Heart Disease Prediction Using Machine Learning Algorithms

1st Sri Jyothirmaye Chowdary.T
Department Of Information Technology
Vardhaman College Of Engineering
Hyderabad, India
srijyothirmayechowdary17it@vardhaman.org

2nd Rakesh.S
Department Of Information Technology
Vardhaman College Of Engineering
Hyderabad, India
rakeshsuda17it@vardhaman.org

3rd Hanish Rao.S
Department Of Information Technology
Vardhaman College Of Engineering
Hyderabad, India
hanishrao17it@vardhaman.org

4th Ravi Kumar.E
Department Of Information Technology
Vardhaman College Of Engineering
Hyderabad, India
ravikumar.e@gmail.com

***Abstract*—Coronary illness is a significant reason for death around the world. The medical services field has a large amount of data, for processing those data certain techniques are used. ,data mining is one of the procedures used. This framework predicts the emerging prospects for coronary heart disease, The results of this framework provide an opportunity for heart disease as YES / NO. This framework examines those boundaries using data mining techniques. Datasets are managed by a python system using the basic Machine Learning Algorithm, Support Vector Classifier, K Neighbors Algorithm, Random Forest Algorithm showing excellent calculation between the above strategies to the appropriate level of coronary disease.**

***Index Terms*—Data Mining, Python Programming, Machine Learning Algorithms.**

I. Introduction

The extraction of information and knowledge from huge amount of data is known as Data Mining. Data mining is an essential step in discovering knowledge from databases. There are numbers of databases, data marts, data warehouses all over the world. Data Mining is used to extract concealed information from a huge database. Data mining is also called as Knowledge Discovery Database (KDD). The data mining has four primary procedures namely Classification, Clustering, Regression, and Association rule. Data mining strategies can quickly mine immense measure of information [1]. Data mining is predominantly required in numerous fields to extricate valuable data from a lot of information [2]. The fields like the clinical field, business field, and instructive field have a tremendous measure of information, consequently these fields information can be mined through those procedures with more helpful data. In proposed framework, We created project, in which the framework predicts the potential outcomes of event of coronary illness as percentage's, It additionally figures out which calculation is best for the forecast of heart disease.It is truly appropriate for coronary illness expectation continuously.

II. Heart Disease

A heart is an imperative organ of the human body. In the event that a heart doesn't play out its activity appropriately, it will impact the other organ of the human-like kidney, cerebrum, etc. According to the statistical data from WHO, one-third population worldwide died from heart disease; heart disease is found to be the leading cause of death in developing countries by 2017. The heart pumps blood through the blood vessels of the circulatory system. Blood furnishes the body with oxygen and supplements, just as aiding the evacuation of metabolic squanders. In the event that blood in the body is inadequate, numerous organs like frontal cortex endure and if heart stops working by, death occurs within minutes.

Heart disease risk factor include:
- High Cholesterol
- High blood pressure
- Diabetics
- Smoking
- Consuming too much alcohol
- Being overweight or obese
- Family history of coronary illness

Symptoms of Heart attack:
- Shortness of breath
- Pain and discomfort in chest
- Pain may spread to left or right arm or to neck, jaw, back or stomach
- Fatigue
- Cold sweat and unsteadiness
- Rapid or irregular heart beat
- Heart burn or abnormal pain

Types of cardiovascular disease:
- Coronary artery disease
- Cardiac arrest

• Congestive heart failure
• Stroke, and more.

III. Literature Survey

A research which was conducted in 2000 by Shusaku Tsumoto, says that human beings were not able of arranging huge size data, so it was suggested to use data mining techniques which are available for finding various kinds of patterns which is available from huge database and then can again be used for clinical research purposes and to perform various operations on it [3].

Tan et al. [4] proposed a hybrid method in which two machine learning algorithms, Support Vector Machine (SVM) and Genetic Algorithm (G.A), are effectively combined with the wrapper approach. They used LIBSVM and the WEKA mining tool to analyze the results. Five data sets (Iris, diabetes disease, breast cancer disease, heart disease and hepatitis) are collected from the Irvine UC machine learning repository for this experiment. After applying the hybrid GA and SVM approach, an accuracy of 84.07% is obtained for heart disease. For all diabetes data, 78.26% accuracy is achieved. The 86.12% accuracy is the result of hepatitis disease.

Otoom et al. [5] presented a system for analysis and follow-up. Coronary artery disease is detected and monitored by the proposed system. Cleveland Heart data are taken from the UCI. This dataset consists of 303 cases and 76 attributes/features. 13 features are used out of 76 features. Two tests with three algorithms: Bayes Naive, Support vector machine, and Functional Trees FT are performed for detection purposes. The WEKA tool is used for detection.

