iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41962

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 10 Issue IV Apr 2022- Available at www.ijraset.com

Machine Learning based Model for Heart Disease Prediction

Gopal Murlidhar Kholade[1], Adharsh Vishnu Tayde[2], Kiran Kishor Vaishnav[3], Dhiraj Shyam Jadhav[4], Dhiraj Sheshrao Jadhao[5]
[1, 2, 3, 4, 5]*Department of Information Technology, Anuradha Engineering College, Chikhli- 443201, India*

Abstract: ***Nowadays we see around us that heart diseases and diabetes are the major threats for human health increasing day-by-day due to either lifestyle or inheritance. Nowadays these two diseases are commonly found under every age group above the 25 years of age and the most fatal one. Most of the people are not able to give some time for their health and that is why people get notified too lately when the diseases are not under control. This model will be helpful in determining the risk of heart disease as well as diabetes prediction.***
Keywords: Machine Learning(ML), Heart Disease Prediction, Diabetes Prediction, Datasets, Random Forest Classifier, web apps

I. INTRODUCTION

Machine Learning is a branch of Artificial intelligence which deals with the automation of work for the ease of humans. Here , we are classifying the task using ML. Machine learning provides one of the main features for extracting the data from large databases. Then by various algorithms it makes a set of various results or we can call it as training output. And after that uses the training output data to identify the main results. Machine learning in medical health care is an emerging field of very high importance for providing prognosis and deeper understanding of data. Most of the ML work depends on Datasets and implementation of algorithms. After that presentation is the next step.

II. SYSTEM REQUIREMENTS

System is the main component for the model. Because, for calculating the result using a model and then for presentation purposes. This all needs to be combined using an application.

A. Software Requirements

1) For Running The Project

- dj-database-url==0.5.0
- Django==3.1.12
- django-heroku==0.3.1
- gunicorn==19.9.0
- joblib==0.14.0
- numpy==1.17.2
- pandas==0.25.1
- psycopg2==2.8.3
- python-dateutil==2.8.0
- pytz==2019.3
- scikit-learn==0.21.3
- scipy==1.3.1
- six==1.12.0
- sqlparse==0.3.0
- whitenoise==4.1.4

2) for Running the Web App

- Browser - Any with web 2.0

B. Hardware Requirements

1) CPU - Intel 9th gen and above
2) Ram - 2GB and above
3) Storage - 100GB and above

III. RELATED LITERATURE SURVEY

1) S. Indhumathi. etl presents a prediction of high risk cardiovascular disease employing a Naïve Bayes algorithm. The preprocessed data has been considered because the training set. Two phases namely classification and prediction were discussed in this work. Preprocessing is finished within the classification phase. The preprocessing includes cleaning of information, normalisation and reduction of knowledge,etc. within the prediction phase the disease types are classified and predicted, i.e. a training set is created supported the disease type, and also the test set is made supported the questions. the anticipated results are sent to the doctor. ANN,often just called a"neural network", may be a mathematical modeler computational model used for a biological purpose. In Other Words, it is an emulation of a biological neural system.[5]
2) The authors have proposed an information mining model for the prediction of heart conditions. Dataset was taken from a UCI machine learning repository site. Four data processing algorithms like Naïve Bayes, random forest, regression, and Decision tree were applied by the authors to predict heart conditions. Among these algorithms, the random forest gives a decent accuracy of 90.16% compared to other algorithms.[6]

IV. MACHINE LEARNING

Machine learning may be a method of knowledge analysis that automates analytical model building. it's a branch of AI-supported the thought that systems can learn from data, identify patterns and make decisions with minimal human intervention.[1] So, from the definition, it's clear that Machine Learning has 3 steps during which the entire model works. Those 3 steps/phases are

A. Data Enrichments

Many times data is stored in tabular form which is termed datasets. this can be a noticeable tabular structure consisting of some attributes having values. Now, during this dataset, we have a bunch of redundancy, errors, and null values. Which is hazardous for the model. This inconsistent data set will affect the accuracy of the model and ruin the output. to induce an accurate and consistent result this step is vital.[2]

1) *Dataset:* The dataset may be a plain tabular representation of the non-linear, uneven, and easy types of data. The datasets for this project are collected from Kaggle-
For Heart disease: https://www.kaggle.com/hosamwajeeh/heart-disease-91-8-4-models/data
For diabetes: https://www.kaggle.com/code/ahmetcankaraolan/diabetes-prediction-using-machine-learning/data

B. Feature Engineering

Visualize your data as a full to work out if there are any links between columns. By using charts, you'll be able to see the feature side-by-side and detect any links among features, and between features and labels. Sometimes we'll generate additional features from existing ones during a classification. you'll find yourself with a large number of columns. during this case you wish to decide on the columns, you'll use as features, but if you have got thousands of columns (i.e. potential features), you'll have to apply Dimensionality Reduction. There are several techniques available to try to do this, including Principal Component Analysis or PCA. PCA is an unsupervised learning algorithm that uses existing columns to get new columns called principal components, which might be used later by the classification algorithm.[2]

