BRAC UNIVERSITY

Inspiring Excellence

Project Report

CSE422 (Artificial Intelligence)

Project Title : Brain Stroke Prediction

Lab Section: 10

Faculty: MNH,MUSA

Group 14	
Kazi Abrab Hossain	21201496
Farzia Akhter	21241019

Table of Contents

Introduction

We are using insights based on data in our machine learning effort to tackle the crucial task of stroke prediction. Stroke is a major global health concern that is influenced by a multitude of risk factors. Through the use of machine learning, our goal is to create a prediction model that can recognize people who are more likely to have a stroke based on a variety of important characteristics. These characteristics include medical history variables like hypertension and heart disease, as well as demographic information like age and gender. Our study also heavily relies on lifestyle parameters such as BMI, smoking status, average glucose level, work type and marital status.

Our project's main goal is to develop a reliable predictive tool that can precisely predict a person's chance of having a stroke based on their particular risk factor combination. Using a large dataset, machine learning algorithms are trained in this manner to identify intricate linkages and patterns that could indicate a higher risk of stroke. By doing this, we hope to equip medical professionals with a resource that will help with targeted preventive care and early intervention for those who have been classified as high-risk.

Dataset Description

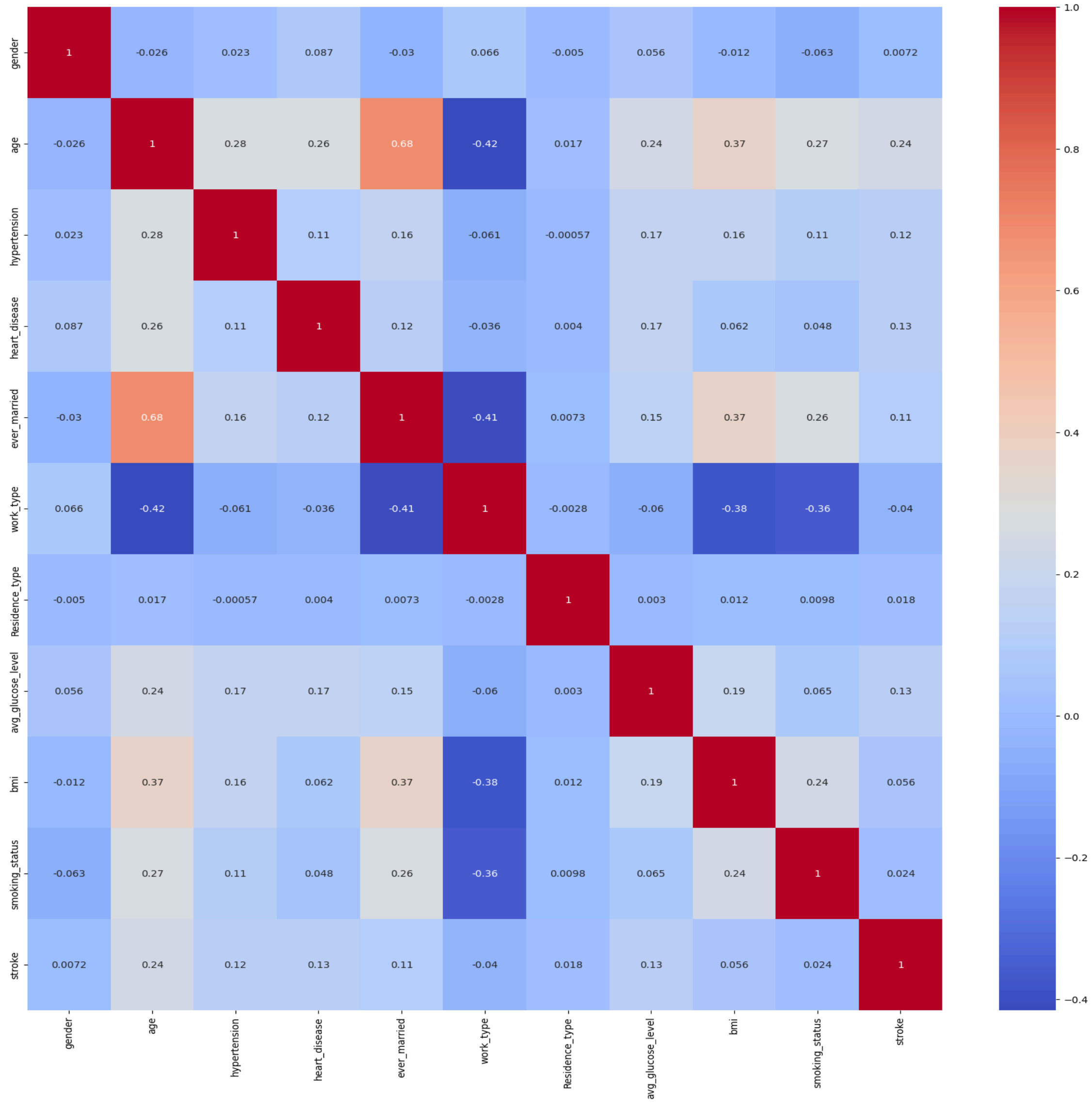
Link: Brain Stroke Dataset
Reference: *Brain Stroke Dataset*. (2022, August 4). Kaggle.
https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset?rvi=1

Dataset Description:
The dataset for the BigMart Sales has 11 features which are described below:

- Gender: Gender of the patient
- Age: Age of the Patient.
- Hypertension: If a patient has hypertension.
- Heart_disease: If a patient has heart disease.
- Ever_married: If a patient has ever married.
- Work_type: What type of work patients do.
- Residence_type: Where a patient lives.
- Avg_glucose_level: Glucose level of the patient.
- Bmi: BMI of the patient.
- Smoking_status: If patient smokes or not
- Stroke: If patient suffered from a stroke or not.

Our brain stroke dataset contains 4981 patients and 11 features. The project needed to have at least 12 features but unfortunately we did not find any usable dataset about Brain Stroke with 12 features that is why we used this dataset. Our dataset has the Classification problem because it has the qualitative value as target (1 means had brain stroke, 0 means not had brain stroke). Our dataset has both quantitative and categorical data included.

The Correlation of all features including both input and output features are shown by applying heatmap using the seaborn library:



This matrix shows the relationship by representing the correlation coefficient between each pair of variables in the dataset. Correlation coefficients measure the strength of a linear relationship between two variables which range from-1 to 1, where:

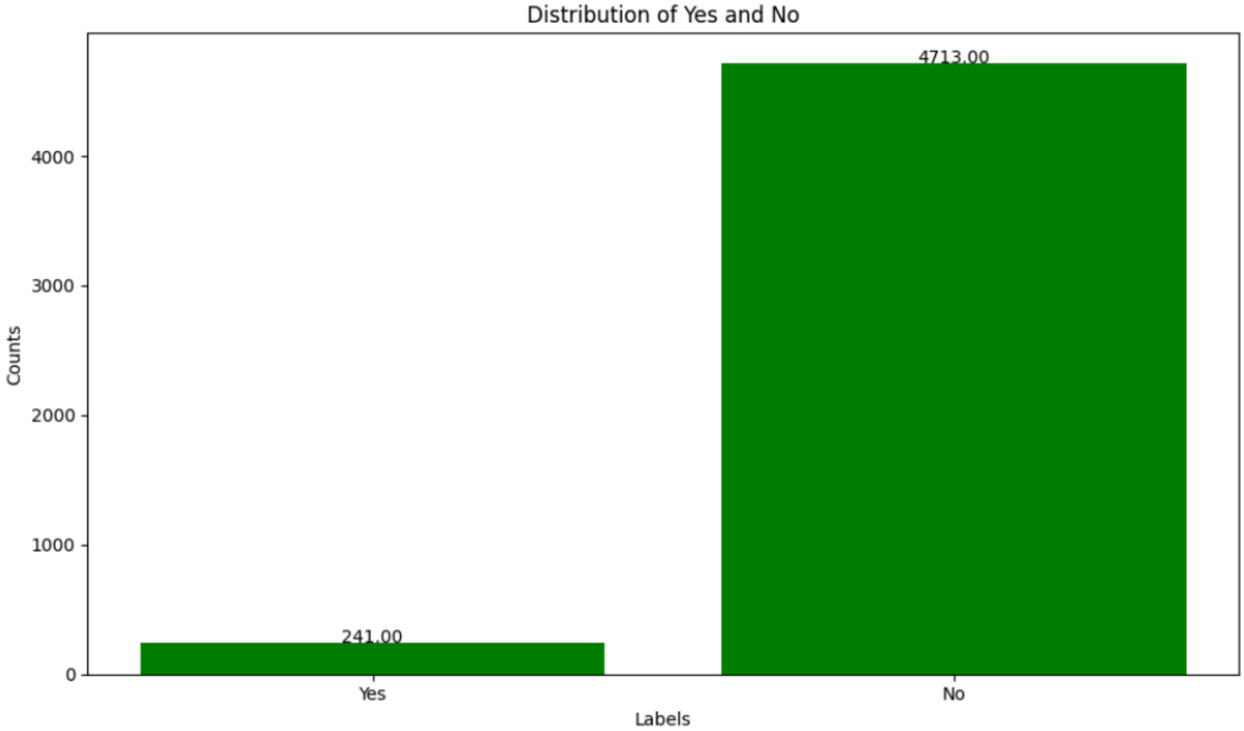
1: Perfect positive correlation

0: No correlation

-1: Perfect negative correlation

With the help of correlation coefficient, we can get the idea of the most important features of the dataset.