Chaurasia et al. [6] suggested using data mining approaches to detect heart disease. The WEKA data mining tool is used which contains a set of machine learning algorithms for mining purposes. A few tests were led on the clinical informational indexes utilizing different characterizes and highlight determination methods. There is an examination on coronary illness information set.Many of those contain great precision.

Medical services has came long way for treatment of patients with various types of diseases.
Finding of patients accurately and managing viable medicines have become a significant test.
Determination of the condition exclusively relies on In the current system,practical utilization of different gathered information is tedious.

IV. Proposed System

In proposed framework, We created project, in which the framework predicts the potential outcomes of coronary illness as percentage's, It additionally figures out which algorithm is best for the forecast of heart disease [7]. The primary errand of the information expectation is finished utilizing four strategies. Main Method used for the prediction is Decision tree,KNN algorithm,SVM and Random forest. The systems use 13 medical attributes or parameters as inputs, the datasets go through mining and displays the accuracy of each algorithm [8]. By implementation of computerized system, limitations in the proposed system will be reduced. These models are executed using python Programming Language.

V. METHODOLOGY

A. Data Source

The data used to anticipate Heart disease was taken from the UCI ML database. The database is a real dataset and contains about 300 instances of information covering 14 clinical boundaries. Database parameters are related to tests that are considered to indicate coronary heart disease such as bp level, type of chest pain and electrocardiography effect and so on.

This data set comprises of 76 attributes,all the distributed investigations allude to the utilization of a subset of 14 of them. In Particular, Cleveland data set is the solitary information base that was utilized by the ML analysts to date. In this project, apart from Cleveland, we also used other database like Switzerland and Budapest.

The "target" issue shows the presence of coronary coronary heart disorder in a patient,Its either int value 0 (Not present) or 1 (presence).

B. Architecture Diagram

Fig. 1. Architecture Diagram

C. Data Preprocessing

```
df = pd.get_dummies(df, columns = ['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'])

standardScaler = StandardScaler()
columns_to_scale = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']
df[columns_to_scale] = standardScaler.fit_transform(df[columns_to_scale])
```

Fig. 2. Data Preprocessing

D. Algorithms Used

1) Decision Tree:
It is one of the supervised ML algorithms. It is used for handling both categorical and numerical data. Based on the conditions, it gives solution as categorical Yes/No,True or False, 1 or 0. For handling the medical dataset, this algorithm is widely used. The output depends on the conditions and dependent variables and it is in the form of horizontal and

vertical line splits. This model analyzes the data on the basis of three nodes namely:

1) Root node:Based on this node, all others perform their functions
2) Interior node: this node handles the conditions of dependent variables.
3) Leaf node: final result is carried on to a leaf node.

To find root node:

$$InformationGain = ClassEntropy - EntropyAttribute \quad (1)$$

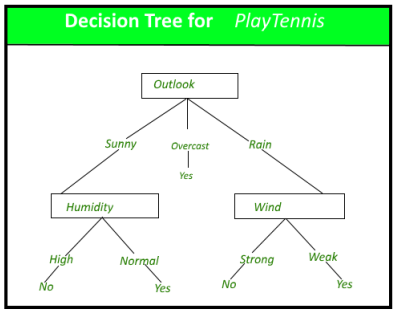

Fig. 3. Decision Tree

Decision Tree

```
[ ] dt_scores = []
    for i in range(1, len(X.columns) + 1):
        dt_classifier = DecisionTreeClassifier(max_features = i, random_state = 0)
        dt_classifier.fit(X_train, y_train)
        dt_scores.append(dt_classifier.score(X_test, y_test))

[ ] plt.plot([i for i in range(1, len(X.columns) + 1)], dt_scores, color = 'green')
    for i in range(1, len(X.columns) + 1):
        plt.text(i, dt_scores[i-1], (i, dt_scores[i-1]))
    plt.xticks([i for i in range(1, len(X.columns) + 1)])
    plt.xlabel('Max features')
    plt.ylabel('Scores')
    plt.title('Decision Tree Classifier scores for different number of maximum features')
```

Fig. 4. Decision Tree

2)K Nearest Neighbours:

• It is one of the Supervised ML algorithms that usesthat uses label inputted dataset to predict the output of data points.
• It is one of the most simple ML algorithms, it can be easily implemented for different set of problems.
• It is based on feature similarity. KNN checks the similarity between a data point and its neighbor and classifies the data point into the class it is most similar to.
Example:

