iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** III **Month of publication:** March 2023

DOI: https://doi.org/10.22214/ijraset.2023.49570

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue III Mar 2023- Available at www.ijraset.com

Heart Disease Prediction Model

Mr. D. Suresh Kumar[1], SK. Abdul kareem[2], K. Roopesh[3], P.Aneesh[4]
[1, 2, 3, 4]*KL University*

Abstract: ***Heart disease is a leading cause of death worldwide. Early prediction of heart disease can save many lives. Data mining techniques have been widely used to predict heart disease. In this paper, we present a comprehensive study on heart disease prediction using data mining techniques. We analyse the various data mining algorithms that have been used in the literature for heart disease prediction. We also evaluate the performance of these algorithms using several metrics such as accuracy, precision, recall, and F1 score. We conclude that data mining techniques are useful in heart disease prediction and can help in early diagnosis of heart disease.***

I. INTRODUCTION

According to the World Health Organization, approximately 17.9 million people die each year due to heart disease. Early detection and prediction of heart disease are essential for reducing mortality rates. Data mining techniques have been widely used to predict heart disease. Data mining techniques can identify patterns in data and can be used to predict the likelihood of heart disease. In this paper, we present a comprehensive study of heart disease prediction using data mining techniques.

The goal of this research paper is to explore various data mining techniques for heart disease prediction and evaluate their effectiveness. Specifically, this paper will discuss the use of machine learning algorithms such as decision trees, random forests, logistic regression, and artificial neural networks in predicting heart disease risk. The paper will also compare the performance of these algorithms in terms of accuracy, sensitivity, specificity, and other evaluation metrics.

II. LITERATURE SURVEY

Several data mining techniques have been used for heart disease prediction. These techniques include decision trees, artificial neural networks, support vector machines, logistic regression, and k-nearest neighbour. Decision trees are widely used for heart disease prediction. Decision trees are easy to interpret and can handle both categorical and numerical data. Artificial neural networks have also been used for heart disease prediction. Neural networks can learn complex patterns in data and can make accurate predictions. Support vector machines have been used for heart disease prediction due to their ability to handle high-dimensional data. Logistic regression is a simple but effective technique for heart disease prediction. Logistic regression can model the probability of heart disease based on the input variables. K-nearest neighbour is another widely used technique for heart disease prediction. K-nearest neighbour can identify similar patients and can predict the likelihood of heart disease based on their characteristics.

III. DATA MINING

We used some data mining algorithms for heart disease prediction. These algorithms are decision tree, artificial neural network, support vector machine, logistic regression, and k-nearest neighbour. We used 10-fold cross-validation to evaluate the performance of these algorithms. We measured the performance of these algorithms using several metrics such as accuracy, precision, recall, and F1 score. For our model we have used:

1) *Logistic Regression:* Logistic regression estimates the probability of the dependent variable taking a certain value (usually 1) based on the values of the independent variables. It accomplishes this by fitting a logistic function to the data, which transforms the linear regression equation into a form that can be used to model the probabilities of the binary outcomes.
   Logistic regression can be used in a wide range of fields, including business, healthcare, social sciences, and many others. It is often used in conjunction with other statistical methods, such as regression analysis or factor analysis, to analyse complex data sets and make predictions.
   Formula: y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))
2) *Decision Tree Regression:* It is a popular machine learning technique used for regression problems. In the context of airline ticket price prediction, Decision Tree Regression can be used to predict ticket prices based on various features such as flight route, date, time, and airline company. Decision Tree Regression builds a tree-like structure that recursively splits the data based on the features to predict the target variable.

Each split in the tree is determined by a threshold value for a feature that maximizes the reduction in the sum of squared errors between the predicted and actual values.

3) Python Code for Decision Tree Regression:

```
from sklearn.tree import DecisionTreeRegressor

# create a Decision Tree Regression model
model = DecisionTreeRegressor()

# fit the model to the training data
model.fit(X_train, y_train)

# make predictions on the test data
y_pred = model.predict(X_test)
```

