JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Prediction of Cardiovascular diseases using Logistic regression: Machine Learning classification algorithm

[1]Sravan Kumar Gurram, [2]Dr.Arathi Chitla, [3]Harshith Raj Boddu
[1]Assistant Professor(C), [2]Professor, [3]Big Data Python Developer
[1]Department of CSE, [2]Department of CSE, [3]SAS2PY Pvt.Ltd.
[1]JNTUH University College of Engineering Jagtial, Telangana State, India

Abstract : Heart disease is one of the complex diseases and globally many people are suffering from this disease. Early detection of heart disease can aid people in making lifestyle changes and, if necessary, ensuring effective medical care. The correct prediction of heart disease can prevent life threats. Machine Learning is one of the prominent techniques used in healthcare domain. Many of the machine learning techniques were even working on identifying heart diseases. The logistic regression algorithm is employed as a predictive modeling tool to analyze and interpret the complex relationships within this dataset, ultimately yielding a robust and interpretable model for CVD risk assessment. To achieve the accuracy, we used Logistic Regression with UCI Cleveland data set with different splits.

IndexTerms - Cardiovascular diseases, UCI Cleveland data set, logistic regression, machine learning.

I. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death globally. An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke. Over three quarters of CVD deaths take place in low- and middle-income countries. It is important to detect cardiovascular disease as early as possible so that management with counselling and medicines can begin at the earliest [1]. One of the major challenges of the medical industry is to predict the cardio vascular disease with less expensive and more reliable method to avoid the compounding effect of the disease in low income or developing countries. The early detection not only reduce the cost but also improves the quality of life.

Through applying the technology of data mining, a new idea is provided for the prediction of heart disease, extracting clinical attributes and pathological data from large medical data sets, and generating biological hypotheses. At present, some studies have applied machine learning technology to the prediction of heart disease, but there are limited studies on the important features of cardiovascular disease. Whereas logistic regression can extract the risk factors of disease and predict the incidence probability of patients in real time. The main objective of this research is to develop a heart disease prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set. Heart disease prediction system aims to exploit machine learning techniques on medical data set to assist in the prediction of the heart diseases. Machine learning is an important area of artificial intelligence. The objective of machine learning is to discover knowledge and make intelligent decisions [2]. Machine learning technology provides an immeasurable platform to the medical sector for resolving health problems effectively. It allows the construction of models to quickly analyze data and deliver results quickly. Using machine learning models, doctors can make a good decision of patient diagnosis and treatment options, leading to improvement of patient health care services [3].

We use UCI Cleveland heart disease dataset. This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them with 303 observations. Specifically, the Cleveland database is the only one that ML researchers have utilized up to this point. In these 14 more relevant features, 13 are independent features and one is target variable which is used to predict the presence of heart disease by using 0 for affected 1 for not affected [4]. Exploratory Data Analysis (EDA) is a method to analyze data using advanced techniques to expose hidden structure, enhances the insight into a given dataset, and identifies the anomalies and outliers. With help of EDA we can remove the noise which helps to improve the accuracy of the model [5]. After EDA we develop the machine learning model by using logistic regression which is classification algorithm. For developing this model, we use sklearn library, working with data frames we use pandas, for data visualization we use matplotlib and seaborn. IDE we use for this work is Jupyter notebook, programming Language we use to develop the code is Python.

In this paper, we have tried to study all the risk factors that influence on the heart and can cause heart disease. In the II section, we have briefed the literature survey in this area. In section III, we have explained the methodologies and all the techniques which we have applied for the prediction of the heart disease. We have shown the results of our work in section IV. We concluded our paper with our findings in section V followed by future work in section VI.

II. Literature Survey

There are many works in literature which diagnoses heart diseases using machine learning as well as deep learning. K. Karthick et al. used UCI data set and split it into the ratio 80:20, they used multiple machine learning techniques finally they concluded random forest gives best accuracy with their experiment [6].

Ashir Javeed et al. used Random Forest model technique, Cleveland heart failure dataset that is freely available on machine learning UCI repository. They used accuracy, sensitivity, specificity and MCC are the evaluation metrics. Finally, they obtained 88.4%. for coronary artery disease (CAD) diagnosis [7].

sarria e. et al. used cardiovascular patients' dataset from the UCI Machine Learning Repository, used the ensemble model with a majority voting technique to construct hybrid classifiers. Their results indicated that the proposed ensemble classifier model achieved a classification accuracy of 98.18% [8].

Mohan et al. used Cleveland data set. The first step is data pre-processing step. In this the tuples are removed from the data set which has missed the values. Attributes age and sex from data set are also not used as the authors thought that it is personal information and has no impact on predication. The remaining 11 attributes are considered important as they contain vital clinical records. They have proposed their own Hybrid Random Forest Linear Method (HRFLM) which is the combination of Random Forest (RF) and linear method (LM). In the HRFLM algorithm, the authors have used four algorithms. RF and LM are giving better results than other algorithms, both the algorithms are put together and new unique algorithm HRFLM is created. They suggested further improvement in accuracy by using combination of various machine learning algorithms [9].

