IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

Heart Stroke Prediction using ML

R. Indu[1], R. Triveni[2], N. Kavya Sri[3], V. Sasi Kumar[4]
B. Tech Students, Department of Information Technology[1,2,3,4]
Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India

Abstract: *As we all know, the human body functions through a variety of organs, with the heart being one of the most important. We can see that the number of deaths from heart attacks has increased in recent days. Even at a young age, people are being affected by heart attacks. People are realizing that they have a heart attack at the end stage, when there are fewer chances of curing it. As a result, many people are losing their lives because they are unable to detect it at an early stage. If we can predict whether a person will have a heart attack or not at an early stage, we may be able to cure that person and save their life. This paper focuses on developing a prediction model for heart stroke using age, hypertension, previous heart disease status, average body glucose level, bmi, and smoking status as parameters. A random forest algorithm is used to create the prediction model.*

Keywords: Healthcare dataset stroke data, NumPy, Pandas, Sklearn, Flask, Random Forest algorithm

I. INTRODUCTION

Heart disease is even being dubbed a "silent killer," as it can kill a person without causing obvious symptoms. The disease's nature is the source of growing concern about the disease and its consequences. As a result, ongoing efforts are being made to forecast the possibility of this deadly disease in advance. As a result, various tools and techniques are being tested regularly to meet today's health needs. Machine Learning techniques can be extremely useful in this regard. Even though heart disease can manifest itself in various ways, there is a common set of core risk factors that influence whether someone is at risk for heart disease or not. We can conclude by collecting data from various sources, categorizing it under appropriate headings, and then analyzing it to extract the desired data. This technique can be very well adapted to heart disease prediction. As the well-known adage goes, "Prevention is better than cure," and early prediction and control of heart disease can help to prevent and reduce death rates from heart disease[4].

II. PROPOSED SYSTEM

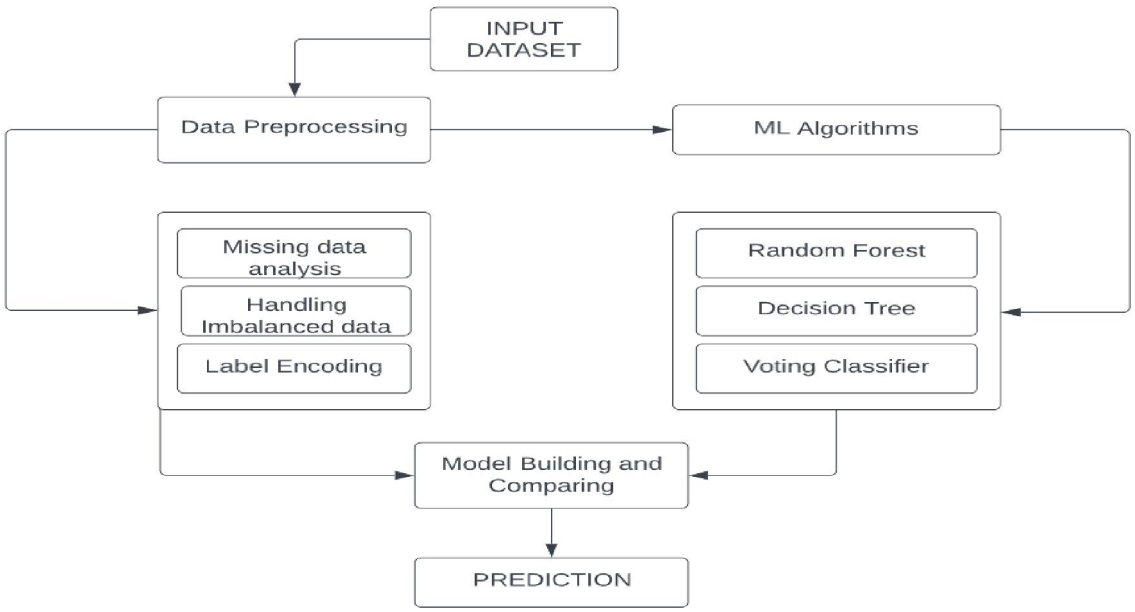

Figure 1.1: Proposed system model

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

One of the most difficult problems in medicine is predicting heart attacks. Every year, the number of people dealing with this issue grows. Several factors increase the risk of having a heart attack. Heart stroke is the leading cause of death in the world today. The World Health Organization (WHO) estimates that heart stroke will kill 12 million people worldwide each year. Machine Learning Algorithms are used to analyze the stroke prediction causes with greater accuracy. This paper focuses on the Random Forest Algorithm, which is one of several Machine Learning Algorithms used for the accurate prediction of heart strokes. The system's operation begins with data collection. The necessary data is then pre-processed into the required format. The data is then separated into training and testing data. Using the training data, the algorithm is applied and the model is trained. The system's accuracy is determined by testing it with the testing data. Finally, using a flask, the model is deployed in PyCharm.

2.1 Attribute Description

To make a stroke prediction, we used the healthcare dataset stroke data from the Kaggle website. There are 12 attributes in the dataset. The following is a description of the attributes:

- **id:** A distinct identifier. It's a patient's id.
- **gender:** This describes the gender of a patient. Male, female, or other
- **age:** This attribute describes the gender of a patient.
- **hypertension:** This characteristic describes the patient's hypertension. If the patient is hypertensive, the value is 0; otherwise, the value is 1.
- **heart disease:** This attribute indicates whether or not the patient has heart disease**.**
- **ever married:** This attribute indicates whether or not the patient is married.
- **work_type:** This attribute describes the patient's work situation.
- **residence_type:** The residence type attribute describes the patient's living situation**.**
- **avg_glucose_level:** This attribute describes the glucose level of the patient.
- **bmi:** This attribute describes the patient's body mass index.
- **smoking status:** This attribute describes the patient's smoking status.
- **stroke:** The patient's stroke is described by this attribute. If the patient has a stroke, the value is 1, otherwise, it is 0[1].

2.2 Random Forest Classifier

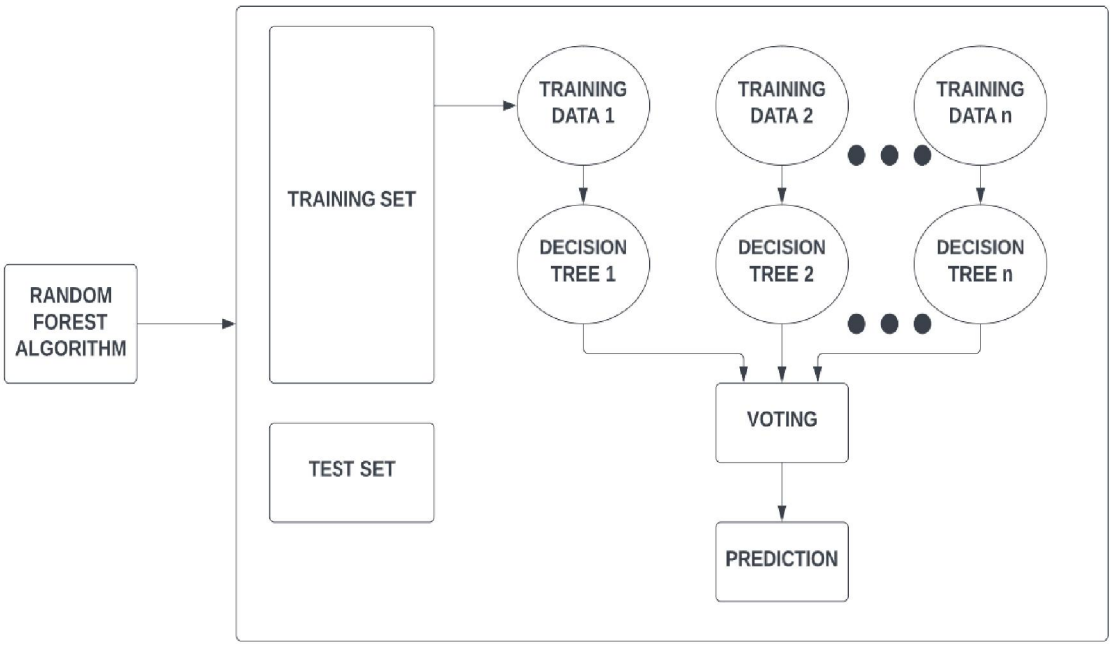

Figure 1.2: Working principle of Random Forest Classifier

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

The above diagram depicts the operation of a random forest classifier. The training dataset is divided into n subsets, each of which contains n decision trees that predict an output. The final output is predicted by voting on the output of each decision tree [3].

III. TECHNOLOGIES USED

3.1 Jupyter Notebook

The original web application for creating and sharing computational documents is Jupyter Notebook. It provides a straightforward, streamlined, document-centric experience. The Jupyter Notebook is an extremely powerful tool for developing and presenting data science projects interactively[5].

3.2 Python Libraries

The libraries of python used here are NumPy, pandas, and sklearn. The NumPyis to operate the arrays and matrices, pandas are used to deal with the ML tasks and data analytics and also to load the data from the CSV file to the data frame of the panda. Sklearn provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering, andso on.[6] Flask is a web framework, it's a Python module that lets you develop web applications easily[7].

3.3 PyCharm

PyCharm is an Integrated Development Environment (IDE) that provides a platform for developers to create Python, web, and data science applications[8].

3.4 Data

The data is taken in the form of a CSV file which contains the data in the form of rows and columns. The data has columns named age, gender, heart_disease, residence_type, work_type, avg_glucose_level, bmi, smoking_status, ever_married, and stroke. Here, we focus on bmi, age, avg_glucose_level, and smoking_status in major[1].

IV. RESULTS

4.1 Loading Dataset

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Figure 1.3: First five rows of the dataset.

The above figure gives an overview of the columns of the dataset that is taken for the prediction of heart stroke[2].

4.2 Data Description

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

Figure 1.4: Description of the dataset.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

The table above describes the dataset that is taken by us. The description includes the total count, mean, standard deviation (std), minimum value, maximum value, and percentiles (25%, 50%, 75%) of every column present in the dataset[2].

4.3 Data Pre-Processing

```
df.isnull().sum()
```

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
stroke               0
dtype: int64
```

Figure 1.5: Data without null values

The above figure describes that the data has no more null values and missing values. The data is cleaned using the python library pandas[2].

4.4 Splitting the Dataset

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test=train_test_split(X,Y,test_size=0.2,random_state=10)
X_train
```

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
2285	1	49.0	0	0	1	2	0	79.64	28.893237	3
4733	1	67.0	0	0	1	2	0	83.16	25.500000	1
3905	1	78.0	0	0	1	2	1	208.85	24.400000	1
4700	1	47.0	0	0	1	2	0	110.14	30.500000	3
4939	0	59.0	0	0	1	2	1	71.08	28.100000	2
...	...	...	...	...	...	...	...	...	...	...
1180	0	62.0	0	0	1	2	0	82.57	36.000000	1
3441	0	59.0	0	0	1	3	1	90.06	28.900000	3
1344	1	47.0	0	0	1	2	0	86.37	39.200000	3
4623	1	25.0	0	0	1	0	1	166.38	23.100000	2
1289	0	80.0	0	0	1	3	0	72.61	27.600000	2

4088 rows × 10 columns

Figure 1.6: Splitting the data into training data and testing data

The above figure describes that the dataset is split into the training dataset and testing dataset by using the train_test_split which is imported from the sklearn. The training data is 80% and the testing data is 20%[2].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

4.5 Training The Model

```
from sklearn.ensemble import RandomForestClassifier
rf=RandomForestClassifier()
```

```
rf.fit(X_train_std,Y_train)
```

```
RandomForestClassifier()
```

```
rf.feature_importances_
```

```
array([0.02863354, 0.23305129, 0.0266067 , 0.02595681, 0.01800435,
       0.05078092, 0.03496413, 0.27226903, 0.24031943, 0.06941381])
```

```
X_train.columns
```

```
Index(['gender', 'age', 'hypertension', 'heart_disease', 'ever_married',
       'work_type', 'Residence_type', 'avg_glucose_level', 'bmi',
       'smoking_status'],
      dtype='object')
```

```
prediction=rf.predict(X_test_std)
```

Figure 1.7: Training the Random Forest Algorithm

The above figure describes that the random forest classifier is trained with the training dataset. We have imported the random forest classifier from sklearn[2].

4.6 Testing the Model

```
Y_test
```

```
2413    0
1141    0
146     1
3883    0
1044    0
       ..
2261    0
4712    0
4971    0
2224    0
4825    0
Name: stroke, Length: 1022, dtype: int64
```

Figure 1.8: Prediction on the testing data.

The above figure describes the random forest algorithm predicting the stroke on testing data. The output is represented in 0's and 1's[2].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

4.7 Web Deployment in the Local Host

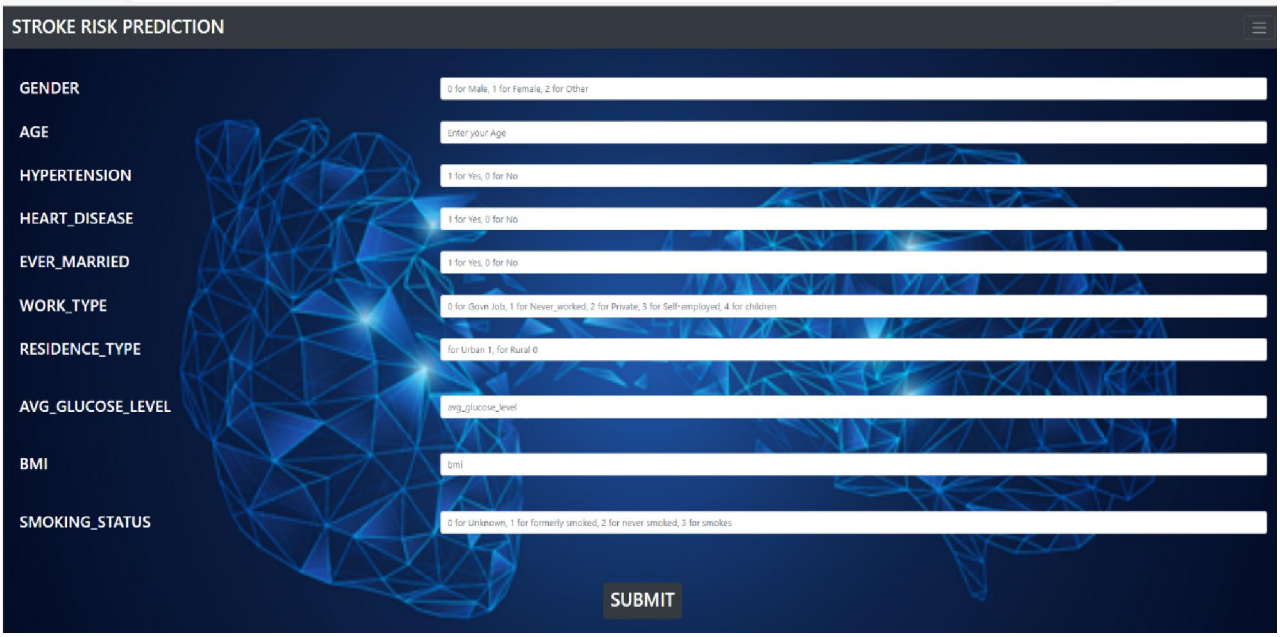

Figure 1.9: Home Page

The above figure describes the home page just like a user interface where the user can give custom inputs in the form and submit it to get the result. We have used flask which is used as a web framework[2].

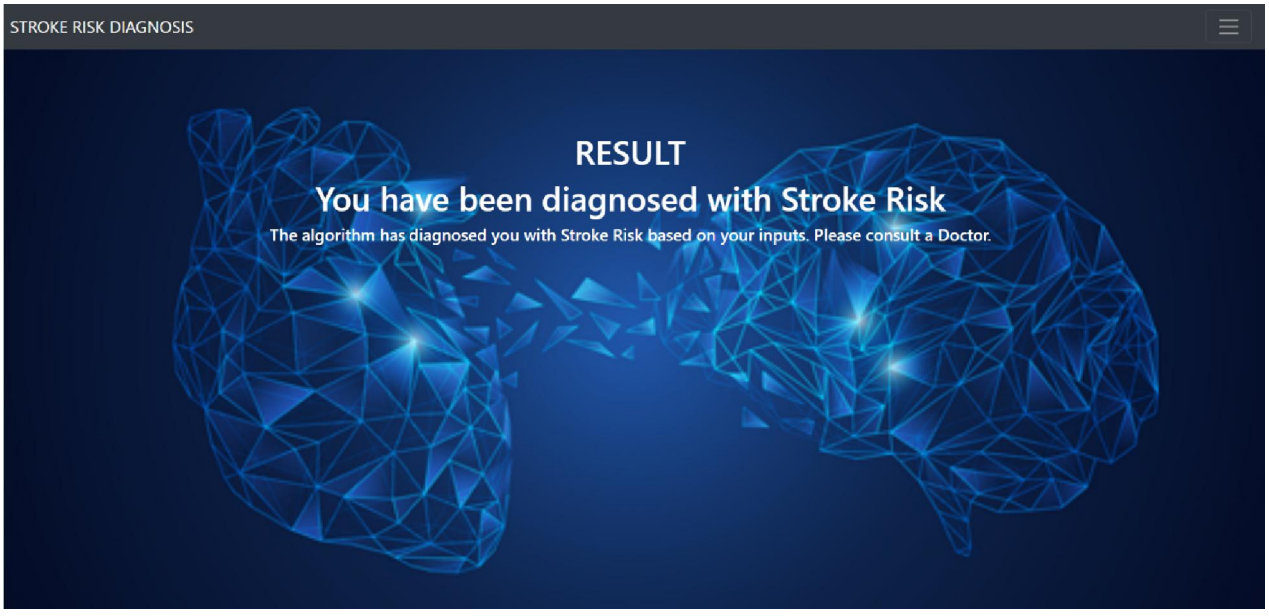

Figure 2.0: Result page for stroke

The above picture describes that the algorithm has diagnosed the patient with a heart stroke and instructs the patient to meet the doctor immediately[2].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, November 2022
Impact Factor: 6.252