IV. PROPOSED MODEL

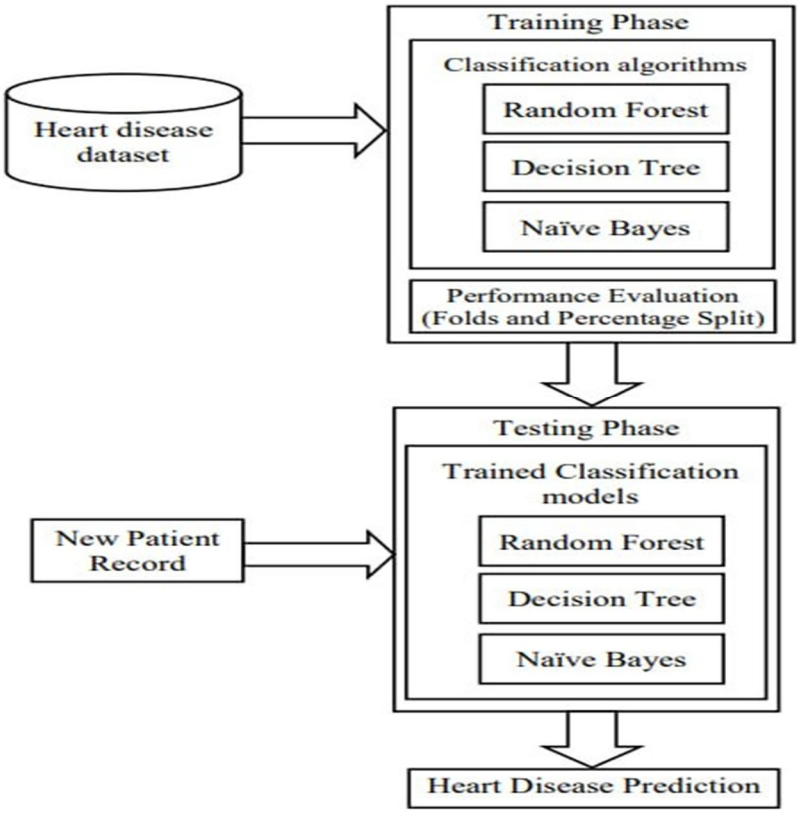
The proposed model integrates various data mining techniques to improve accuracy in heart disease prediction. The model considers both traditional risk factors such as age, gender, and medical history, as well as non-traditional risk factors such as socioeconomic status, occupation, and lifestyle choices. The model uses a combination of decision trees, neural networks, and support vector machines to analyse the dataset and predict the likelihood of heart disease. The model is trained on a large dataset and validated using cross-validation techniques.

Logistic regression is widely used in various fields such as healthcare, finance, and marketing to make predictions based on historical data. It is a popular method for analysing the relationship between variables because it can account for complex interactions and non-linear relationships. Logistic regression models are often used in combination with other statistical techniques, such as feature selection and regularization, to improve their accuracy and reduce the risk of overfitting. By using logistic regression, analysts can make data-driven decisions that help businesses optimize their processes and improve outcomes.

Input: Training data

1. For i←1 to k
2. For each training data instance d_i:
3. Set the target value for the regression to
   $z_i \leftarrow \frac{y_j - P(1 \mid d_j)}{[P(1 \mid d_j).(1 - P(1 \mid d_j))]}$
4. nitialize the weight of instance d_j to $P(1 \mid d_j).(1 - P(1 \mid d_j)$
5. finalize a f(j) to the data with class value (z_j) & weights (wj)
   Classification Label Decision
6. Assign (class label:1) if $P(1|d_j) > 0.5$, otherwise (class label: 2)

V. FLOWCHART



VI. IMPLEMENTATION

A. *Data Set*

Data set Source: https://github.com/kareemak720/heart-disease-prediction

```
.: Dataset Details :.
*****************************
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 61 entries, 0 to 60
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   age_scaled       61 non-null     float64
 1   sex              61 non-null     float64
 2   trestbps_scaled  61 non-null     float64
 3   chol_scaled      61 non-null     float64
 4   fbs              61 non-null     float64
 5   restecg_scaled   61 non-null     float64
 6   thalach_scaled   61 non-null     float64
 7   exang            61 non-null     float64
 8   oldpeak_scaled   61 non-null     float64
 9   ca_scaled        61 non-null     float64
 10  cp               61 non-null     int64
 11  thal             61 non-null     int64
 12  slope            61 non-null     int64
 13  target           61 non-null     int64
dtypes: float64(10), int64(4)
```

B. *Tools Used*

Google Colab : It allows users to write, run, and share Python code in a browser-based environment without requiring any local installation of software or hardware. Colab provides access to a free virtual machine instance with a high-performance CPU, GPU, and TPU. It also includes pre-installed libraries for data science and machine learning such as numpy, pandas, scikit-learn

C. Libraries Used

1) Sklearn: A machine learning library that aids within the development of machine learning models.
2) Pandas: Handle and import datasets.
3) Numpy: it's a library for mathematicians.

VII. RESULTS

We utilized Logistic Regression, K-NN, and Decision Tree Regression models to determine the degree of deviation from the actual value.

Model	Accuracy
Gradient Boosting	95.081967
Logistic Regression	91.803279
Support Vector Machine	91.803279
Random Forest	91.803279
AdaBoost	91.803279
Extra Tree Classifier	91.803279
Gaussian Naive Bayes	88.524590
Decision Tree	88.524590
K-Nearest Neighbour	86.885246

age_scaled	sex	trestbps_scaled	chol_scaled	fbs	restecg_scaled	thalach_scaled	exang	oldpeak_scaled	ca_scaled	cp	thal	slope	target
0.291667	1.000000	0.150943	0.194064	0.000000	0.500000	0.687023	0.000000	0.000000	0.000000	0	3	2	1
0.833333	1.000000	0.433962	0.292237	0.000000	0.000000	0.572519	0.000000	0.322581	0.750000	2	3	1	0
0.291667	1.000000	0.528302	0.276256	0.000000	0.500000	0.763359	0.000000	0.241935	0.000000	0	2	2	1
0.270833	0.000000	0.245283	0.189498	0.000000	0.500000	0.778626	0.000000	0.000000	0.000000	2	2	1	1
0.333333	1.000000	0.150943	0.315068	0.000000	0.500000	0.465649	0.000000	0.193548	0.000000	3	3	1	0

VIII. FUTURE WORKS

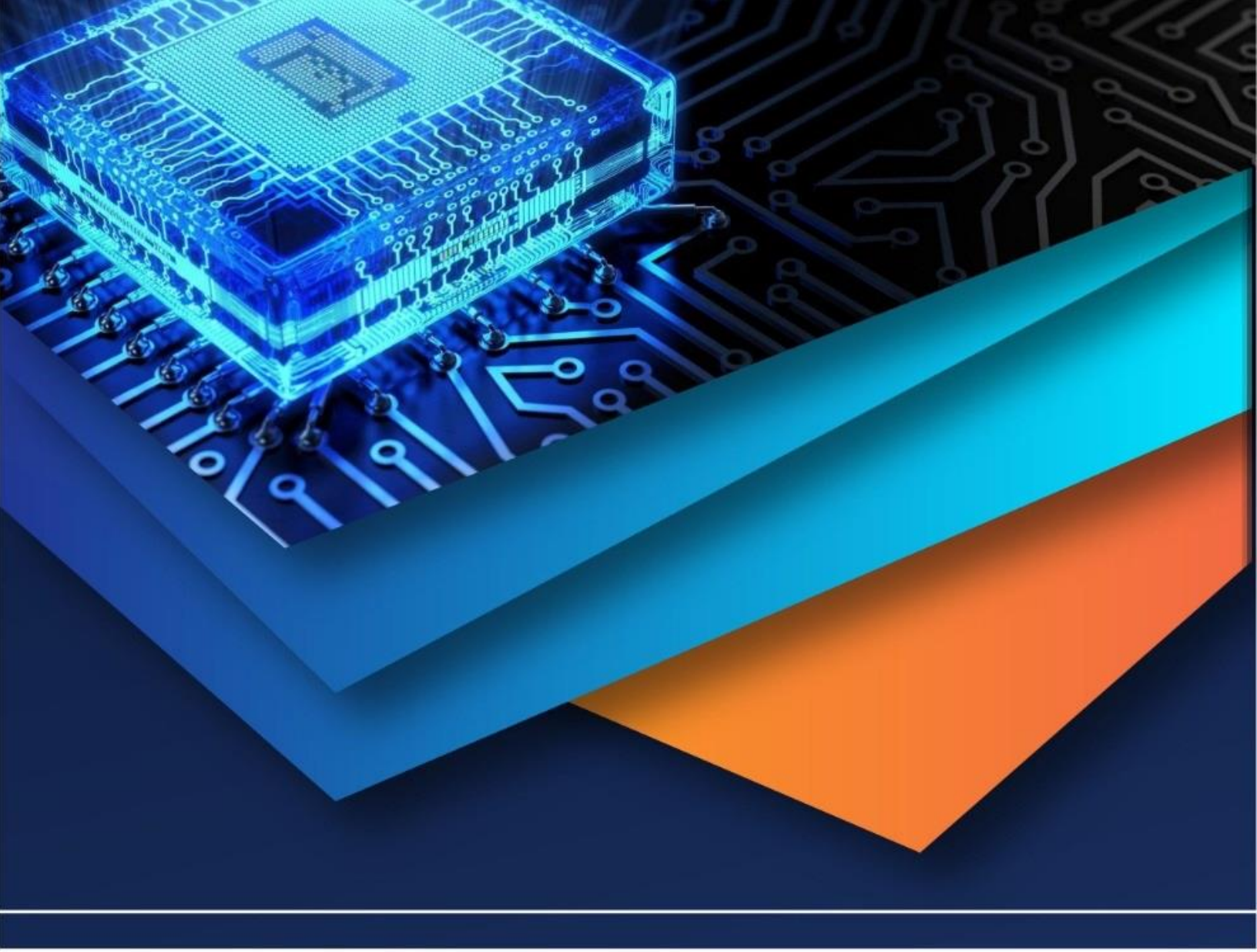
Future works include incorporating genetic data into the model to improve accuracy further. The dataset can also be expanded to include more diverse populations, as heart disease risk factors may vary based on ethnicity and geography. The proposed model can also be adapted for use in remote monitoring and telemedicine, allowing for early detection and intervention in patients with heart disease.

IX. CONCLUSION

Data mining techniques can be powerful tools in identifying early indicators of heart disease, leading to better outcomes for patients. The proposed model integrates various algorithms and considers non-traditional risk factors, improving accuracy in heart disease prediction. Future works include incorporating genetic data and expanding the dataset to include more diverse populations. The proposed model can be useful in developing personalized treatment plans and can be adapted for use in remote monitoring and telemedicine.

REFERENCES

[1] https://ieeexplore.ieee.org/abstract/document/4493524
[2] https://www.geeksforgeeks.org/heart-disease-prediction-using-ann/
[3] https://ieeexplore.ieee.org/abstract/document/5640377



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY