

Stroke Risk prediction using machine learning models

CBIO313: Data Mining and Machine Learning



Sondos ameen awad

211001732

Under the supervision of Dr. Mohammed Elsayeh

**Table of Contents**

[**1. Introduction** 2](#_Toc168063378)

[**1.1. Dataset Overview** 2](#_Toc168063379)

[**2. Methodology** 3](#_Toc168063380)

[**2.1. Importing Necessary Libraries** 3](#_Toc168063381)

[**2.2. Data Preprocessing** 3](#_Toc168063382)

[**2.3. Exploratory Data Analysis (EDA)** 4](#_Toc168063383)

[**2.4. Feature Engineering** 4](#_Toc168063384)

[**2.4.1. Label Encoding** 4](#_Toc168063385)

[**2.4.2. Feature Scaling** 4](#_Toc168063386)

[**2.5. Model Development** 4](#_Toc168063387)

[**2.6. Model Hyperparameter Tuning** 4](#_Toc168063388)

[**2.7. Model Deployment** 5](#_Toc168063389)

[**3. Results** 5](#_Toc168063390)

[**3.1. Model Development** 5](#_Toc168063391)

[**3.2. Model Hyperparameter Tuning** 6](#_Toc168063392)

[**3.3. Model Deployment** 6](#_Toc168063393)

[**4. Conclusion** 6](#_Toc168063394)

[**5. Limitations** 6](#_Toc168063395)

[**6. Recommendations** 7](#_Toc168063396)

[**7. References** 8](#_Toc168063397)

# **1. Introduction**

According to the World Stroke Organization, 13 million individuals experience a stroke each year, resulting in around 5.5 million fatalities (Dritsas & Trigka, 2022). Strokes are the largest cause of mortality and disability globally, affecting every aspect of life. A stroke is an acute neurological condition caused by an interruption in blood flow to the brain, which deprives brain cells of oxygen (Dritsas & Trigka, 2022).

A previous stroke or transient ischemic attack, heart diseases such as myocardial infarction and atrial fibrillation, hypertension, carotid stenosis from atherosclerosis, smoking, high cholesterol, diabetes, obesity, a sedentary lifestyle, alcohol consumption, blood clotting disorders, estrogen therapy, and the use of substances such as cocaine and amphetamines are all risk factors for stroke (Dritsas & Trigka, 2022). Stroke symptoms can vary and advance fast or slowly, including memory, attention, and speech problems, emotional concerns, loss of balance and walking abilities, loss of feeling on one side of the body, and trouble swallowing (Dritsas & Trigka, 2022).

While some people recover from a stroke, many continue to suffer long-term consequences, depending on the severity of the stroke. Advances in artificial intelligence (AI) and machine learning (ML) have become critical for early disease prediction in healthcare.

This project aims to develop and evaluate various machine learning (ML) models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, KNN, Naïve Bayes, XGBoost, AdaBoost, Gradient Boost, and Decision Tree, to create an accurate framework for predicting stroke occurrence over time.

## **1.1. Dataset Overview**

The dataset used in this project is available on google datasets. The dataset consisted of 43400 instances or patients along with 12 features or columns which were mostly nominal. The features were:

|  |  |
| --- | --- |
| ID | Refers to the patient’s ID. |
| Gender | Refers to the patient’s gender (Male, Female, others). |
| Age | Refers to the patient’s age in years. |
| Hypertension | Refers to whether the patient is hypertensive. |
| Heart diseases | Refers to whether the patient suffers from any heart diseases. |
| Ever married | Refers to the patient’s marital status. |
| Work type | Refers to the patient’s marital status. |
| Residence type | Refers to the participant’s residence status (urban, rural). |
| Average glucose level | Refers to the participant’s average glucose level. |
| BMI | Refers to the participant’s body mass index. |
| Smoking status | Refers to the participant's smoking status (never\_smoked, formerly\_smoked, smokes). |
| Stroke | Refers to the occurrence of stroke. |

# **2. Methodology**

## **2.1. Importing Necessary Libraries**

All necessary libraries for dataset reading, cleaning, scaling, model development and deployment were imported.

## **2.2. Data Preprocessing**

Data preprocessing is an essential phase in the success of any project, as it guarantees that the dataset is of sufficient quality for subsequent analysis. This process encompasses several key steps, including addressing missing or null values, managing duplicated entries, handling outliers, and performing data scaling and normalization. For this specific dataset, all null values and outliers were removed, no duplicated values were detected, and the "id" column was excluded from the dataset.

## **2.3. Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is a vital step for machine learning algorithms, as it allows us to understand the dataset, its features, and the distribution of values. Typically, this step involves creating various plots to explore the data. For this dataset, visualization techniques included count plots for stroke occurrences, gender distribution, smoking status, and heart diseases, as well as a pie chart for stroke occurrences across different genders and a box plot to identify outliers. The count plot for stroke occurrences revealed a significant data imbalance, which was addressed using oversampling techniques like SMOTE. The gender distribution plot identified seven entries labeled as "other," which were subsequently removed.

## **2.4. Feature Engineering**

The Feature Engineering process is used to convert or transform the data in our dataset into a format that is more suitable for further analysis to improve model accuracy and efficiency.

### **2.4.1. Label Encoding**

This step involves transforming categorical columns (gender, ever\_marries, work\_type, Residence\_type, smoking\_status) into numerical ones using label encoding to aid in subsequent analysis.