III. Methodology

We used existing data set on that we apply logistic regression and experimented with different splits.

3.1 Data Collection

The UCI data repository provided the data set used in this study. Access to UCI Machine Learning Repository is unrestricted. Many studies have discovered that the Cleveland and Hungarian databases, in particular, are good.

For creating a machine learning model due to their lower levels of missing information and outliers. Before the data is sent into the proposed algorithm for training and testing, it is cleaned and preprocessed. The machine learning community uses the UCI Machine Learning Repository, which is a collection of databases, domain theories, and data generators, for the empirical investigation of machine learning algorithms. Our work's main goal is to increase the accuracy of heart disease detection. To obtain more precise results, the UCI repository dataset is used in this work. A subset of 14 are used in all published experiments. Specifically, the Cleveland database is the only one that ML researchers have utilized up to this point. The patient's heart condition is indicated in the "Target" field. It has an integer value between 0 and 1 (0 denoting absence and 1 presence). Since the majority of machine learning algorithms demand integer values, attributes having category values were changed to numerical values. For variables with more than two categories, dummy variables were also made.

There are 14 attributes and 303 records in the collection. The output value, or the forecast value of the patients with heart disease, is one of the thirteen factors that were utilized as the eigen values for the forecast of heart disease.

3.2 Data Analysis

In this work, we used Jupyter notebook and python programming language to develop the code. In data processing we removed missing values and outliers. We need to check the presence of null values.

```
# checking for sum of missing values
df.isnull().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

We checked the distribution to the target variable using countplot method of seaborn library.

sns.countplot(x="target", data=df)

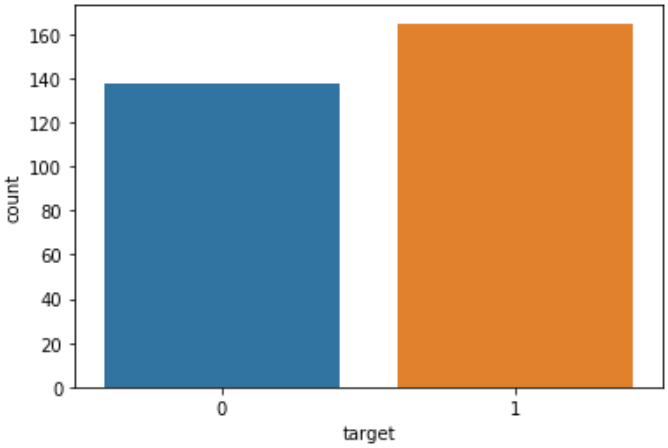

Fig.1: Distribution of Target Variable

We also checked number of the persons affected by the heart disease based on gender using sex feature in the dataset.
sns.countplot(x="target", hue="sex", data=df)

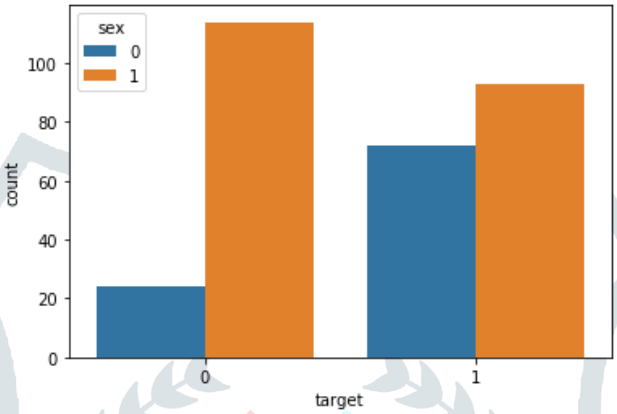

Fig.2: Number of the classes in sex feature vs target feature.

We checked the distribution of each feature using hist method of matplotlib library. df['age'].plot.hist()

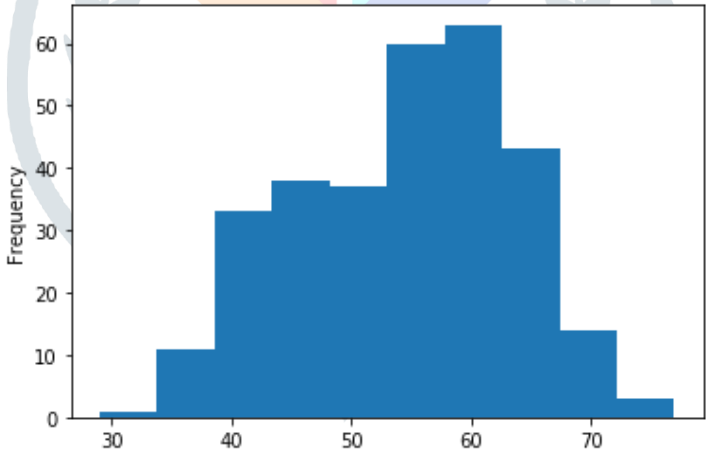

Fig.3: Data distribution of age attribute

IV. Results

After analysis of the data, that processed data given to the machine learning model. To divide the data into training and testing sets, we employ the train_test_split method of sklearn package. The model is trained using training data. With help of this trained knowledge, we test the developed model using the unknown data to model i.e., testing data.

In this work we split the dataset in to multiple ratios and find the results for all the ratios. Finally, we found best split. For each and every split we use confusion matrix which is used to evaluate the model. We use seaborn library to create confusion matrix and with heatmap method we can visualize it.

sns.heatmap(confusion_matrix(Y_test, X_test_prediction), annot=True

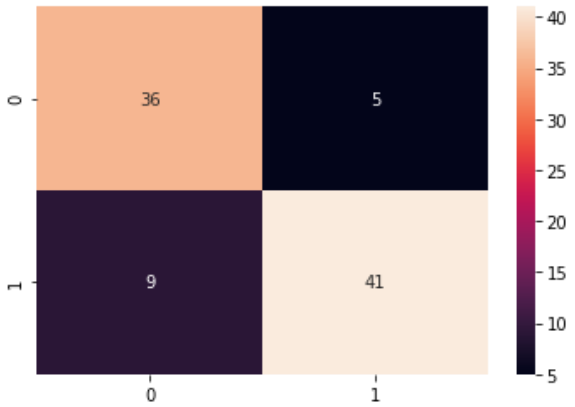

Fig.4: Confusion matrix for 70:30 split.

We use classification_report method of sklearn machine learning library to know the evaluation metrics precision, recall, f1-score values for each split.

Split Ration		Precision	Recall	F1 Score
90:10	0	73	79	76
	1	81	76	79
85:15	0	74	81	77
	1	83	76	79
80:20	0	79	82	81
	1	84	82	83
75:25	0	81	86	83
	1	87	83	85
70:30	0	88	88	84
	1	89	82	85
65:35	0	79	82	81
	1	84	82	83
60:40	0	83	80	82
	1	84	86	85

Table 2. Evaluation metrics values for each split

With help of the confusion matrix we found the accuracy for training data and testing data also. Finally, we found best split as 70:30 with 84.61 accuracy.

S.No.	Split Ration	Training Accuracy	Testing Accuracy
1	90:10	84.92	77.41
2	85:15	84.82	78.26
3	80:20	85.12	81.96
4	75:25	84.58	84.21
5	**70:30**	**84.43**	**84.61**
6	65:35	85.12	81.96
7	60:40	87.84	83.60

Table 2. Accuracy values for each split

V. Conclusion

With this we have concluded applying logistic regression on UCI Cleveland dataset with 14 attributes with different splits got good results in finding the heart diseases. This approach is less time consuming and less complexity and allows us to predict the possibility of occurring heart disease at the early stage itself. Hence people can go for the medication with less cost.

VI. Future Work

The main goal of this research is to develop highly accurate cardiac disease prediction tools. We used the logistic regression algorithm in machine learning implemented with sklearn using python programming language, to predict cardiac disease. The

prediction of cardiac illnesses utilizing cutting-edge methods and algorithms with even lower temporal complexity is the paper's future focus.

REFERENCES

[1] World Health Organization, Cardiovascular diseases, key facts, June 2021, https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] Lidong Wang, Cheryl Ann Alexander, Machine Learning in Big Data, 2016.

[3] Yash D. Patel, Janushi Shastri, Shraddha Tandel, Machine Learning in Healthcare Sector, May 2021.

[4] UCI Machine Learning Repository, Heart Disease Data Set, https://archive.ics.uci.edu/ml/datasets/heart+disease

[5] R. Indra Kumari, T. Poongodi, Soumya Ranjan Jena, Heart Disease Prediction using Exploratory Data Analysis, 2020.

[6] K. Karthick, S. K. Aruna, Ravi Samikannu, Ramya Kuppusamy, Yuvaraja Teekaraman and Amruth Ramesh Thelkar , Implementation of a Heart Disease Risk Prediction Model Using Machine Learning, 2022.

[7] Ashir javeed, shijie zhou, liao yongjian, iqbal qasim, adeeb noor and redhwan nour, An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection, 2019.

[8] Sarria e. a. ashri, m. m. el-gayar, eman m. el-daydamony, Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm, 2021.

[9] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava, “Effective heart disease prediction using hybrid machine learning techniques” IEEE Access (2019).