International Journal of Science Academic Research
Vol. 04, Issue 05, pp.5600-5603, May, 2023
Available online at http://www.scienceijsar.com

ISSN: 2582-6425

Research Article

PREDICTING RISKS OF DEVELOPING HEART DISEASES USING MACHINE LEARNING ALGORITHMS

***Shreya Podishetti, Sri Ganesh Kakkerla and Surya Podishetti**

Department of Information Technology, Kakatiya Institute of Technology and Science, Warangal, India

Received 29th March 2023; **Accepted** 24th April 2023; **Published online** 26th May 2023

Abstract

Heart diseases are known by many names around the world such as arrhythmias, heart failure, cardiac arrest, cardiovascular diseases, etc. People around the world are victims of developing a risk to catch these diseases in the long term go. Keeping these risk-developing patterns in mind, it is possible to train a machine in predicting whether a person will develop a heart condition in the next ten years. The machine learning models used in this paper are decision trees, random forests, KNN, and K-Means. This research paper aims to explore the application of machine learning techniques to better identify heart diseases in advance without the interference of any medical practitioner.

Keywords: Machine Learning, Decision Trees, Random Forest, K-means, KNN.

INTRODUCTION

In recent times, cardiovascular health problems are something that can be heard in any household. Unlike the older times, health trends are changing rapidly due to changing lifestyles and human habits. Some such hazardous habits include smoking, vaping, increasing alcohol intake, consuming fast foods, and lack of exercise. Some people might understand the consequences of developing any heart-related problems and hence choose to take the needed precautions for it. Trying to help such desired people, this research helps predict whether a person has the risk of being affected by cardiac trends. In doing so, we are using a machine learning model which takes the details of the patient which include whether the person smokes, and if he does then how many cigarettes he smokes per day, what is the person's education, the person's age, gender, BMI, BP, heart rate and related medical information. We next try to feed this data to a couple of models namely decision trees, random forests, KNN, and K-Means. Each of these models has its pros and cons which causes the accuracy rate of individual model's performance to vary. Comparing these values among each other, helps us to identify the best-suited Machine Learning model for solving our problem. The models we are using in this research include supervised learning models as well as unsupervised learning models which can be considered as an important distinction in choosing them particularly. Through such diversities, the decision in choosing the appropriate and suitable model for solving our problem at the end is the main aim of this research. By the end of this process, an accuracy of 85.07% can be reached to correctly identify the health risk of the patient in the coming ten years.

LITERATURE REVIEW

There are several research surveys exploring different ways to deal with cardiovascular problems in patients using many emerging technologies such as IOT, Data Mining, Artificial Intelligence, Machine Learning, and Deep Learning.

Corresponding Author:** ***Shreya Podishetti,
Department of Information Technology, Kakatiya Institute of Technology and Science, Warangal, India

An article by Patel, Jaymin. (2015) proposed a data mining approach for the early detection of heart failures. The authors used advanced and special data mining techniques to properly mine industry healthcare data for better decision making and later utilize machine learning models such as Decision Trees and Naïve Bayes. This way of approach has drastically improved the accuracy of the prediction process. Another study by S. Mohan. (2019) proposed hybrid Machine Learning models to address the same issue. This method is also known as the ensemble model. As the name is suggesting, it is the combination of more than one model. By following this methodology, the flaws of one model can be covered and strengthened by the perks of another model. As a result, the overall accuracy scores of these hybrid models are largely highlighted. The proposed ensemble models include KNN and SVS which could show prediction accuracy which reaches up to 90%.

METHODOLOGY

Collecting Data

The methodology opted for this research is not very different from any other machine learning models. In any predictive project, to obtain the utmost perfection, it is important to have abundant data. Accordingly, gathering a precise and sufficient quantity dataset happens to be the first step in this research. A very well know online community known as Kaggle has a wide range of datasets available for aspiring and professional data scientists and machine learning developers to download and use for. Making use of such great availability, a dataset named as *Framingham dataset* is being used for this research. The size of the dataset is 4240 rows and the 16 features present in this dataset are shown in the figure below.

Data Cleaning

It is well known that in reality, datasets can be messy or noisy. To make sure that the results from training the model are on point, the dataset requires some data cleaning and data transformation to be more effective. The dataset used here contains some missing values as well as some outliers.

```
male               int64
age                int64
education        float64
currentSmoker      int64
cigsPerDay       float64
BPMeds           float64
prevalentStroke    int64
prevalentHyp       int64
diabetes           int64
totChol          float64
sysBP            float64
diaBP            float64
BMI              float64
heartRate        float64
glucose          float64
TenYearCHD         int64
dtype: object
```

Fig. 1. Features of Dataset

To deal with the missing values, there are a couple of ways which include dropping all the rows which contain missing values or dropping an entire column if it has too many null spaces, or collecting the exact data by talking to the specialists in that specific field and filling in those values or finally filling the mean, median or mode values of the entire feature. Dropping columns or rows in this case can cause the size of the dataset to decrease which is not recommended as having more data can help the model to be trained on more cases and hence show equal results in the testing phase also. Keeping this in mind the method opted to handle the missing data in this research is by filling in the mean values of the features. Using data visualization techniques, we can identify the outliers in the entire dataset, and with that reference, those rows are dropped from the dataset which produces clean data.

Feature Selection and Feature Scaling

Apart from the basic data cleaning performed so far, there are two more techniques called feature selection and feature scaling which will help us identify and use only the important features in our training process. Applying a feature selection check among the input features, showed that the currentSmoker and heartRate fields are of minute importance, which gives us the freedom to eliminate these two features from the training dataset. The method we are using for feature scaling is the min-max scaler which will prevent the results to be biased.

KNN Model

The next step in this process is fitting the data to each model. Before fitting the data, we are going to split the dataset into two parts where one part is used for the training and the other part is used for the testing of the model. The training data will be 80% of the whole dataset leaving 20% for testing. We first choose to train and test the data on KNN supervised model. KNN stands for K- Nearest Neighbours. In this model, it is important to select the right value for K which is the number of nearest neighbours. To do so, we will check the error rate associated which each value of K by plotting them against each other and then estimate the right value. The graph is shown below.

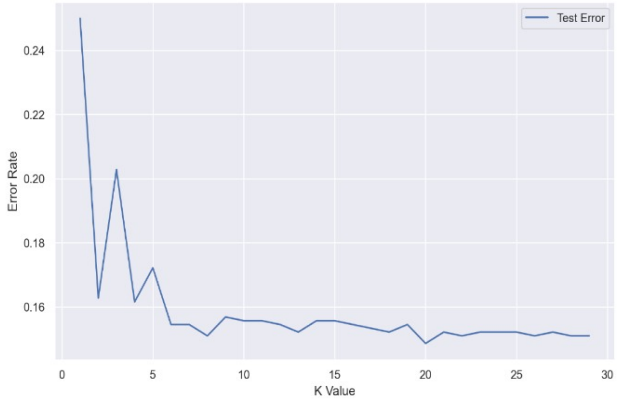

Fig 2 K-Value Vs Error Rate graph of KNN

By carefully observing the K value from where there are fewer spikes, the best value to give K is 14. In the next step, we are going to pass K as 14 and start training the data on the KNN model using the training data. By finally comparing both the test value and predicted value, the KNN classifier is giving an overall accuracy of 84.39999999999999%.

```
In [48]: acc = accuracy_score(y_test, knn_predict)
         print(f"The accuracy score for KNN is: {round(acc,3)*100}%")

         The accuracy score for KNN is: 84.39999999999999%
```

Fig 3 Accuracy of KNN

Decision Trees

Decision Trees are one of the most fundamental machine learning models which work on given conditional rules. This can be implemented on both classification and regression tasks using supervised learning methods. Even in this case, we will be dividing the dataset into 80% training data and 20% testing data and then fitting the training data to the model. Comparing the predicted result to the test result, the following confusion matrix is obtained which can be useful in understanding the models, accuracy, precision, F1 score, and recall values to judge the model.

```
In [50]: print(classification_report(y_test,base_pred))

                   precision    recall  f1-score   support

              0.0       0.86      0.81      0.83       720
              1.0       0.18      0.23      0.20       128

         accuracy                           0.72       848
        macro avg       0.52      0.52      0.52       848
     weighted avg       0.75      0.72      0.74       848
```

Fig. 4. Confusion Metrics of Decision Trees

This model has given an accuracy score of 72.39999999999999%, which is lower than the accuracy score of the KNN model. The reason for this can be that it overfitted the data.

```
In [51]: acc = accuracy_score(y_test, base_pred)
         print(f"The accuracy score for Decision Trees is: {round(acc,3)*100}%")

         The accuracy score for Decision Trees is: 72.39999999999999%
```

Fig. 5. Accuracy of Decision Trees

Random Forests

In the decision trees, we have seen the case of overfitting which lead the lower accuracy than expected. To improve that accuracy, we have another model named Random Forests which in simple terms means the combinations of more than one decision trees structures. This can lead to a more powerful version of the decision tree algorithm. In this model, we have to provide the number of trees we need in advance. To select the correct value for this model, we can do the same procedure we did for the KNN model which is plotting the number of trees against the error rate associated. The resultant graph can be seen in the figure
below.

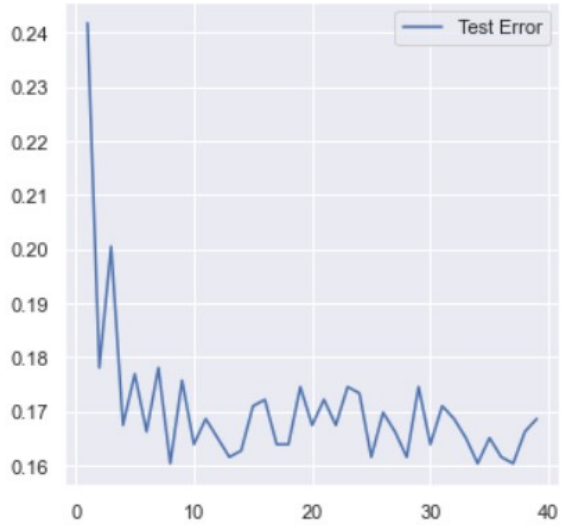

Fig. 6. Number of Trees Vs Error Rate graph of Random Forests

Studying the graph above shows the best value to select is 2. Now we can take this and start training the data. The final confusion matrix of this model shows that the performance has improved from the decision tree model.

	precision	recall	f1-score	support
0.0	0.85	0.98	0.91	720
1.0	0.33	0.05	0.09	128
accuracy			0.84	848
macro avg	0.59	0.52	0.50	848
weighted avg	0.78	0.84	0.79	848

Fig. 7. Confusion Metrics of Random Forests

Now the accuracy rate can be seen to be 84.1%.

```
In [72]: acc = accuracy_score(y_test, predictions)
         print(f"The accuracy score for Random Rofests is: {round(acc,3)*100}%")

         The accuracy score for Random Rofests is: 84.1%
```

Fig. 8. Accuracy of Random Forests

K-Means

Till now, the three models that we worked on are supervised learning algorithms that need us to provide the model with the answers during the training process. But now we are going to check with the case of an unsupervised learning model called K-Means. This model does not require us to provide the results of the classification in advance. The logic behind this concept is the calculation of the Euclidean distance between the plotted data points. Based on the distances, the model will form a K number of clusters and classify the data accordingly. These clusters are nothing but the group of similar types which in our case the value of K will be 2 as we are classifying whether a patient will develop a cardiovascular disease. So here the 2 cluster values are “yes” and “no”. The plot for the K value against the distance is shown below.

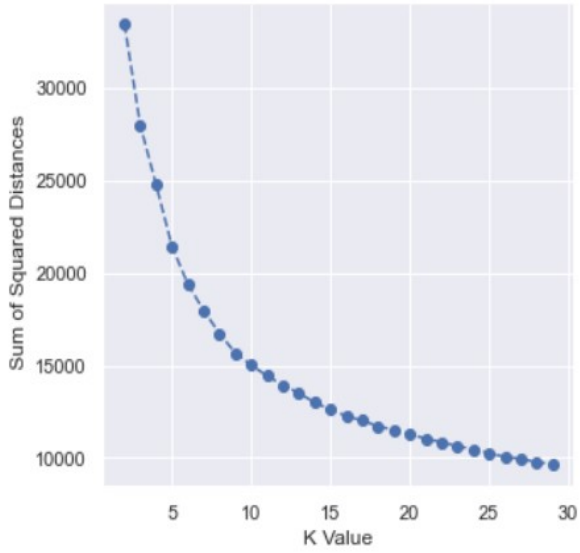

Fig. 9. Elbow Graph to Calculate k value

Just like the supervised learning process, we are going to follow the same procedure in fitting, training, and testing the dataset on the model. By following the process, the following confusion matrix and accuracy rates are resultant as output.

	precision	recall	f1-score	support
0.0	0.89	0.72	0.80	720
1.0	0.25	0.52	0.33	128
accuracy			0.69	848
macro avg	0.57	0.62	0.57	848
weighted avg	0.80	0.69	0.73	848

Fig. 10. Confusion Metrics of K-Means

The accuracy of this unsupervised learning model is 68.89999999999999%. Unsupervised learning can be a little less in performance results as the model is learning by itself without receiving any correction. Due to this reason, the model can be a little inferior to the other models we are comparing with.

```
In [58]: acc = accuracy_score(y_test, base_pred)
         print(f"The accuracy score for K-Means is: {round(acc,3)*100}%")

         The accuracy score for K-Means is: 68.89999999999999%
```

Fig. 11. Accuracy score of K-Means

RESULTSANDDISCUSSION

By comparing the four models based on the accuracy which is how well the model can predict correctly, we get the following results.

ALGORITHM	ACCURACY
KNN	84.6
DECISION TREE	73.2
K-MEANS	68.8
RANDOM FOREST	84.1

Fig. 12. Comparing Accuracy Scores

Observing the above table, we can see that the KNN model is best suited for our classification problem. Taking this into consideration, we can finally apply our questionnaire to the KNN model. The final result will look as shown below.

```
Input Patient Information:
Patient's age: >>> 35
Patient's gender. male=1, female=0: >>> 1
Patient's smoked cigarettes per day: >>> 1
Patient's systolic blood pressure: >>> 77
Patient's diastolic blood pressure: >>> 86
Patient's cholesterin level: >>> 78
Was Patient hypertensive? Yes=1, No=0 >>> 0
Did Patient have diabetes? Yes=1, No=0 >>> 0
What is the Patient's glucose level? >>> 76
Has Patient been on Blood Pressure Medication? Yes=1, No=0 >>> 0


Result:
The patient will not develop a Heart Disease.
```

Fig 13 Final Prediction Result

Conclusion

Cardiovascular diseases are not something that we can ignore and it is important to take the required precautions to stay healthy and safe. This research identified the best machine learning algorithm to apply to our use case and reached an accuracy rate of 84.5%. In the future, this intends to add more complex algorithms and ensemble models into the picture to improve the overall performance. Also, a mobile app development can help people access this procedure at any time and anywhere which can have additions such as sending alerts if the smoking limit has been reached and contacting emergency contacts if any risks have been identified.

REFERENCES

1. Mohan S., C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
2. Patel, Jaymin, Dr TejalUpadhyay, and Samir Patel. "Heart disease prediction using machine learning and data mining technique." Heart Disease 7.1 (2015): 129-137.
3. Katarya, R., Meena, S.K. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. *Health Technol.* 11, 87–97 (2021).
4. Nashif, S., Raihan, Md.R., Islam, Md.R. and Imam, M.H. (2018) Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System. *World Journal of Engineering and Technology,* 6, 854-873.
5. Juan-Jose Beunza, Enrique Puertas, Ester García-Ovejero, GemaVillalba, Emilia Condes, Gergana Koleva, Cristian Hurtado, Manuel F. Landecho, Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease).
6. Jagtap, Abhijeet, et al. "Heart disease prediction using machine learning." International Journal of Research in Engineering, *Science and Management* 2.2 (2019): 352-355.
7. Ramalingam, V. V., AyantanDandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." *International Journal of Engineering & Technology* 7.2.8 (2018): 684-687.

****