Heart Disease Prediction using Machine Learning Algorithms.

Varun M. Parekh Vanshit M. Shah
Ushmay H. Patel Madhvendra K. Jhala

***Abstract*— Heart disease is a leading cause of mortality worldwide. Early diagnosis and accurate prediction of heart disease can significantly improve patient outcomes. Machine learning algorithms have shown promising results in predicting heart disease risk. Our project aims to develop a machine learning-based prediction model for heart disease using a dataset of patients' clinical and demographic information.**

KeyWords— Classification; feature selection; training and testing; heart disease prediction; Naïve Bayes; Logistic Regression; Accuracy;

I. Introduction:

We explored the application of machine learning algorithms in predicting heart disease and improving the accuracy of diagnosis. We review the current state of research in this field, discuss the challenges and opportunities for using machine learning in heart disease prediction, and highlight some of the most promising algorithms and approaches. Some of the methodologies used for Heart-Disease Prediction in this paper are Naive Bayes, Random Forest, K-Nearest Neighbors and Decision tree. This report aims to provide a comprehensive overview of the use of the mentioned machine learning algorithms in predicting heart disease, and also to positively contribute to the development of effective and accurate diagnostic tools for the very concerning disease(i.e., Heart-Disease).

II. Literature Survey

Sonam Nikhar et al proposed paper " Prediction of Heart Disease Using Machine Learning Algorithms" their research gives point to point explanation of Naïve Bayes and decision tree classifiers that are used especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has higher accuracy than Bayesian classifier.

Avinash Golande et al, proposed "Heart Disease Prediction Using Effective Machine Learning Techniques" in which few data mining techniques are used that support the doctors to differentiate the heart disease. Usually utilized methodologies are k-nearest neighbor, Decision tree and Naïve Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel selfarranging guide and SVM (Bolster Vector Machine).

Anjan N. Repaka et al, proposed a model stating the performance of prediction for two classification models, which is analyzed and compared to previous work. The experimental results show that accuracy is improved in finding the percentage of risk prediction of our proposed method in comparison with other models.

III. Implementation

The dataset used in this study was obtained from the UCI Machine Learning Repository, and it included various attributes such as age, gender, cholesterol levels, and blood pressure, among others. Our dataset consisted of 303 instances. We started our project by importing the required libraries such as numpy, pandas and matplotlib etc. As already before the mid-sem we found the dataset with 14 most significant features out of the 76 features and had implemented Naive Bayes and Logistic Regression. After that we plotted all the features and differentiated them categorically based upon the features and their categories. Then we plotted the distribution and the box-plots for all the features and then categorized the features based on the count of "has_disease" and "no_disease". After that we plotted the correlation matrix between the features and the condition. We plotted the correlation matrix and found out that exercise induced angina had the strongest correlation with our condition, that is whether the patient

had heart disease or not. Further we implemented the K-Nearest Neighbors algorithm and since the dataset that we used had only 303 values hence we used the k-fold(k is 5) technique in-order to train the model. On a similar basis we implemented the Random Forest Classifier, Gaussian Classifier, Decision Tree Classifier and Support Vector Classifier. We also performed a model check for every algorithm used and calculated parameters like ROC/AUC , accuracy and F1 score Mean and found out that decision tree based algorithms were overfitted since their train and test accuracy mean had a large difference. So we plotted the feature importance for each of the features for the Random Forest Classifier and the Decision Tree Classifier.

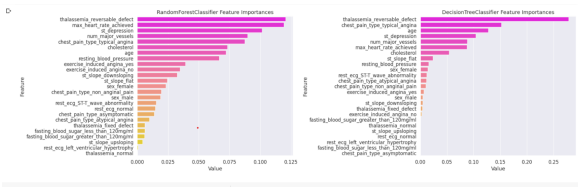
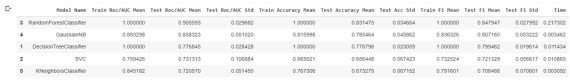


Fig. 1 Depicting feature importance

We also plotted the confusion matrix for each of the classifiers in-order to calculate the accuracy of the model and finally we calculated the accuracy of all the models and also calculated the best hyperparameters for each of the models implemented and calculated the best-score achieved for each of the models.

