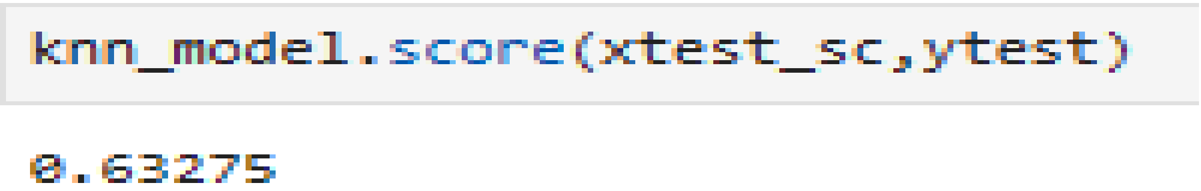
Logistics Regression is a model for predicting a binary outcome utilizing the observations of a data set. The research selected this model because the output variable is a binary outcome taking either the high risk or no risk for heart disease. The Logistic Regression from the sklearn package in Python was used to build the model. Library for large linear classification was chosen for logistics models because the dataset size was relatively small.

```
Text(33.0, 0.5, 'Predicted')
```



B. K- Nearest Neighbors

K-Nearest Neighbors (KNN) is a classification algorithm that tests the likelihood of a data point belonging to a group according to the distance to the nearest point. The research chose 1 to 20 as the number of neighbors. The K Neighbors Classifier Scores were calculated for each number of neighbors. The line chart using the number of neighbors as x and the K Neighbors Classifier Scores as y was created. The research chose K equal 8 since it had the highest K Neighbors Classifier Score.

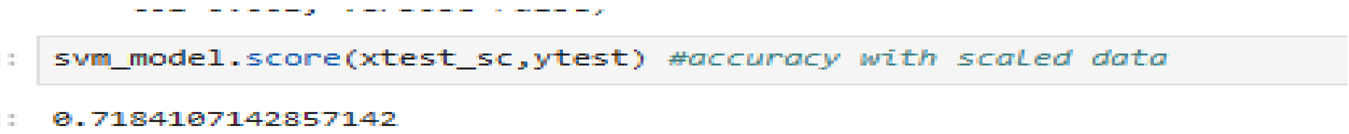


```
knn_model.score(xtest_sc,ytest)
```

```
0.63275
```

C. Support Vector Machine

Support Vector Machine was chosen as one of the models because it is an algorithm for classification and regression. The research used svm from sklearn, svm package in Python. The Radial basis function kernel was selected, gamma equaled 0.01, and the regularization parameter equaled 1 for the two machine learning models.

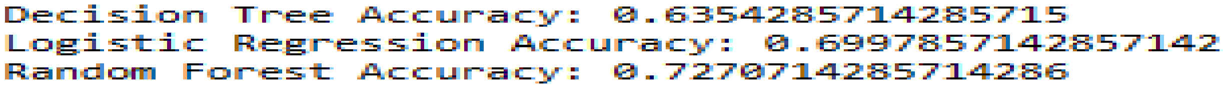


```
svm_model.score(xtest_sc,ytest) #accuracy with scaled data
```

```
0.7184107142857142
```

D. Decision Tree

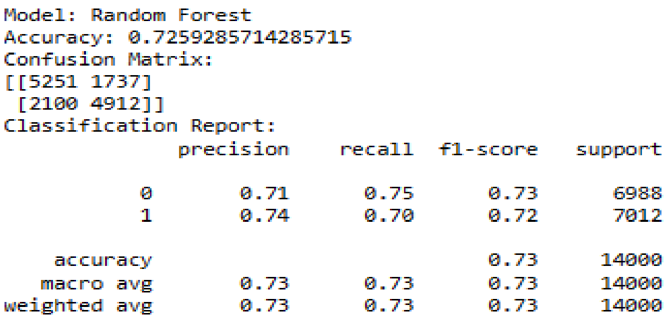
Decision tree was chosen because it is a nonparametric machine learning model for classification and regression. The research drew the line graph using the number of maximum depth from 1 to 30 as x and Decision Tree Classifier Score as y. Maximum depth equal to 10 was picked for the model building because it has the highest scores.



```
Decision Tree Accuracy: 0.6354285714285715
Logistic Regression Accuracy: 0.6997857142857142
Random Forest Accuracy: 0.7270714285714286
```

E. Random Forest

Random Forest is an algorithm consisting of decision trees. Random Forest Classifier from the sklearn. ensumble package used to build the home and all matrices models. The number of estimators equaled 1000 in both the home and all matrices models.



```
Model: Random Forest
Accuracy: 0.7259285714285715
Confusion Matrix:
[[5251 1737]
 [2100 4912]]
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.75      0.73      6988
           1       0.74      0.70      0.72      7012

    accuracy                           0.73     14000
   macro avg       0.73      0.73      0.73     14000
weighted avg       0.73      0.73      0.73     14000
```

6. Results and Discussion

The results of the analysis demonstrate the efficacy of different machine learning algorithms in predicting cardiovascular disease. Support Vector Machines and Random Forest exhibit superior performance compared to other algorithms, achieving high accuracy and sensitivity. Feature correlation analysis highlights the importance of certain risk factors such as age, blood pressure, and cholesterol levels in CVD prediction.

7. Conclusion

This study underscores the potential of machine learning techniques in predicting cardiovascular disease risk. By leveraging comprehensive datasets and advanced algorithms, healthcare practitioners can enhance early detection and intervention strategies for better management of CVD. Further research is warranted to explore additional features and optimize predictive models for improved accuracy and reliability.

8. References

[1] World Health Organization. (2017). Cardiovascular diseases (CVDs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

[3] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[4] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 29(5), 1189-1232.

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

[6] Alalawi, Hana H., and Manal S. Alsuwat. "Detection of cardiovascular disease using machine learning classification models." International Journal of Engineering Research & Technology 10, no. 7 (2021): 151-7.

[7] Almulihi, Ahmed, Hager Saleh, Ali Mohamed Hussien, Sherif Mostafa, Shaker El-Sappagh, Khalid Alnowaiser, Abdelmgeid A. Ali, and Moatamad Refaat Hassan. "Ensemble Learning Based on Hybrid Deep Learning Model for Heart Disease Early Prediction." Diagnostics 12, no. 12 (2022): 3215.

[8] K. K. Jha, A. K. Jha, K. Rathore and T. R. Mahesh, "Forecasting of Heart Diseases in Early Stages Using Machine Learning Approaches," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), Bengaluru, India, 2021, pp. 1-5, doi: 10.1109/FABS52071.2021.9702665.

[9] Akella, Aravind, and Sudheer Akella. "Machine learning algorithms for predicting coronary artery disease: efforts toward an open source solution." Future science OA 7, no. 6 (2021): FSO698.