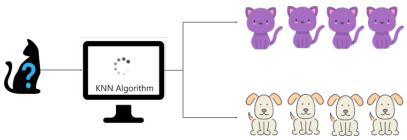

Fig. 5. KNN example

K neighbours

```
[ ] knn_scores = []
    for k in range(1,31):
        knn_classifier = KNeighborsClassifier(n_neighbors = k)
        knn_classifier.fit(X_train, y_train)
        knn_scores.append(knn_classifier.score(X_test, y_test))

    plt.plot([k for k in range(1, 31)], knn_scores, color = 'red')
    for i in range(1,31):
        plt.text(i, knn_scores[i-1], (i, knn_scores[i-1]))
    plt.xticks([i for i in range(1, 31)])
    plt.xlabel('Number of Neighbors (K)')
    plt.ylabel('Scores')
    plt.title('K Neighbors Classifier scores for different K values')
```

Fig. 6. KNN

3)Support Vector Machine:

It is one of the Supervised ML algorithms, which is used for Classification as well as Regression problems.The objective of the SVM algorithm is to create best line or the decision boundary that can isolate n-dimensional spaces into classes in order to put the new data point in the correct category in the future. This best decision boundary is called as a hyperplane.

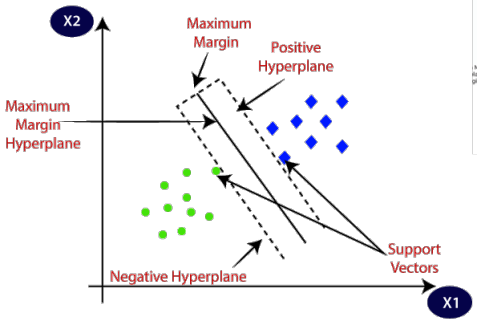

Fig. 7. SVM Example

Support Vector Classifier

```
[ ] kernels = ['linear', 'poly', 'rbf', 'sigmoid']

    svc_scores1 = []
    for i in range(len(kernels)):
        svc_classifier = SVC(kernel = kernels[i])
        svc_classifier.fit(X_train1, y_train1)
        svc_scores1.append(svc_classifier.score(X_test1, y_test1))

    svc_scores2 = []
    for i in range(len(kernels)):
        svc_classifier = SVC(kernel = kernels[i])
        svc_classifier.fit(X_train2, y_train2)
        svc_scores2.append(svc_classifier.score(X_test2, y_test2))

    svc_scores3 = []
    for i in range(len(kernels)):
        svc_classifier = SVC(kernel = kernels[i])
        svc_classifier.fit(X_train3, y_train3)
        svc_scores3.append(svc_classifier.score(X_test3, y_test3))
```

Fig. 8. SVM

4)Random Forest:

It is one of the Supervised ML algorithms,which firstly creates decision trees on data-samples and then gets the prediction from each of those and then selects the best solution by voting. It is known as an ensemble method, which is better than a single-decision-tree as it reduces the over fitting by averaging the results.

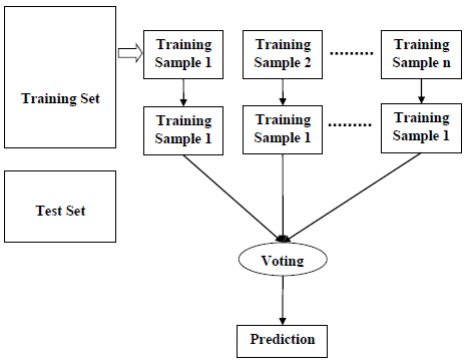

Fig. 9. Random Forest

Random Forest

```
[ ] rf_scores = []
    estimators = [10, 100, 200, 500, 1000]
    for i in estimators:
        rf_classifier = RandomForestClassifier(n_estimators = i, random_state = 0)
        rf_classifier.fit(X_train, y_train)
        rf_scores.append(rf_classifier.score(X_test, y_test))

[ ] colors = rainbow(np.linspace(0, 1, len(estimators)))
    plt.bar([i for i in range(len(estimators))], rf_scores, color = colors, width = 0.8)
    for i in range(len(estimators)):
        plt.text(i, rf_scores[i], rf_scores[i])
    plt.xticks(ticks = [i for i in range(len(estimators))], labels = [str(estimator) for estimator in estimators])
    plt.xlabel('Number of estimators')
    plt.ylabel('Scores')
    plt.title('Random Forest Classifier scores for different number of estimators')
```

Fig. 10. Random Forest

All the above algorithms [9] will be applied to the dataset to find the best algorithm for prediction heart disease.

VI. RESULT DISCUSSION

We have applied all the four algorithms on three different datasets namely Cleveland, Switzerland, Budapest. The following are the accuracy obtained for each algorithm.
As we can see from the below plot that KNN has the highest accuracy, So KNN algorithm is used to build the GUI to predict the occurrence of heart disease.

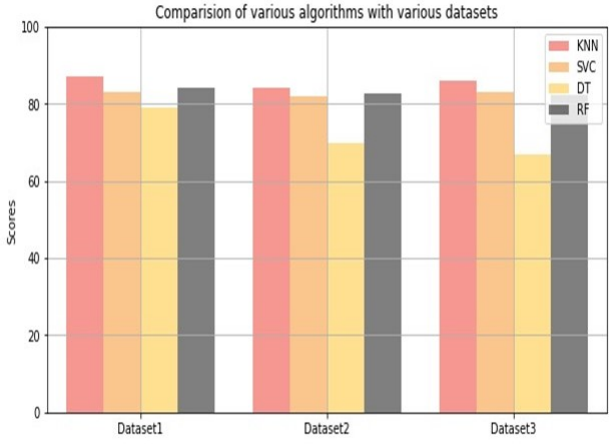

Fig. 11. Result

In this, KNN Algorithm has been used to build the GUI using python programming language and tkinter for the prediction of heart disease.

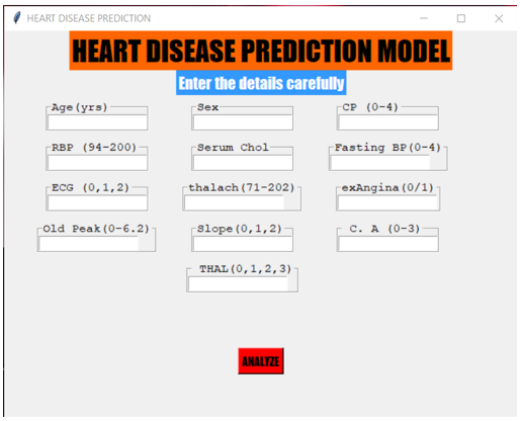

Fig. 12. GUI

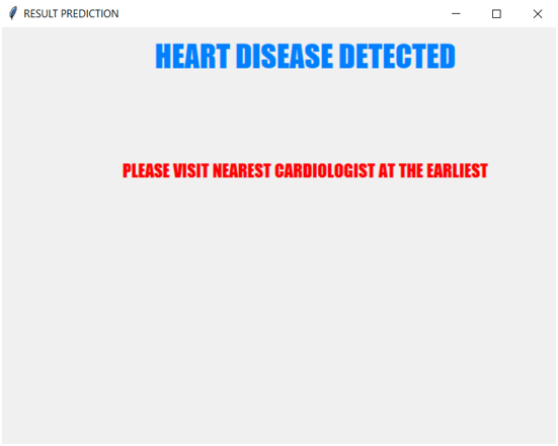

Fig. 13. Final Output

VII. CONCLUSION

In this paper, four supervised data mining algorithms are applied on various datasets to predict the possibilities of occurrence of heart disease. These algorithms are applied to the same datasets in order to analyze the best algorithm in terms of accuracy. The decision tree model predicted the heart disease with 79% where as KNN has predicted with 87% , SVM predicted with 83% and Random forest with 84%. So, we can conclude that KNN has the highest accuracy.
We used KNN algorithm to build a GUI to predict heart disease.
In the future, the designed system with the classified ML algorithms can be used to predict any other diseases. The work can also be improved or extended for the automation of heart disease analysis including any other or some other ML Algorithms.

REFERENCES

[1] Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., Lamgunde, A. (2017, June). Heart disease prediction using data mining techniques. In 2017 International Conference on Intelligent Computing and Control (I2C2) (pp. 1-8). IEEE

[2] Using Data Mining Techniques to Predict Diabetes and Heart Diseases. In 2018 4th International Conference on Frontiers of Signal Processing (ICFSP) (pp. 150- 154). IEEE.

[3] Tsumoto, S. (2000, October). Discovery of clinical knowledge in hospital information systems: Two case studies. In International Symposium on Methodologies for Intelligent Systems (pp. 573-581). Springer, Berlin, Heidelberg.

[4] L. Van Cauwenberge, “Top 10 Machine Learning Algorithms”, Data Sci. Cent., 2015.

[5] S. Thirumuruganathan, “A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm”, [Online]. https//saravananthirumuruganathan.wordpress.c om/2010/05/17/a-detailed- Introd., 2010.

[6] Introduction to Machine Learning Using Python [Online]. Available:

https://www.geeksforgeeks.org/introduction-machine-learning-using-python/

[7] Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., Lin, E. J. (2011,September). HDPS: Heart disease prediction system. In 2011 Computingin Cardiology (pp. 557-560). IEEE.

[8] Sultana, Marjia, Afrin Haider, and Mohammad Shorif Uddin. ”Analysisof data mining techniques for heart disease prediction.” In 2016 3rdInternational Conference on Electrical Engineering and Information-Communication Technology (ICEEICT), pp. 1-5. IEEE, 2016.

[9] Introduction to Supervised Machine Learning Algorithms [Online]. Available:

https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/