Department of Computer Science Savitribai Phule University , Pune

Masters of Computer Science

A Project Report

On

Brain Stroke Prediction

Name:- Piyusha Chetan Nikam

Roll No. :- N20111058

Project Guide: Prof. R. Korpal

PUNE UNIVERSITY

Computer Science
Department

INTRODUCTION

Burden of Stroke in the World-Stroke is the second leading cause of death and leading cause of adult disability worldwide with 400-800 strokes per 100,000, 15 million new acute strokes every year, 28,500,000 disability adjusted life-years and 28-30-day case fatality ranging from 17% to 35%. The burden of stroke will likely worsen with stroke and heart disease related deaths projected to increase to five million in 2020, compared to three million in 1998. This will be a result of continuing health and demographic transition resulting in increase in vascular disease risk factors and population of the elderly.

Causes of mortality from stroke Death from stroke is as a result of comorbidities and/ or complications. Complications of stroke may arise at different time periods. The beginning of stroke symptoms and the first month following the stroke onset is the most critical period for survival with the highest number of fatalities in the first week. Complications of stroke include hyperglycemia, hypoglycemia, hypertension, hypotension, fever, infarct extension or rebreeding, cerebral edema, herniation, coning, aspiration, aspiration pneumonia, urinary tract infection, cardiac dysrhythmia, deep venous thrombosis and pulmonary embolism among others. During the first week from stroke onset, death is usually due to transtentorial herniation and hemorrhage, with death due to hemorrhage happening within the first three days and death due to cerebral infarction usually occurring between the third to sixth day. One week after the onset of stroke, death is usually due to complications resulting from relative immobility such as pneumonia, sepsis and pulmonary embolism.

Prevention of stroke - More than 70% of strokes are first events, thus making primary stroke prevention a particularly important aspect. Interventions should be targeted at behavior modification, which however requires information about the baseline perceptions, knowledge and prevalence of risk factors in defined populations.

Proposed System / Scope of work

This project helps to predict the stroke risk using prediction model in older people and for people who are addicted to the risk factors as mentioned in the project.

This Project can also be used to find the stroke probabilities in young people and underage people by collecting respectative risk factor information and doctors consulting.

Objectives

The main objective of the project is to find stroke by using machine learning technique and various algorithms. Such as SVM, RFA, DTA.

Here we analysing the accuracy level and evaluate the model performance to choose the best one of the several algorithms. By using this method we can predict the stroke of the patient by observing their symptoms like age, gender, Hypertensions.

Methodology

- To proceed with the implementation, different datasets were considered . Out of all the existing datasets, an appropriate dataset was collected for model building.
- After collecting the dataset, It lies in preparing the dataset to make the data more clear and easily understood by the machine. This step is called as Data preprocessing.
- It includes handling of missing values, handling imbalanced data and performing label encoding that are specific for this particular dataset.
- For model building, preprocessed dataset along with machine learning algorithms are required. Logistic Regression, Decision Tree Classification algorithm, Random Forest Classification algorithm, K-Nearest Neighbor algorithm, Support Vector algorithm are used.

Techniques

- Python-For Programming Logic
- Application:-Used in application for GUI
- Python :- Provides machine learning process

Hardware and software

Hardware:

- Intel Processor
- 4/8 GB RAM
- Min 1 GB Hard Disc

Any Operating System

Software:

- Anaconda, Jupyter Notebook,PyCharm
- Python:

Python is a high-level programming language designed to be easy to read and simple to implement. It is open source, which means it is free to use, even for commercial applications. Python can run on Mac, Windows, and UNIX systems and has also been ported to Java and .NET virtual machines.

Python is considered a scripting language, like Ruby or Perl and is often used for creating Web applications and dynamic Web content. It is also supported by a number of 2D and 3D imaging programs, enabling users to create custom plugins and extensions with Python.

Describe the software / algorithm to be implemented (I/O screens, any statistical / mathematical models that will be implemented)

1] DECISION TREE ALGORITHM:-

- Decision tree learning is one of the prescient displaying approaches utilized in measurements, information mining, and Machine Learning. It utilizes a choice tree to go from perceptions about a thing to decisions about the thing's objective worth.
- Tree models where the objective variable can take a discrete arrangement of qualities are called characterization trees,in these tree structures, leaves speak to class names, and branches speak to conjunctions of highlights that lead to those class names.
- Choice trees where the objective variable can take constant qualities are called relapse trees. In the choice investigation, a choice tree can be utilized to outwardly and expressly speak to choices and dynamics. In information mining, a choice tree depicts information This page manages choice trees in information mining.

2] SUPPORT VECTOR ALGORITHM:-

- In Machine Learning, are managed to learn models with related learning calculations that break down information utilized for characterization and relapse investigation.
- An SVM model is a portrayal of the models as focuses in space, mapped with the goal that the instances of the different classifications are isolated by an unmistakable hole that is as wide as could reasonably be expected.
- New models are then mapped into that equivalent space and anticipated to have a place with a classification dependent on the side of the hole on which they fall. Notwithstanding performing straight characterization, SVMs can productively play out a non-direct grouping utilizing what is known as the bit stunt, verifiably mapping their contributions to high-dimensional component spaces.

3] K-NEAREST NEIGHBOUR ALGORITHM:-

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

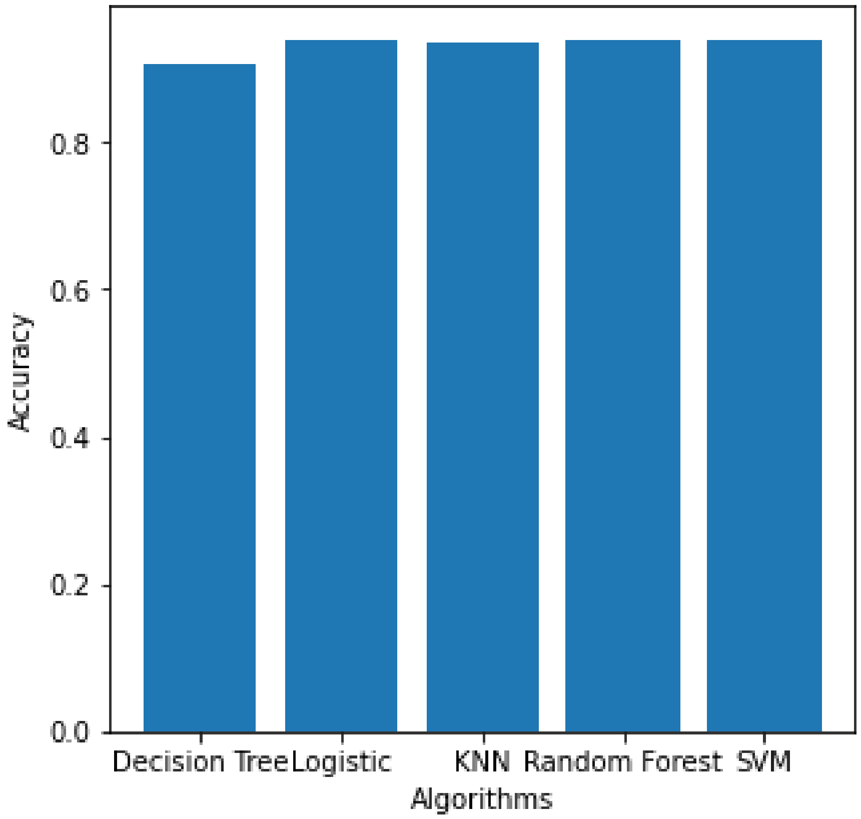
4] LOGISTIC REGRESSION ALGORITHM:-

- The Logistic Regression Predicts the output of a categorical dependent variable. The outcome must be categorical or discreate value. It can yes or No or 0 or 1.
- It is used to classify the observation using different types of data and can easily determine the most effective variable used for the classification.

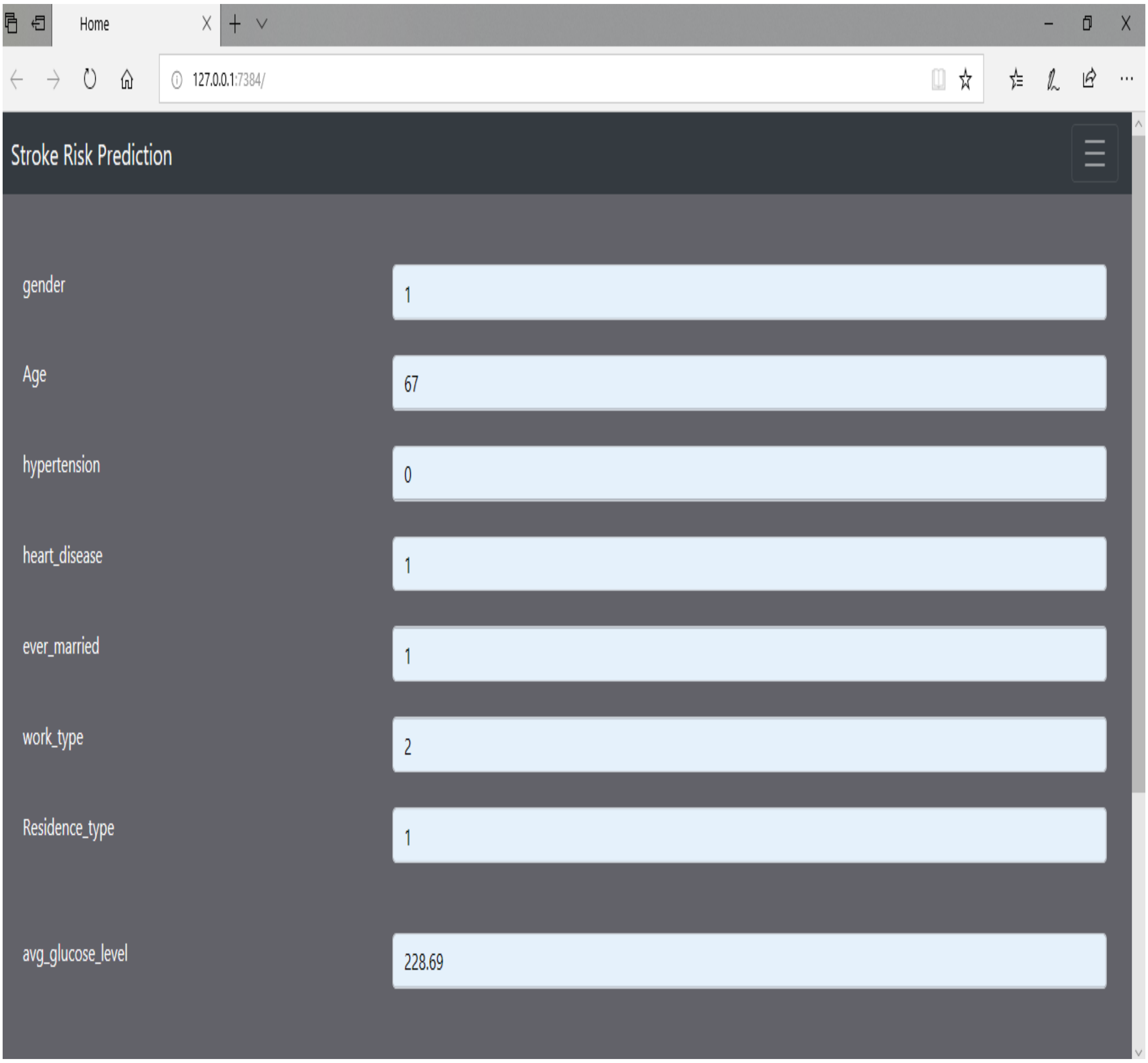
5] RANDOM FOREST ALGORITHM:-

- Random forests or random decision forests are a gathering learning strategy for arrangement, relapse and different assignments that work by building a large number of choice trees at preparing time and yielding the class that is the method of the classes or mean forecast of the individual trees.
- Random choice woodlands right for choice trees' propensity for overfitting to their preparation set. The principal calculation for irregular choice backwoods was made by Tin Kam Ho utilizing the arbitrary subspace method, which, in Ho's plan, is an approach to actualize the "stochastic segregation" way to deal with the order proposed by Eugene Kleinberg.

Visualiazation

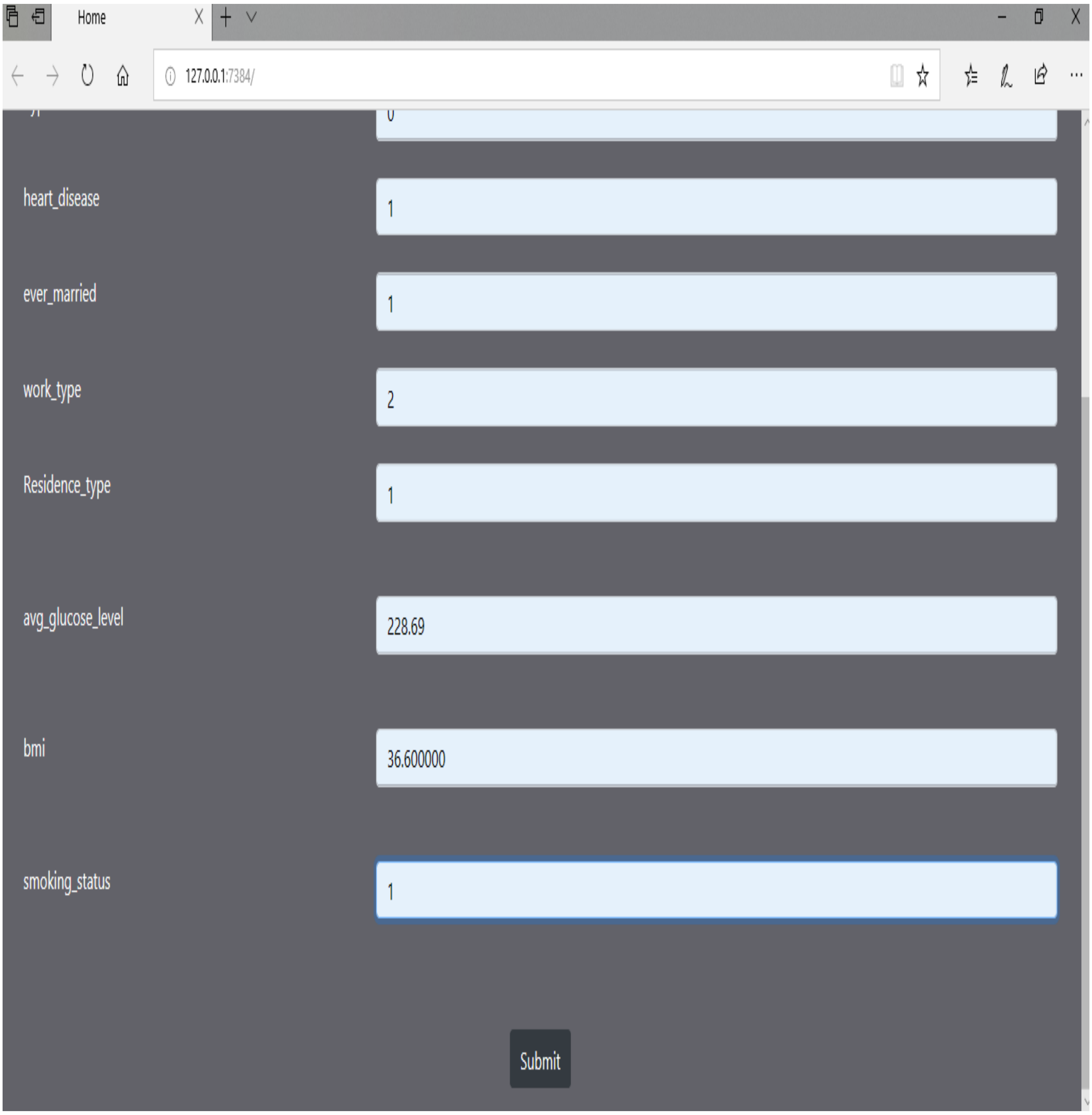


Webpages

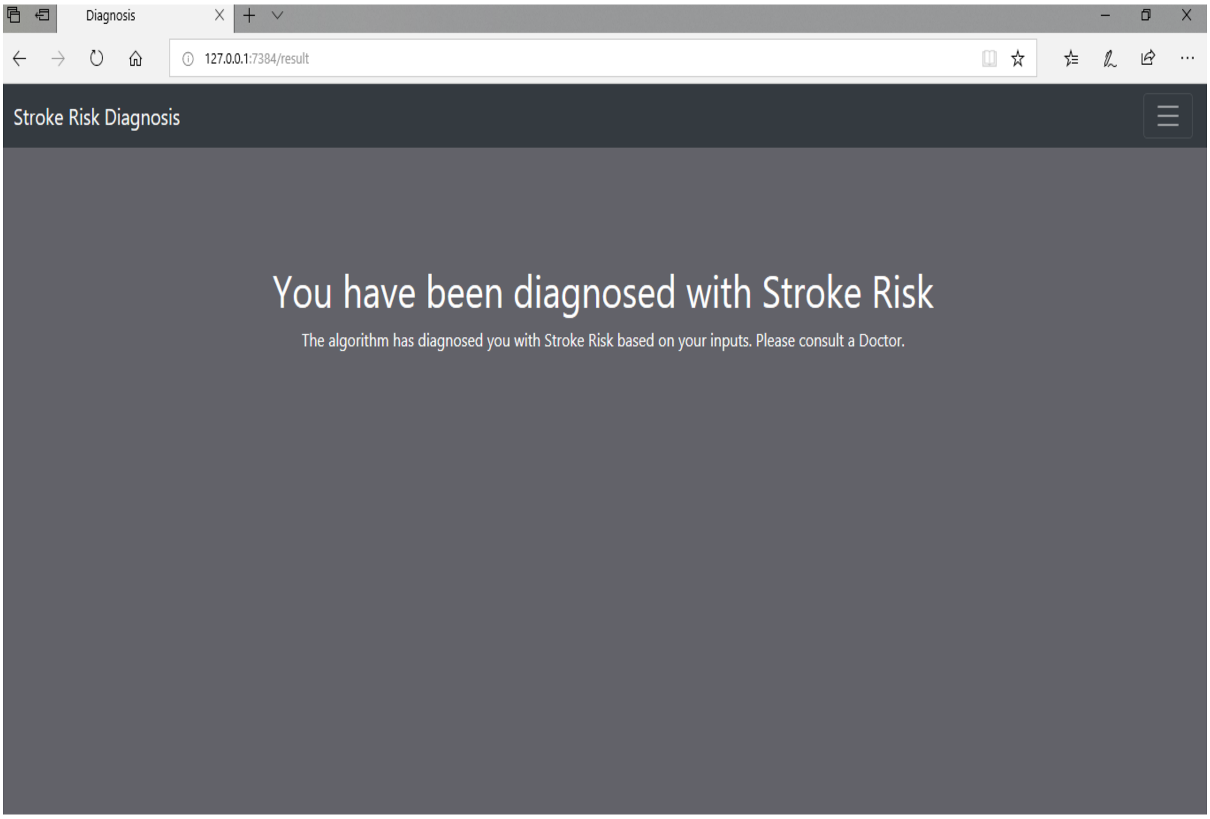


Stroke Risk Prediction

gender	1
Age	67
hypertension	0
heart_disease	1
ever_married	1
work_type	2
Residence_type	1
avg_glucose_level	228.69



	0
heart_disease	1
ever_married	1
work_type	2
Residence_type	1
avg_glucose_level	228.69
bmi	36.600000
smoking_status	1



Stroke Risk Diagnosis

You have been diagnosed with Stroke Risk

The algorithm has diagnosed you with Stroke Risk based on your inputs. Please consult a Doctor.

Output of algorithms:

Algorithm	Accuracy
DECISION TREE ALGORITHM	0.9080234833659491
SUPPORT VECTOR ALGORITHM	0.9393346379647749
K-NEAREST NEIGHBOUR ALGORITHM	0.9344422700587084
LOGISTIC REGRESSION ALGORITHM	0.9383561643835616
RANDOM FOREST ALGORITHM	0.9373776908023483

From the above algorithms logistic regression fits our dataset best.

GitHub Link: **https://github.com/Piyusha14**

References

- https://www.kaggle.com/diamonds
- https://www.w3schools.com/python/ss
- https://www.geeksforgeeks.org