C. Model Construction-

In this stage, we get a divorce the info is set in three parts-

1) Training data are going to be wont to train your chosen algorithms,
2) Testing data are accustomed check the performance of the results.
3) Validation data are used at the very end of the method, or if necessary, rarely checked out or used before that, to avoid introducing any bias to the result.
4) Choose the relevant algorithms and take a look at various algorithms and their combinations for improving the performance. Use Hyper-parameters (ex. GridSearch in Python) procedure to undertake many combinations, to seek out the one that yields the simplest result (do not try manual combinations).
5) Continue testing after every new model. If the results aren't satisfactory.[2]

V. ALGORITHM AND MODEL

An algorithm is a set of rules that must be followed when solving a particular problem. Here, the Random Forest Algorithm which has ample accuracy is used.

A. Random Forest Algorithm

Random forest could be a Supervised Machine Learning Algorithm that's used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average just in case of regression.[3]
One of the foremost important features of the Random Forest Algorithm is that it can handle the information set containing continuous variables within the case of regression and categorical variables within the case of classification. It performs better results for classification problems.[3]

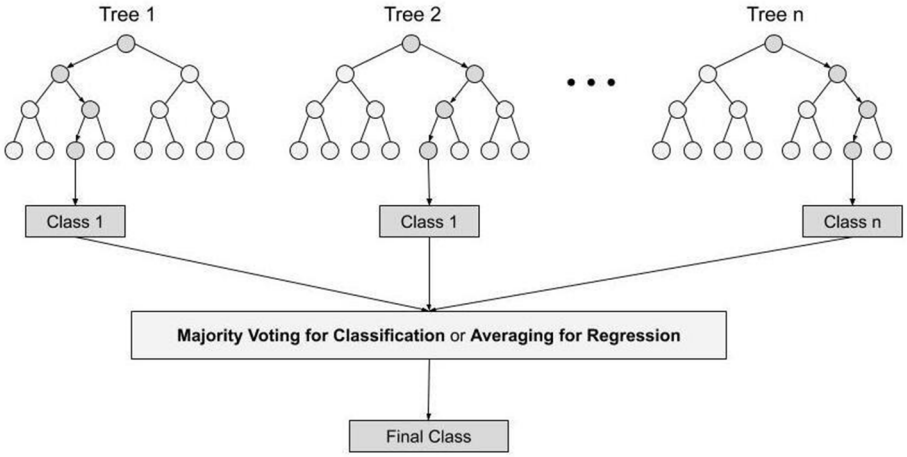


Fig. 1: Random Forest Algorithm

Accuracy Score, here accuracy score refers to the accuracy classification score which is in multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels. In machine learning, it's important to understand the success rate of the algorithm that's used as a metric for that.

The accuracy Classification score for the present model is: 84.60%

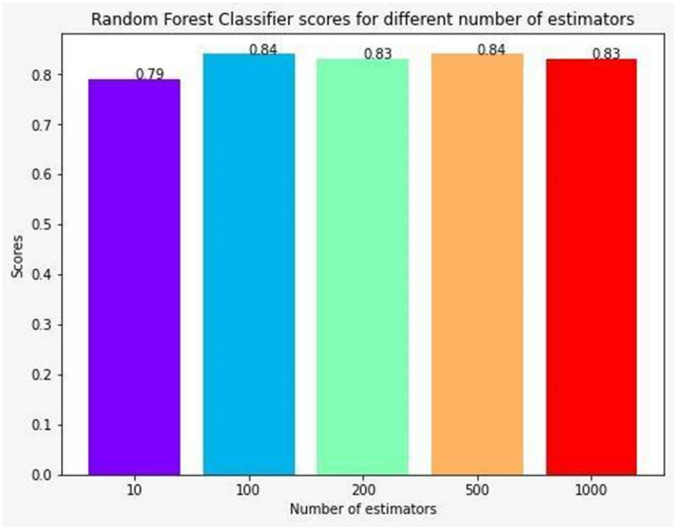


Fig. 2: Random forest classifier score for various numbers of estimators

B. Code

https://drive.google.com/drive/folders/1yqpbaJLts__eX4ea_mX9RAka2HNOayPe?usp=sharing