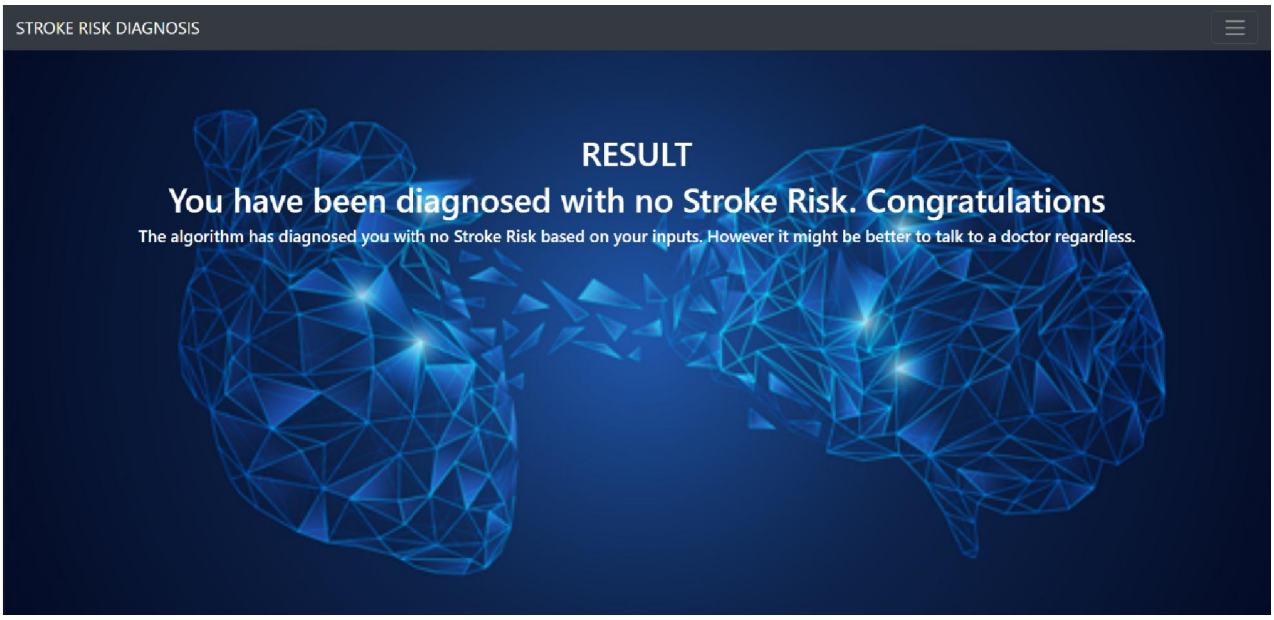

Figure 2.1: Result page for no stroke

The above picture describes that the algorithm has diagnosed the patient with no heart stroke[2].

V. FUTURE SCOPE

We can develop this by using various other ML algorithms which provide much more accurate results along with ease of usage. Apart from the feature importances, we can also use other attributes for the prediction. In the future, we can also use cloud resources like Azure ML Studio, etc.,

VI. CONCLUSION

Therefore, wehave used the Random Forest Classifier to predict the heart stroke based on various inputs like age, gender, heart_disease, residence_type, work_type, avg_glucose_level, bmi, smoking_status and ever_married. The Random Forest classifier has predicted the heart stroke with an accuracy of 95%. Through the feature importances avg_glucose_level, bmi and age are used as major data in predicting heart stroke. So, by predicting heart stroke we could be able to prevent ourselves from heart stroke in the future.

REFERENCES

[1]. Dataset - https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
[2]. Notebook - https://github.com/19501A1295/Heart-Stroke-Prediction-using-ML
[3]. https://www.javatpoint.com/machine-learning-random-forest-algorithm
[4]. https://ijcrt.org/papers/IJCRT2106047.pdf
[5]. https://jupyter.org/try-jupyter/retro/notebooks/?path=notebooks/Intro.ipynb
[6]. https://medium.com/personal-project/numpy-pandas-and-scikit-learn-explained-e7336baecedc#:~:text=The%20great%20thing%20about%20Numpy,Scikit%20Learn%20function(s).
[7]. https://pythonbasics.org/what-is-flask-python/
[8]. https://www.jetbrains.com/help/pycharm/quick-start-guide.html#:~:text=PyCharm%20is%20a%20dedicated%20Python,web%2C%20and%20data%20science%20development.