### **2.4.2. Feature Scaling**

This step involves scaling the features to ensure all instances have comparable scale or range.

## **2.5. Model Development**

The dataset was initially split into training and testing sets following the selection of the appropriate features and target variables. Nine machine learning models were then trained using the training set: Logistic Regression, Support Vector Machine, Random Forest, KNN, Naïve Bayes, XGBoost, AdaBoost, Gradient Boost, and Decision Tree. The accuracy scores of these models were calculated, and the model with the highest performance was chosen.

## **2.6. Model Hyperparameter Tuning**

A range of evaluation metrics, including accuracy, precision, recall, F1 score, and confusion matrix, were employed to assess the performance of the best-performing model, which in this case was the Decision Tree. A dictionary of hyperparameters (max\_depth, min\_samples\_leaf, min\_samples\_split, and criterion) was created, and the Decision Tree model was fine-tuned using a grid search cross-validation object fitted to the training set to identify the best estimator and determine its accuracy score. The same evaluation metrics were then used to re-evaluate the model's performance. Subsequently, all models were retrained using the most important features, and their accuracies were calculated.

## **2.7. Model Deployment**

To ensure the Python classifier model is readily available for use, we first save it to an external file using the pickle library. Then, we utilized the Streamlit library to develop a web application that forecasts the occurrence of strokes based on user-provided features. Initially, we load the pre-trained classifier with Pickle. We then create several text input fields for users to input values for various attributes. These inputs are used to construct a dataframe, where the feature names correspond to columns and the user inputs are the column values. Finally, the model predicts the occurrence of strokes, and the results are displayed to the user.

# **3. Results**

## **3.1. Model Development**

After training all nine models, the accuracy scores were as follows:

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| Logistic Regression | 75.39581340983955 % |
| Support Vector Machine (SVM) | 75.65614706194879 % |
| Random Forest | 94.81989161619381 % |
| K Nearest Neighbor (KNN) | 90.02231431303794 % |
| Naïve Bayes | 72.29837424290724 % |
| XGBoost | 87.58899160556795 % |
| AdaBoost | 76.89406014238656 % |
| Gradient Boost | 82.63202635214111 % |
| **Decision Tree** | **97.15758155350123 %** |

Since the Decision Tree classification model had the best accuracy, it was chosen for model deployment.

## **3.2. Model Hyperparameter Tuning**

Before fine-tuning the Decision Tree model, the classification report showed precision, recall, and F1-score values of 0.97, 0.98, and 0.97 for stroke samples, and 0.98, 0.97, and 0.97 for no-stroke samples. After hyperparameter tuning, the model's accuracy dropped from 97.15% to 76.62%, likely due to overfitting or issues with the dataset, as accuracy did not improve despite parameter adjustments. Consequently, all models were retrained using only the most important features. The Decision Tree model again demonstrated the best performance, achieving an accuracy of 97.205%, and was therefore selected for deployment.

## **3.3. Model Deployment**

The deployed model was successfully able to predict the occurrence of strokes in patients in accordance with the given variables.

# **4. Conclusion**

This project aimed to develop a machine learning model to predict stroke occurrences in patients based on information such as average glucose level and BMI, thereby helping to prevent misdiagnoses that could lead to severe consequences, including death. Our goal was to provide a tool that could enhance early detection accuracy, enabling more effective treatment plans and better patient outcomes.

Through a series of steps including data collection, preprocessing, feature engineering, model development, tuning, and deployment, we successfully created a model using the Decision Tree classifier. This model achieved an accuracy score of 97.2% in predicting stroke occurrences.

# **5. Limitations**

The project encountered several significant challenges throughout its development. One major issue was the imbalance in the dataset, which made it difficult to train the model effectively. Another limitation was the restricted number of features available, which constrained the model's ability to make accurate predictions. Additionally, the dataset contained a considerable number of outliers and null values, further complicating the analysis and necessitating extensive preprocessing efforts. Furthermore, the model demonstrated a high susceptibility to overfitting, making it challenging to achieve consistent and reliable predictions. These obstacles required careful attention and strategic adjustments to ensure the robustness and accuracy of the final model.

# **6. Recommendations**

We recommend addressing these issues by implementing several key strategies. Firstly, proper data cleaning is essential to handle the significant number of outliers and null values, ensuring the dataset's integrity and reliability. Secondly, balancing the data is crucial to mitigate the effects of dataset imbalance, thereby enhancing the model's ability to learn effectively from the data. Thirdly, employing a variety of models and incorporating regularization techniques would help avoid overfitting, promoting more robust and generalizable predictions. These measures collectively aim to improve the model's performance and ensure its reliability in predicting stroke occurrences.

# **7. References**

Dritsas, E., & Trigka, M. (2022). Stroke Risk Prediction with Machine Learning Techniques. *Sensors*, *22*(13), 4670. https://doi.org/10.3390/s22134670

Dutta, G. (n.d.). *Hyperparameter Tuning in Decision Trees*. Kaggle.com. https://www.kaggle.com/code/gauravduttakiit/hyperparameter-tuning-in-decision-trees

KLEPPEN, E. (2023, May 11). *How To Find Outliers Using Python [Step-by-Step Guide]*. Careerfoundry.com. https://careerfoundry.com/en/blog/data-analytics/how-to-find-outliers/

*Stroke Prediction*. (n.d.). Kaggle.com. Retrieved May 31, 2024, from https://www.kaggle.com/code/cthng123/stroke-prediction