C. Presentation

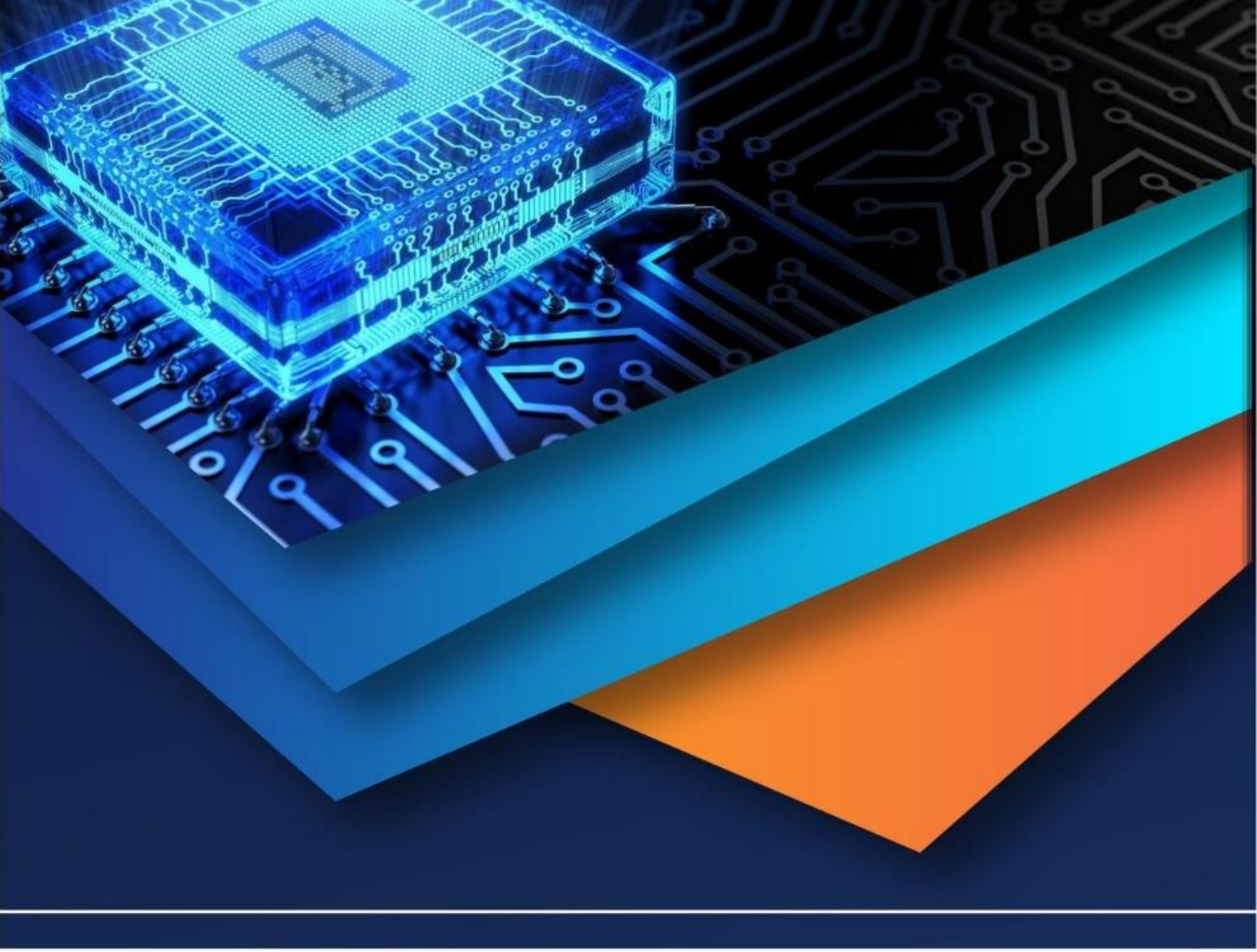
It's the last step in any project. Finally finally the event and testing the foremost crucial step is presenting the model during a simple UI. Where it should be easy to use. The frontend development for this project is finished on Django(python).
Link for project UI:
https://drive.google.com/drive/folders/10hR1EokQo4BUweEuSmSXVW1EqOmMAYPG

D. Results

Here the results the result is obtained from model was in the form of 1's and 0's as yes and no. So, it needs to be simplified and converted into textual form.

REFERENCES

[1] ML definition
https://www.sas.com/en_in/insights/analytics/machine-learning.html
[2] Machine Learning Steps
https://www.spiria.com/en/blog/artificial-intelligence/the-3-vital-steps-of-machine-learning/
[3] Random Forest Classification
https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
[4] Dataset
For Heart disease:
https://www.kaggle.com/hosamwajeeh/heart-disease-91-8-4-models/data
For diabetes:
https://www.kaggle.com/code/ahmetcankaraolan/diabetes-prediction-using-machine-learning/data
[5] S.Indumathi, Mr.G.Vijayabaskar, "Web-Based Healthcare Detection Using naive Bayes algorithm" International.
[6] Rajdhan Apurb, Agarwal Avi, Sai Milan, Ravi Dundigalla, Ghuli Poonam." Heart Disease Prediction using Machine Learning" INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY