DS&AI 2023
DATA SCIENCE AND ARTIFICIAL INTELLIGENCE CONFERENCE 2023
1ST - 3RD FEBRUARY 2023

Stroke prediction system using machine learning

Thalma Thandie

Sponsored by Artificial Intelligence Journal, Google TensorFlow

Background

- **A stroke** is a serious medical condition that occurs when the blood supply to the brain is interrupted, which can cause brain cells to die. Stroke is a leading cause of death and disability worldwide and early detection and treatment is critical for minimizing the risk of long-term disabilities and death.

- Traditionally, the diagnosis of stroke has been based on clinical symptoms, imaging studies, and laboratory tests. However, these methods have limitations, and there is a growing interest in developing more accurate and efficient methods for stroke prediction. One approach that has been gaining popularity in recent years is the use of machine learning (ML) techniques.

- **Machine learning** is a type of artificial intelligence that allows systems to learn from data and make predictions or decisions without being explicitly programmed. Machine learning algorithms can be used to analyze large amounts of data and to identify patterns and relationships that may not be apparent to human experts.

Background-*continued*

- There are various machine learning methods that can be used for stroke prediction, such as supervised learning, unsupervised learning, and deep learning.

- The use of machine learning techniques in stroke prediction has the potential to improve the accuracy and efficiency of stroke diagnosis, and to identify patients at high risk of stroke who may benefit from early intervention.

- It's important to note that, even though ML can help with the prediction of stroke, it's not a replacement for the clinical judgement of a healthcare professional. The results of the ML model should always be interpreted by a healthcare professional, who will make the final decision about the patient's treatment.

Problem

- It is difficult to identify individuals who are at high risk of experiencing a stroke, making it challenging to prevent or mitigate the occurrence and severity of strokes. This can lead to increased morbidity and mortality rates, as well as strain on healthcare systems.

Project Objectives

- Identify individuals who are at high risk of experiencing a stroke, so that preventative measures can be taken to reduce their risk.

- Reduce the overall burden of stroke on healthcare systems and society, by reducing the number of strokes and the severity of strokes that occur.

- Develop accurate, efficient, and non-invasive methods for stroke prediction, to minimize the burden on patients and healthcare providers.

Background Literature

- Existing works in the literature have investigated various aspects of stroke prediction. Jeena et al. provides a study of various risk factors to understand the probability of stroke. It used a regression-based approach to identify the relation between a factor and its corresponding impact on stroke. The risk factors identified in this work were divided into four groups — demographic, lifestyle, medical/clinical and functional.

- Similarly, Luk et al. studied 878 Chinese subjects to understand if age has an impact on stroke rehabilitation outcomes

- Hung et al. in compared deep learning models and machine learning models for stroke prediction from electronic medical claims database.

Background Literature

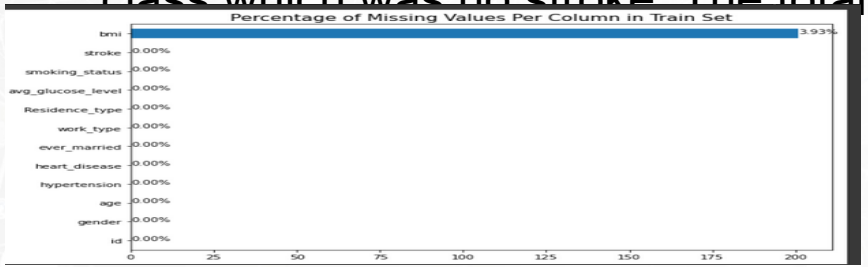
- The results from the various techniques are indicative of the fact that multiple factors can affect the results of any conducted study. These various factors include the way the data was collected, the selected features, the approach used in cleaning the data, imputation of missing values, randomness and standardization of the data will have an impact on the outcome of any study carried.

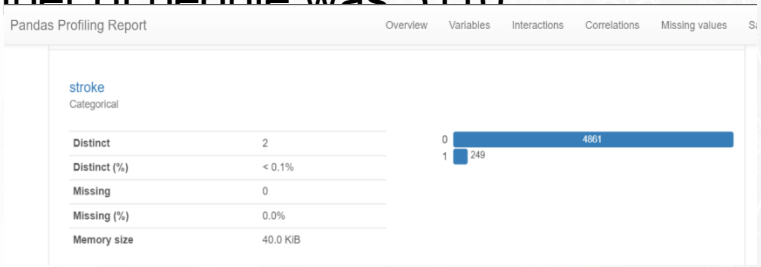
- Studies in related areas demonstrate that identifying the important features impacts the final performance of machine learning framework. It is important for us to identify the perfect combination of features, instead of using all the available features in the feature space. As indicated in redundant attributes and/or totally irrelevant attributes to a class should be identified and removed before the use of a classification algorithm

Methodology

1) **Data acquisition:** A dataset of individuals with information on their blood pressure, age, gender, heart disease, and smoking status was obtained from the source, in this case Kaggle.

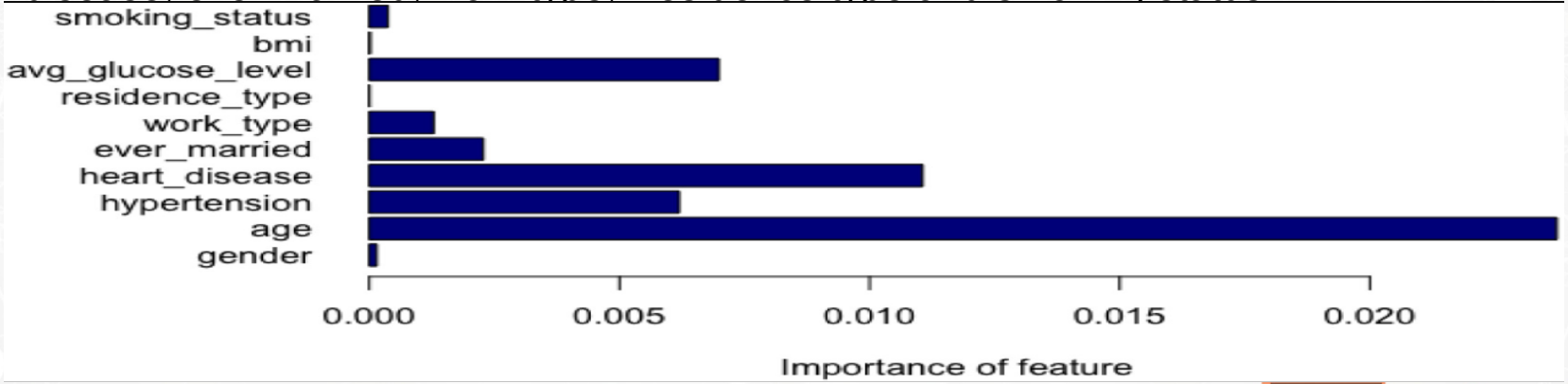
2) **Data cleaning and preprocessing:** In this section the data was Cleaned and preprocessed to ensure that it is suitable for use in machine learning. This involved removing missing values, converting categorical variables into numerical ones and since It has been observed that our target class has an imbalanced. Resampling was therefore done to down-sample the majority class which was no stroke. The total number of people was 5110





Methodology

3) **Feature selection:** In this step a set of relevant features was selected from the dataset will be used to train the machine learning model. The data set had a total of 12 variables which were id, gender, age, hypertension, heart disease, ever married, work type, Residence type, average glucose level, BMI, smoking status and stroke. The features selected to be used to train the model were: id, gender, age, hypertension, heart disease, ever married, work type, Residence type and smoking status.

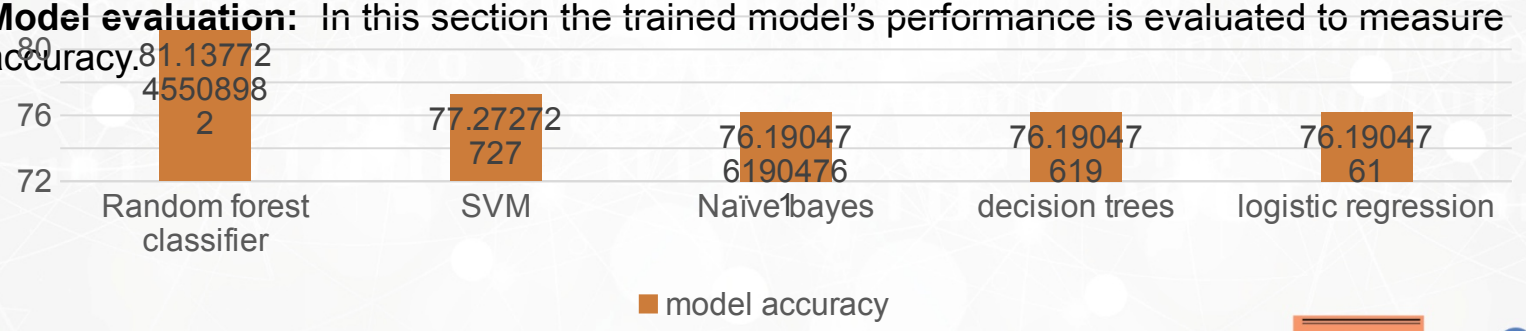


Methodology

- **4) Model selection:** Select a machine learning algorithm that is suitable for this task. This is a classification problem and therefore the classification machine learning algorithms were selected for use

- **5) Model training:** Train the selected machine learning model using the selected features and the cleaned and preprocessed dataset. The model was trained using different classification algorithms in python programming language.

6) Model evaluation: In this section the trained model's performance is evaluated to measure accuracy.



Methodology

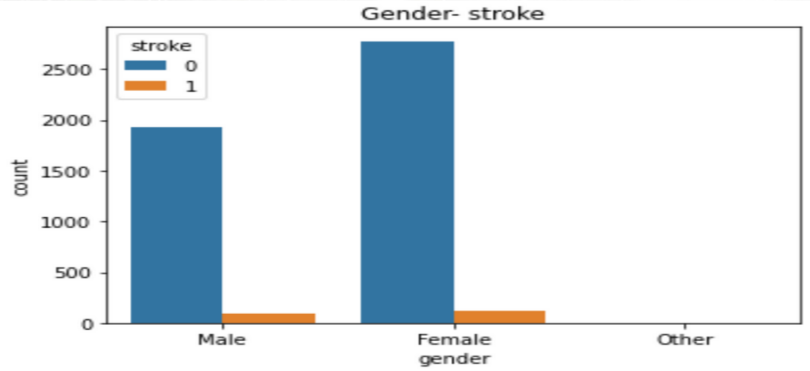
- It can be observed that the random forest classifier has the highest accuracy among all the models. This suggests that the random forest classifier is the best performing model for this particular dataset and task.

7) **Model Deployment:** Once the model's accuracy was satisfactory, it was deployed in a production environment using a combination of Streamlit and PyCharm. The model was first exported in a format that can be used in production such as '.pkl'. Then a Streamlit app was created to serve as an interface for the deployed model, and was configured to run locally. Users can access the Streamlit app and make predictions via a web.

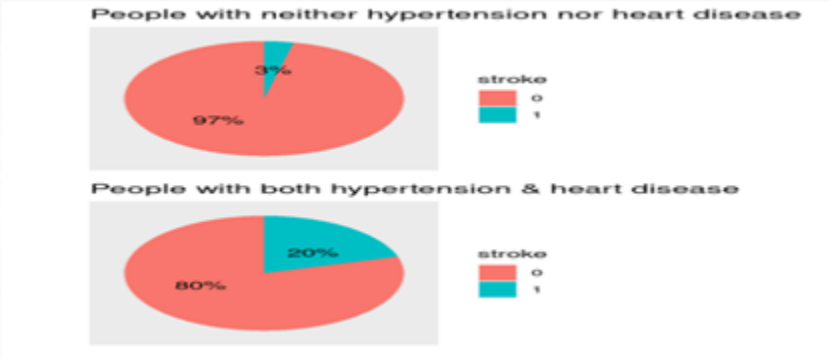
Results / Outputs

The study found that :-

- The number of women who experienced a stroke is higher than men. These findings highlight the need for further research to better understand the reasons behind this difference.

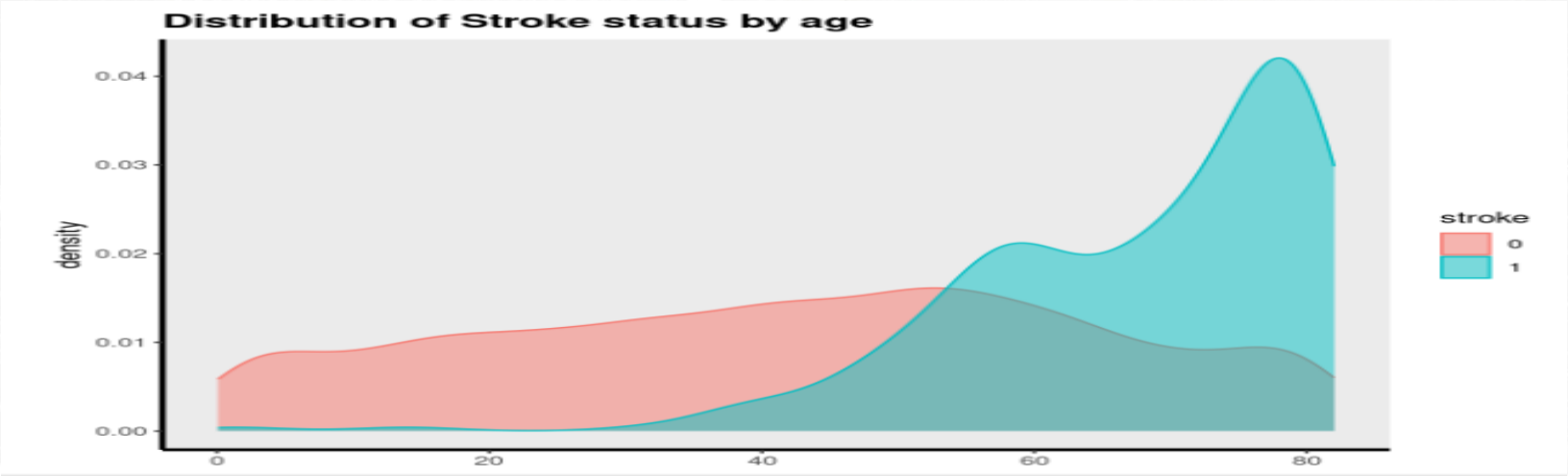


- As expected, having both Hypertension and Heart disease increases the chances of having stroke even more. About 20% of patients in the data with these two diseases have stroke.



Results / Outputs

Age is a significant risk factor for stroke. As people age, their risk of having a stroke increases. The dataset have shows that the incidence of stroke increases with age, with the highest rates occurring in those over the age of 65.



Discussion / Implications

- The results confirm the existing knowledge that age is a major risk factor for stroke. The results show that the incidence of stroke increases with age, with the highest rates occurring in those over the age of 65. This information could have important implications for healthcare, as it could help identify individuals at high risk for stroke, allowing for earlier intervention and prevention.

- Having both hypertension and heart disease increases the chances of having a stroke, with about 20% of patients in the dataset with these two diseases experiencing a stroke. This confirms the known relationship between hypertension and heart disease as risk factors for stroke and emphasizes the importance of controlling these diseases to reduce stroke risk.

Discussion / Implications

- The study found that the number of women who experienced a stroke is higher than men. This indicates that women may be at a greater risk of stroke than men.
- According to https://www.stroke.org.uk/what-is-stroke/are-you-at-risk-of-stroke/women-and-stroke
- Some aspects of women's lives can increase our risk of a stroke, like the contraceptive pill, pregnancy and having migraines. But for most women, taking care of your health and managing your risk factors will help you avoid a stroke. For example, if you are a younger woman who doesn't smoke and is physically active, your risk of stroke will probably be very low.

Conclusions

- In conclusion, the development of a machine learning model for stroke prediction is a challenging task that requires a combination of domain knowledge, data science expertise, and the ability to use advanced machine learning techniques. The accuracy of the model is a crucial factor in determining its clinical utility.

- The model developed in this study achieved an accuracy of 80% and demonstrated its ability to predict stroke in a reliable manner. However, there is always room for improvement and further research is needed to increase the accuracy of the model.

Future Work / Directions

I. Gather more data: One of the most effective ways to improve model performance is to gather more training data. This can be especially important in the medical field, where large datasets are often needed to accurately predict complex conditions like stroke.

II. Feature engineering: Another way to improve model performance is to carefully design new features that are better suited to the task at hand. This can involve using domain knowledge to extract new features from the raw data, or combining existing features in new ways.

III. Hyperparameter tuning: The performance of many machine learning algorithms is controlled by a set of hyperparameters. By carefully tuning these parameters, it is possible to improve model performance.

IV. Ensemble methods: One can use ensemble methods like bagging, boosting and stack ensemble to improve the performance of the model.

THANK YOU!