Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF ELECTRONICS ENGINEERING [SENSE]

WINTER SEMESTER 2022-23

ECE3502

IOT Domain Analyst

J Component

Health Monitoring and Heart Stroke Prediction

By:

20BEC0609 - B. Chetan Sai

20BEC0784 - Gurujala Jaswanth

20BEC0266 - M. Sree Dhathri

Submitted to:

Professor – Dr. Biswajit Dwivedy

Aim:

To monitor the health of the patient and predict the approach for stroke prediction using machine learning and neural networks.

Abstract:

A stroke occurs when the blood supply to part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die in minutes. A stroke is a medical emergency, and prompt treatment is crucial. Early action can reduce brain damage and other complications. In 2015, there were about 42.4 million people who had previously had a stroke and were still alive. In 2015, stroke was the second most frequent cause of death after coronary artery disease, accounting for 6.3 million deaths. About 3.0 million deaths resulted from ischemic stroke while 3.3 million deaths resulted from haemorrhagic stroke. Hence, correct detection and finding presence of stroke inside a human becomes essential. There are various medical instruments available in the market for predicting brain stroke, but they are very much expensive, and they are not efficient enough to be able to calculate the chance of having a brain stroke. So, there is a need to find better and efficient approach to diagnose brain strokes at an early stage.

Heart disease patients have many of those visible risk factors in common which may be used very effectively for diagnosis. A system based on such risk factors would not only help medical professionals, but it would give patients a warning about the probable presence of heart disease even before the patient visits a hospital or goes for costly medical check-ups. Hence this paper presents a technique for prediction of heart disease using major risk factors with help of different Classifying Algorithms. This technique involves four major classification algorithms such as K Neighbours, Support Vector, Decision Tree, Random Forest algorithms.

Keywords – Heart Disease, Machine Learning, Logistic Regression, KNN, Prediction.

Introduction:

Heart disease cases are rising at an alarming rate, and it's critical and to be able to predict these diseases in advance. The project focuses on predicting which patients are more likely to have heart disease based on a variety of medical factors. To predict and identify patients with heart disease, we used different algorithms such as logistic regression and KNN. The proposed model's accuracy was quite good, and it was able to predict signs of heart disease in a person. This heart disease predictive method improves patient treatment and makes diagnosing the disease easier along with allowing exploring large data at once.

With tons of new healthcare technology start-ups, IoT is rapidly revolutionizing the healthcare industry. Keeping track of the health status of your patient at home is a difficult task because of the busy schedules and our daily life work. Specially old age patients should be periodically monitored. So, we propose an innovative system that automated this task with ease. We will use the MAX30100/102 Pulse Oximeter sensor to measure the Heart Rate/Pulse (BPM) as well as the Blood Oxygen Level (SpO2).

Artificial Intelligence techniques have been widely used in clinical decision support systems for prediction and diagnosis of various diseases with good accuracy. These classifying techniques are very effective in designing clinical support systems due to their ability to get hidden patterns and relationships in medical data provided by medical professionals. One of the most important applications of such systems is in the diagnosis of heart diseases because it is one of the leading causes of deaths all over the world. Almost all systems that predict heart diseases using clinical dataset having parameters and inputs from complex tests conducted in labs. None of the systems predicts heart diseases supporting risk factors like age, case history, diabetes, hypertension, high cholesterol, tobacco smoking, alcohol intake, obesity, or physical inactivity, etc.

Cardiovascular disease (CVD) remains the leading cause of death for adults in the United States (US) with an estimated 85.6 million Americans experiencing some form of CVD [1,2]. The term CVD is used to describe disorders of the heart and blood vessels such as coronary heart disease, stroke, congestive heart failure, and arrhythmias. African Americans comprise 13.3% of the US population (46.3 million people) yet have a three-fold greater risk of developing CVD and a two-fold greater risk of CVD related mortality than that of non-Hispanic whites and other ethnic groups.

Heart diseases are often used in exchange for cardiovascular diseases. These kinds of diseases mainly refer to the conditions of blocked or narrowed blood vessels, resulting in a stroke, chest pain or angina, and heart attack.

MAX30100:

Pulse Oximeters are low-cost non-Invasive medical sensors used to continuously measure the Oxygen saturation (SPO2) of haemoglobin in blood. It displays the percentage of blood that is loaded with oxygen.

The principle of pulse oximetry is based on the differential absorption characteristics of oxygenated and the de-oxygenated haemoglobin. Oxygenated haemoglobin absorbs more infrared light and allows more red light to pass through. Whereas Deoxygenated haemoglobin absorbs more red light and allowing more infrared light to pass through.

Each pulse oximeter sensor probe contains two light emitting diode one emitting red light and the other emitting near infrared light, it also has a photo-detector. The photodetector measures the intensity of transmitted light at each wavelength. And using the differences in the reading the blood oxygen content is calculated. The probe is placed on a suitable part of the body, usually a fingertip or ear lobe.

Literature survey:

1. "A predictive analytics approach for stroke prediction using machine learning and neural networks."

Methodology:

Stroke diseases can be divided into ischemic stroke and haemorrhagic stroke, and they should be minimized by emergency treatment such as thrombolytic or coagulant administration by type. First, it is essential to detect in real time the precursor symptoms of stroke, which occur differently for everyone, and to provide professional treatment by a medical institution within the proper treatment window.

Objective:

The choice of feature-selection methods can help in improving the prediction accuracy of the model and efficient data management of the archived input features.

Introduction:

Among them, 6 million deaths were due to cerebrovascular disease, which was reported to be the second leading cause of death. A timely detection and prevention of stroke has become very essential to avoid its adverse consequences. The field of medical sciences has observed tremendous improvements due to the rise in technological advancements over time.

Advantages:

➢ Explained Principal Component Analysis very well as per the Work.

➢ Prediction is Good.

➢ Collected Good Electronic Health Records (EHR).

Dis-Advantages:

➢ No Disadvantages. Explained very well as per problem statement.

Conclusion:

Efficient stroke-detection methods have been increasingly studied in recent years. The selection of important features from the high dimensional medical dataset is critical to the prediction model's performance. Existing research on automatic detection of stroke risk through data mining techniques faces a significant challenge in the selection of effective features as predictive cues.

2. "Heart Disease Prediction Using Data Mining Classification Algorithms".

Methodology:

This paper systematically analyses the various factors in electronic health records for effective stroke prediction. Using various statistical techniques and principal component analysis, we identify the most important factors for stroke prediction. We conclude that age, heart disease, average glucose level, and hypertension are the most important factors for detecting stroke in patients.

Objective:

Based on the reviews of previous research the objectives of this research are to predict heart diseases using data mining classification techniques such as naive bias, k-nearest Neighbour, logistic regression decision tree.

Advantages:

➢ Explained Principal Component Analysis and Data Mining Techniques very well.

➢ Perfect Use of Decision tree classification in Python.

➢ Correct F1 Score is calculated.

Conclusion:

In this paper, we presented a detailed analysis of patients' attributes in electronic health record for stroke prediction. We systematically analysed different features. We performed feature correlation analysis and a step wise analysis for choosing an optimum set of features. We found that the different features are not well-correlated and a combination of only 4 features (A, HD, HT and AG) might have good contribution towards stroke prediction.

3. AI-Based Stroke Disease Prediction System Using ECG and PPG Bio-Signals.

Methodology:

Stroke is considered one of the most serious diseases in modern society as it can cause death in severe cases, while also leading to physical and mental disorders such as hemiparesis, speech impairment(aphasia), ataxia, visual impairment, consciousness impairment, and dementia. Some of them indicates that they cannot use optimization techniques to improve their model performance.

Objective:

The Objective of this paper is to Diagnosis & Prognostic treatment, by analysing the real – time Predictions.

Advantages:

➢ Good Collection & Pre – processing of Data.

➢ Perfect Multimodal of ECGs & PPGs are used for Machine Learning.

➢ The Deep Learning Techniques used, are best to Perfect Prediction.

Dis - Advantages:

They have certain disadvantages such as a long measurement time, high cost, and radiation exposure. Therefore, in this study, we utilized the clinical results showing that abnormalities in the autonomic nervous system and sympathetic nervous system can occur due to the prognostic symptoms of stroke.

Conclusion:

In Addition to that, we performed principal component analysis (pca). The analysis showed that almost all principal components are needed to explain a higher variance. The variable loadings however showed that the first principal component which has the highest variance might explain the underlying phenomenon of stroke prediction. Finally, three machine learning algorithms were implemented on a set of different features and principal components configurations.

4. Stroke Disease Detection and Prediction Using Robust Learning Approaches

Methodology:

Stroke is a life-threatening medical illness that should be treated as soon as possible to avoid further complications. The development of an ML model could aid in the early detection of stroke and the subsequent mitigation of its severe consequences. The effectiveness of several ML algorithms in properly predicting stroke based on several physiological variables is investigated in this study.

Objective:

The aim of this study was to apply computational methods using machine learning techniques to predict stroke from data set.

Proposed Algorithms:

Random Forest, Decision tree, Voting classifier, Logistic regression

Advantages:

➢ Explained very well as per their proposed work.

➢ Prediction through graphs is good. Used graphical representation for clarity and understanding to user.

➢ Priority of explanation is good

Disadvantages:

We can also use different algorithms like SVC, Decision tree algorithm but it won't produce better accuracy. This method is cost-efficient (saves money) and faster than other algorithms.

Conclusion:

According to the research, the random forest method outperforms other processes when cross-validation metrics are used in brain stroke forecasting. Random forest classification outperforms the other methods tested with a classification accuracy of 96 percent. The future scope of this study is that using a larger dataset and machine learning models, such as AdaBoost, SVM, and Bagging, the framework models may be enhanced. This will enhance the dependability of the framework and the framework's presentation.

5. Predicting Risk of Stroke from Lab Tests Using Machine Learning

Methodology:

They used the National Health and Nutrition Examination Survey data sets with three different data selection methods (ie, without data resampling, with data imputation, and with data resampling) to develop predictive models. They used four machine learning classifiers and six performance measures to evaluate the performance of the models.

Advantages:

More concentrated on specific algorithms. Background verification and classification was too good. Worked on Surveys as well as per Particular selection.

Disadvantages: Not concentrated on separation of data clearly. Some point of view, data was not sufficient to generation.

Result:

The correlations between these different lab tests and stroke were found in several studies. However, this is the first study that used all of these different attributes to build a prediction model using machine learning algorithms. Our results showed that a prediction model can be created using the random forest algorithm and could achieve an accuracy of 0.96.

Conclusion:

he predictive model, built using data from lab tests, was easy to use and had high accuracy. In future studies, we aim to use data that reflect different types of stroke and to explore the data to build a prediction model for each type.

6. Stroke Prediction using Machine Learning Methods.

Methodology:

They presented a detailed analysis of patients' attributes in electronic health record for stroke prediction and systematically analysed different features. They performed feature correlation analysis and a step wise analysis for choosing an optimum set of features and found that the different features are not well-correlated and a combination of only 4 features might have good contribution towards stroke prediction. Additionally, they performed principal component analysis.

Advantages:

Prediction based on the graphs like visualization of data is too good. Component analysis is explained clearly. Relation between principal components and patient attributes was superb.

Disadvantages:

Technology usage and orientation is less. Not explained very clearly about the CNN model and not deep into other technologies

Conclusion:

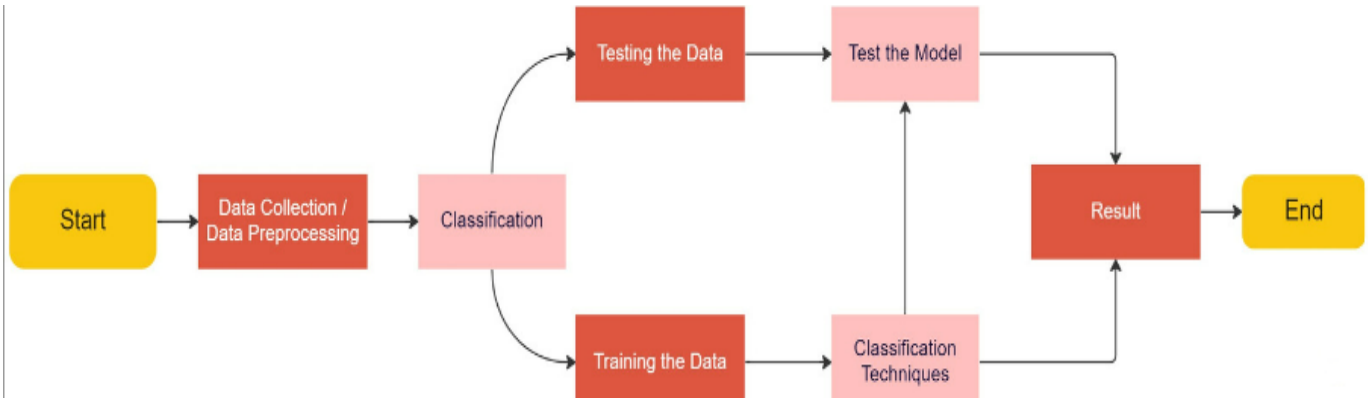
Analysis showed that almost all principal components are needed to explain a higher variance. The variable loadings however showed that the first principal component which has the highest variance might explain the underlying phenomenon of stroke prediction. Finally, three machine learning algorithms were implemented on a set of different features and principal components configurations. We found that neural network works the best with a feature

combination of A, HD, HT and AG. The accuracy and miss rate for this combination are 78% and 19% respectively.

Proposed Work:

- Collecting sensor Data
- Uploading data to ThingSpeak
- Gathering the dataset for prediction
- Data Cleaning - Missing Values and unique values
- Data Visualization - Correlation
- Scatter - between age, average glucose level and bmi, average glucose level
- Heatmap
- Boxplot of BMI
- Outliers for BMI
- Checking for nulls
- Executing the Classification Models
- Comparing the Testing and Training Scores of Models
- Creating the website for user application.

Block Diagram:



The block diagram represents the general flow of a machine learning project. Each block represents a specific phase in the process, as explained below:

- Start: This block indicates the beginning of the machine learning project, where the project goals and objectives are defined.

- Data Collection/Data Pre-processing: In this phase, data is collected from various sources, such as databases, online sources, we used Kaggle for this purpose. We pre-processed to ensure that the data is in a suitable format for analysis. Data pre-processing may involve cleaning, normalization, or feature selection to ensure that the data is suitable for the next stage of the project.

- Classification: In this phase, the pre-processed data is used to train a machine learning algorithm that can classify new data accurately. The classification algorithm uses the pre-processed data to learn patterns and relationships between the input features and the output classes.

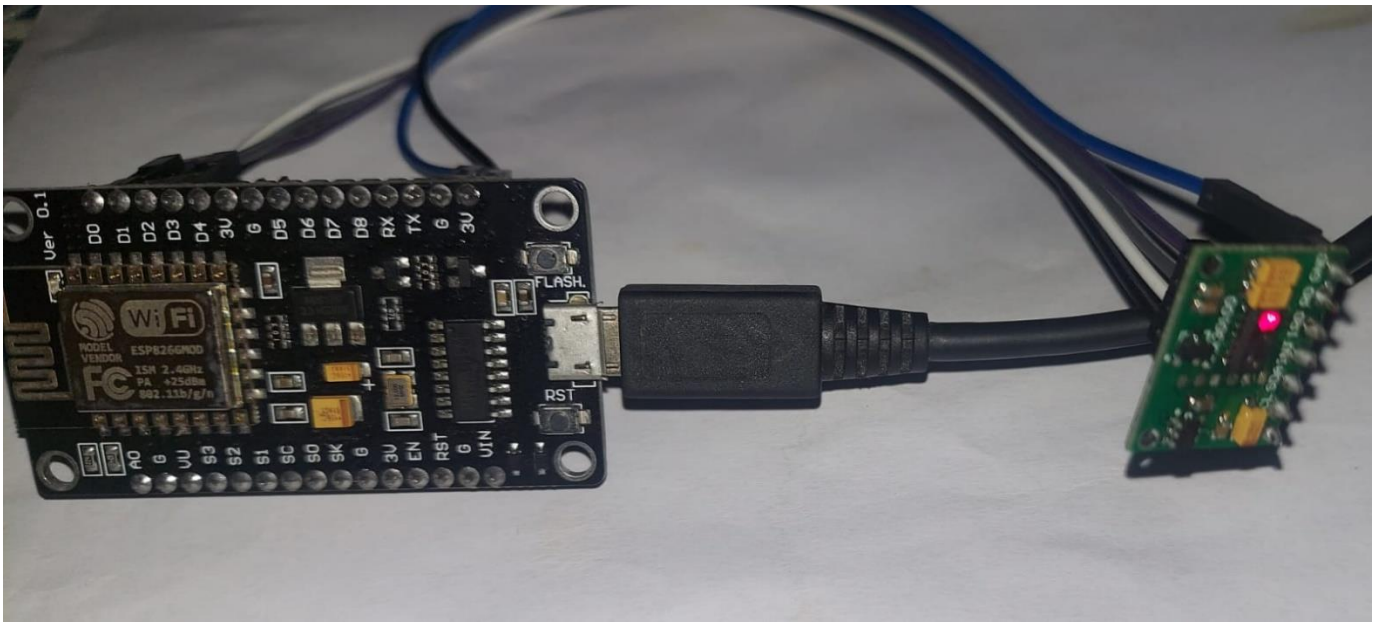
- Testing and Training: In this phase, the trained algorithm is tested on new data to ensure that it can accurately classify new data. The algorithm is evaluated using metrics such as accuracy, precision, recall, and F1 score to measure its performance.

- Result: This block represents the output of the machine learning project, which includes the classification results, as well as any other relevant metrics or visualizations.

- End: This block represents the end of the machine learning project, where the project goals and objectives are evaluated to determine whether they have been achieved successfully.

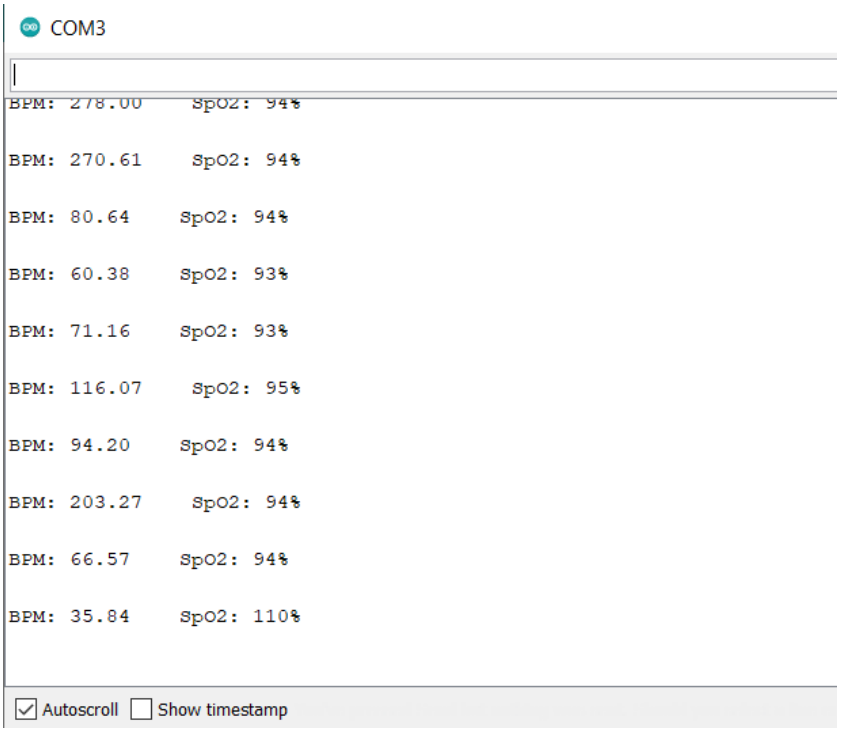
Overall, the block diagram provides a simplified overview of the machine learning project's flow, from data collection and pre-processing to classification, testing, and training. Each block in the diagram represents a specific phase in the project, and the output of each phase is used as input for the subsequent phase. The final output of the project is the classification results and any other relevant metrics or visualizations.

Circuit:



Observations:

The values of the sensors BPM and SpO2 (oxygen) are monitored in the serial monitor continuously .

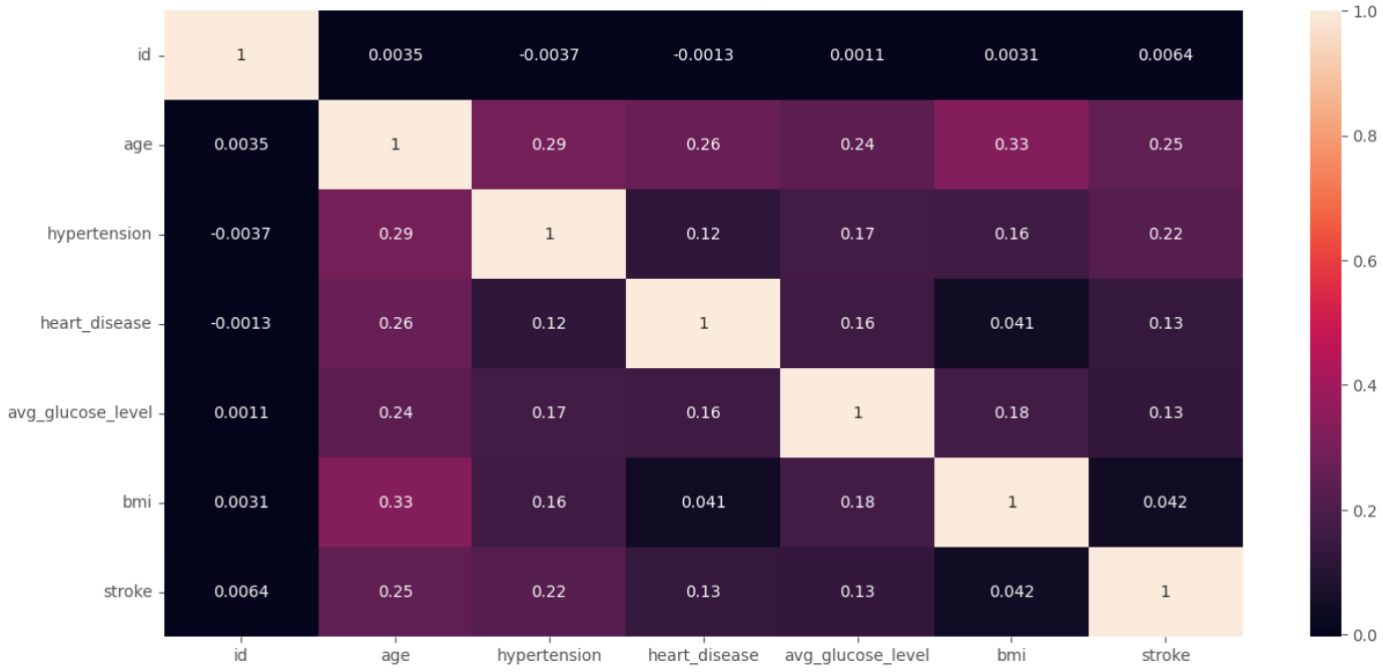


The data from the dataset is loaded for the pre-processing and cleaning of the data.

```
[2] df=pd.read_csv("/content/healthcare-dataset-stroke-data 1.csv")
    df.head()
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	1	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	1	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

The Correlation plot in the heatmap of the plotted to understand the relation between the different variables.



Confusion matrix for different Algorithms:

Algorithm	True Positive	False Negative	False Positive	True Negative
XGBOOST	1457	11	60	5
Random Forest	1464	4	65	0
SVM	1468	0	65	0
Logistic Regression	1467	1	65	0
MLP NN	1429	31	64	9

Comparison between Algorithms:

From Logistic Regression:

Testing Score: 95.694 %

Accuracy of the model: 96%

From MLP NN Classification:

Testing Score: 93.8034

Accuracy of the model: 94%

From SVM:

Testing Score: 95.759 %

Accuracy of the model: 96%

From XGBoost Classifier:

Testing Score: 95.368 %

Accuracy of the model: 95%

From AdaBoost Classifier:

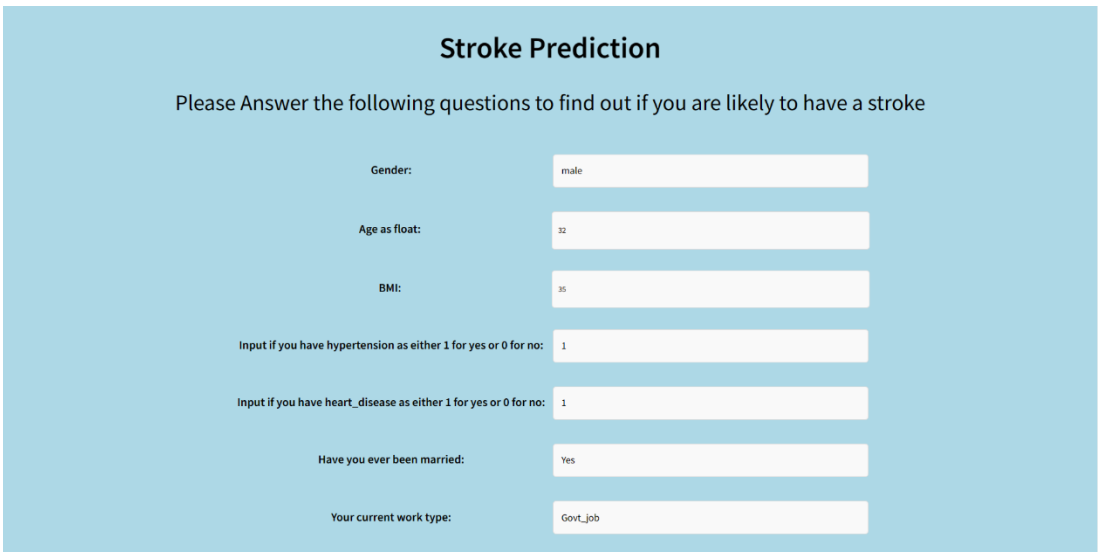
Accuracy of the model: 91 %

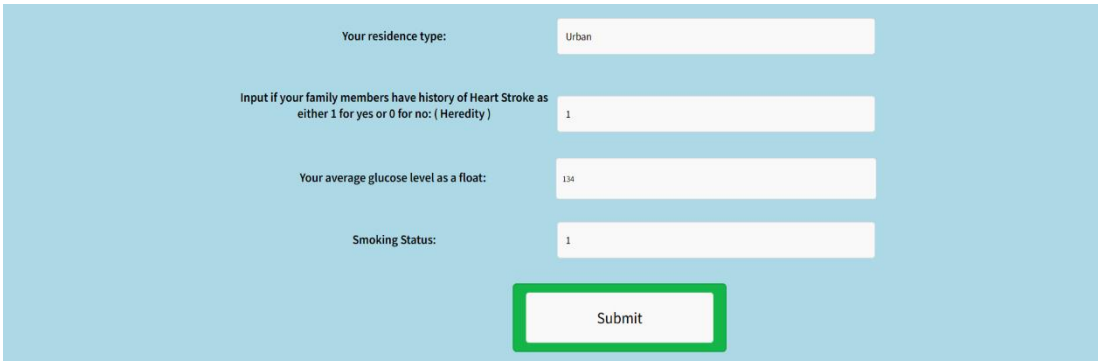
From Random Forest Classifier:

Testing Score: 95.499 %

Accuracy of the model: 95%.

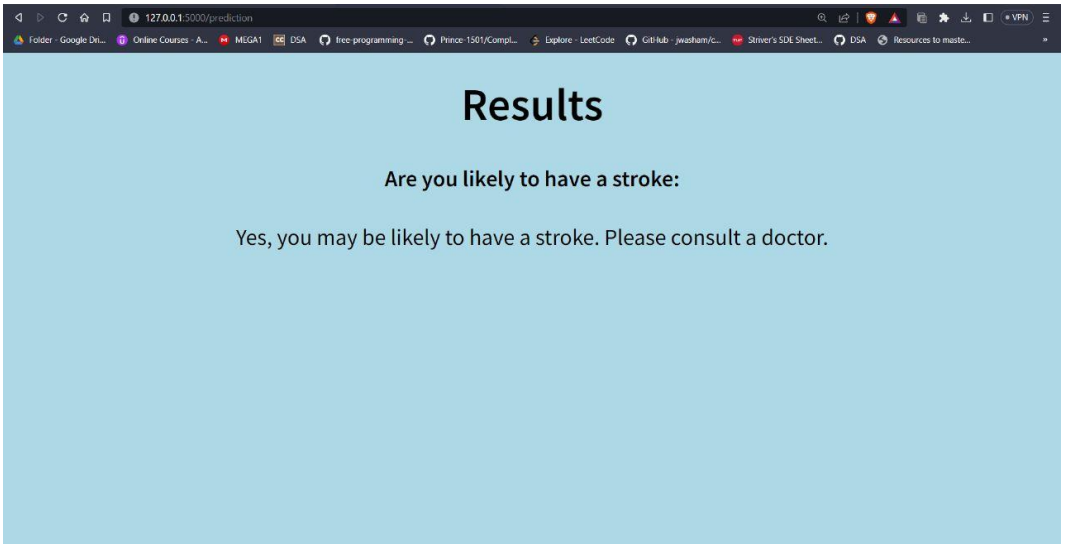
We select the SVM algorithm because of high accuracy. We use it in our code to predict the result weather the person has any chance of getting future stroke.



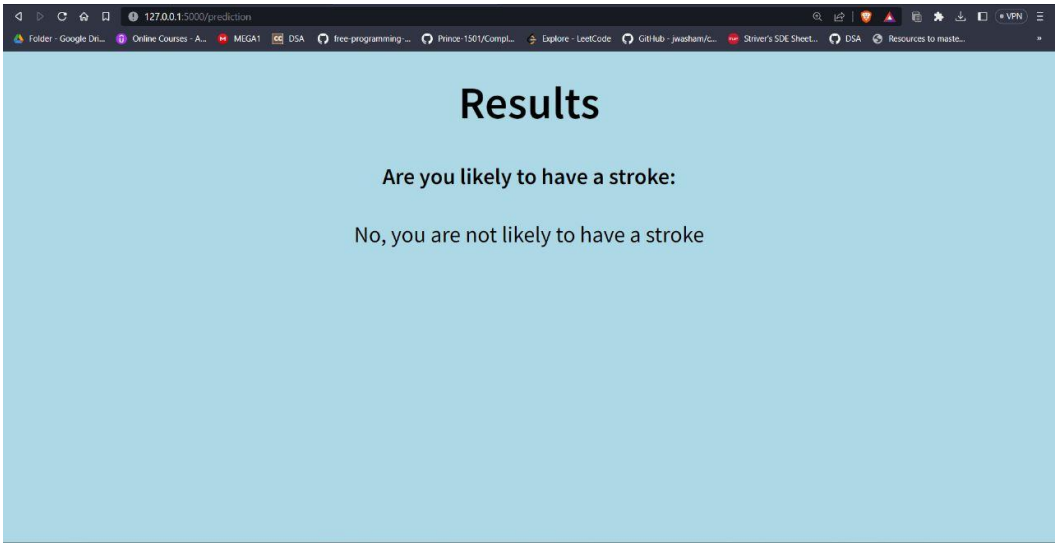


Results:

If the person has any chance of getting the of stroke it displays the below result to the user.



If the person doesn’t have any chance of getting the of stroke it displays the below result to the user.



Conclusion:

In conclusion, this project successfully achieved its objectives of continuously monitoring the patient and predicting if a patient is suffering from heart disease effectively, classifying the patient's risk level by implementing different data mining techniques, presenting a comparative study by analysing the performance of various machine learning algorithms, and creating a working model (website).

The project collected authentic and diverse data from the archives provided by Health Associations and used various learning algorithms such as SVM, Random Forest, XGBoost, AdaBoost, MLP NN, and logistic regression. These algorithms helped in analysing the performance of the system in correctly classifying if a patient is at a risk of stroke. The testing scores and accuracy rates of the models were high, indicating the effectiveness of machine learning in predicting heart disease.

The project provides valuable insights into the best machine learning algorithms for predicting heart disease, which can be used to guide future research in this area. The working model created in the project can be used by the public for early detection and intervention of heart disease, leading to improved patient outcomes.

In summary, this project contributes to improving the healthcare system by enabling constant monitoring and early detection and intervention for heart disease, ultimately leading to better patient outcomes.

References:

[1]. Kaur, Mandeep, Sachin R. Sakhare, Kirti Wanjale, and Farzana Akter. "Early stroke prediction methods for prevention of strokes." Behavioural Neurology 2022 (2022).

[2]. Fernandez-Lozano, Carlos, Pablo Hervella, Virginia Mato-Abad, Manuel Rodríguez-Yáñez, Sonia Suárez-Garaboa, Iria López-Dequidt, Ana Estany-Gestal et al. "Random forest-based prediction of stroke outcome." Scientific reports 11, no. 1 (2021): 1-12.

[3]. Arslan, Ahmet Kadir, Cemil Colak, and Mehmet Ediz Sarihan. "Different medical data mining approaches based prediction of ischemic stroke." Computer methods and programs in biomedicine 130 (2016): 87-92.

[4]. Winzeck, Stefan, Arsany Hakim, Richard McKinley, José AADSR Pinto, Victor Alves, Carlos Silva, Maxim Pisov et al. "ISLES 2016 and 2017-benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI." Frontiers in neurology 9 (2018): 679.

[5]. Liu, Tianyu, Wenhui Fan, and Cheng Wu. "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset." Artificial intelligence in medicine 101 (2019): 101723.

[6] "Diabetes and Stroke Prevention." [Online]. Available: https://www.stroke.org/en/about-stroke/stroke-risk-factors/diabetes-and-stroke-prevention. [Accessed: Apr. 12, 2023].

[7] D. K. D. D. S. K. D. K. D. A. N. A. R. A. D. L. Perera, M. P. Bandara, and N. Wickramasinghe, "Prediction of Stroke in Patients with Type 2 Diabetes Mellitus Using a Novel Risk Score," Journal of Diabetes Research, vol. 2020, p. 8858624, Dec. 2020, doi: 10.1155/2020/8858624.

[8] M. R. Islam, M. T. Islam, and A. Almogren, "Diabetes Monitoring: The Next Big Move for Wearable Industry," Circuit Digest, Aug. 2019. [Online]. Available: https://circuitdigest.com/article/diabetes-monitoring-the-next-big-move-for-wearable-industry. [Accessed: Apr. 12, 2023].

[9] S. Zhang, X. Wu, Y. Liu, X. Liu, J. Wang, and H. Lu, "Early Prediction of Acute Ischemic Stroke with Uncontrolled Hyperglycemia as an Initial Symptom," Journal of Stroke and Cerebrovascular Diseases, vol. 26, no. 2, pp. 324-331, Feb. 2017, doi: 10.1016/j.jstrokecerebrovasdis.2016.10.012.

Thank You!