IV. Results

"Post-midsem we have worked on different datasets and also understood the concept of hyper parameter tuning. We plotted a confusion matrix to evaluate the performance of different classifiers and to calculate performance metrics like accuracy, F1 score, recall and precision.



	Model Name	Train Roc/AUC Mean	Test Roc/AUC Mean	Test Roc/AUC Std	Train Accuracy Mean	Test Accuracy Mean	Test Acc Std	Train F1 Mean	Test F1 Mean	Test F1 Std	Time
3	RandomForestClassifier	1.000000	0.905593	0.029682	1.000000	0.831475	0.034964	1.000000	0.847947	0.027952	0.217302
4	GaussianNB	0.859298	0.856323	0.051020	0.815998	0.798464	0.040962	0.836326	0.807160	0.003222	0.003462
1	DecisionTreeClassifier	1.000000	0.776845	0.028428	1.000000	0.778798	0.023009	1.000000	0.799462	0.019614	0.011434
2	SVC	0.759426	0.731313	0.106884	0.665221	0.656446	0.067423	0.732524	0.721329	0.056517	0.010865
0	KNeighborsClassifier	0.845162	0.720570	0.051450	0.767306	0.673279	0.067152	0.791601	0.708466	0.070601	0.003092

Fig. 2 Model check

After applying that on the selected classifiers we ended up with optimized results. We had string inputs so we had to perform one hot encoding before we could perform any analysis on it. We used Bin discretizer for it. We also found out the confusion matrix for every classifier and it also depicted that the decision tree and random forest were overfitted. We were formulating it to calculate F1 score, accuracy , precision and recall but due to Some classifiers showed a significant boost whereas some had a very little change. Random Forest Classifier worked the best for our dataset giving an accuracy of 88.52%."

Algorithm used for prediction	Accuracy(in %)
Random Forest	88.52
Support Vector Classifier	86.88
K-Nearest Neighbors	83.60
Decision Tree Classifier	80.32
Gaussian Naive Bayes	78.68

V. Conclusion:

In conclusion, the study conducted on heart disease prediction using Machine Learning showed that all classifiers can effectively predict the occurrence of heart disease in patients. However, Random Forest outperformed other classifiers in terms of accuracy, precision, and recall. We can also optimize the results by cleaning the outliers in our data with techniques such as Isolation Forest and Elliptic envelope. The results obtained from the study suggest that Machine Learning models can be utilized as effective tools in predicting heart disease, thereby aiding healthcare professionals in the early detection and prevention of heart disease. Furthermore, the use of Machine Learning models in predicting heart disease can help reduce the high mortality rate associated with heart disease worldwide.

Overall, this study highlights the potential benefits of using Machine Learning models in healthcare and underscores the need for further research in this area to improve the accuracy and reliability of these models.

References

1) A. S. Abdullah and R. R. Rajalaxmi, *"A data mining model for predicting the coronary heart disease using random forest classifier,"* in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25.
2) Chen A H, Huang S Y, Hong P S, Cheng C H & Lin E J (2011, September). *HDPS: Heart disease prediction system.* In 2011 Computing in Cardiology (pp. 557-60). IEEE.
3) J. Thomas and R. T. Princy, ''*Human heart disease prediction system using data mining techniques*,'' in Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT), Mar. 2016, pp. 1–5.
4) S. Radhimeenakshi, ''*Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network*,'' in Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom), New Delhi, India, Mar. 2016, pp. 3107–3111.
5) J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, ''*Association rule mining to detect factors which contribute to heart disease in males and females,*'' Expert Syst. Appl., vol. 40, no. 4, pp. 1086–1093, 2013. doi: 10.1016/j.eswa.2012.08.028.