Unfortunately our dataset was highly imbalanced. The output feature only contained 241 patients who suffered from brain stroke and 4713 patients who did not.



Dataset pre- processing

Null values: The dataset we used had some null values in columns named Residence type, average glucose level and smoking status. The null values were not so much that is why we dropped the rows containing the null values because we had more than enough data to work with.

Categorical values: there are 5 object types of data . As the model can't use object type features so to solve this problem we did categorical encoding by label encoder.

Feature Scaling

Feature scaling is an essential step in preparing our dataset for machine learning models. It involves normalizing the range of features so that they are interpreted on the same scale by the models. As our dataset has many varying ranges,units etc.so to interpret those features on the same scale we did min max scaling.

Dataset Splitting

We divided the column features into x and y variables where y contains the dependent variable ‘stroke’ which will be predicted and x contains the remaining other features. Our dataset is splitted in a 0.30 ratio where 70% is the training set and 30% is the testing set. We also did stratified as the output feature is imbalance. We also used Oversampling for the output feature.

Model Training and Testing

We have used the following three classification models to train and test the dataset:

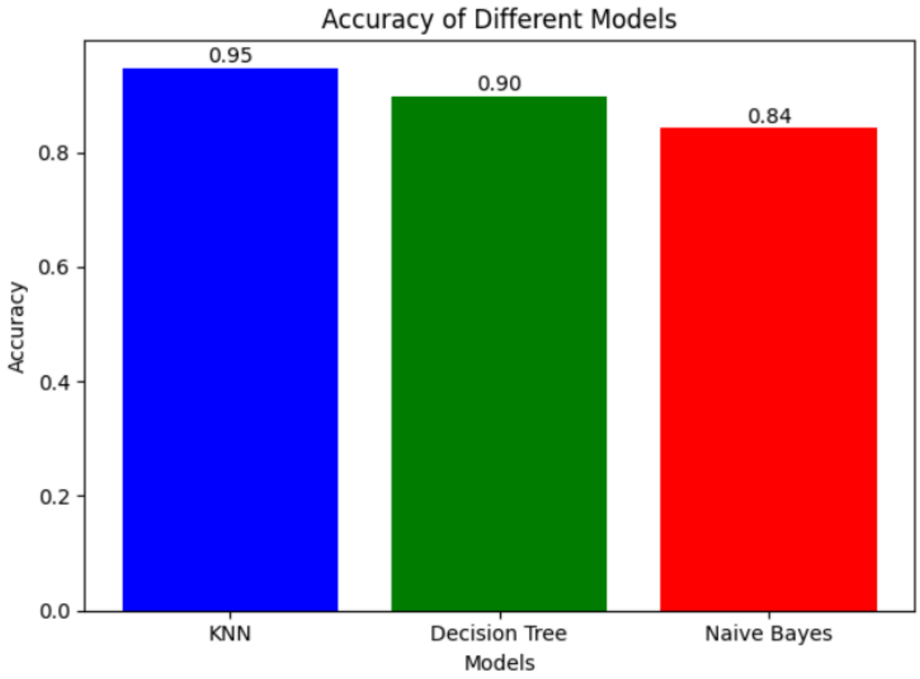
- **KNN:** KNN is a simple yet effective supervised learning algorithm used for both classification and regression tasks. It makes predictions based on the majority class (for classification) or the average value (for regression) of the k-nearest data points in the feature space. KNN relies heavily on a distance metric (such as Euclidean distance) to determine the similarity between data points.
- **Decision Tree:** Decision trees are versatile supervised learning models used for both classification and regression tasks. The model partitions the feature space into regions, assigning labels or values to each region based on training data. Decision trees use various criteria (like Gini impurity or

information gain) to select the best feature and threshold for splitting at each node. Decision trees can handle both numerical and categorical data.

- **Naive Bayes:** Naive Bayes is a probabilistic classifier based on Bayes' theorem with a strong (naive) assumption of feature independence. It calculates the probability of a label given the features using Bayes' theorem and assumes that all features are conditionally independent. Naive Bayes is particularly useful for text classification and spam filtering.

Model Selection / Comparison Analysis

The barchart is showing the accuracy comparison for 3 different models.



Precision, recall and F1 score comparison

1. **KNN Model**

   Precision: 0.2

   Recall: 0.027777777777777776

   F1 Score: 0.048780487804878044

2. **Decision Tree Model**

   Precision: 0.09090909090909091

   Recall: 0.125

   F1 Score: 0.10526315789473685

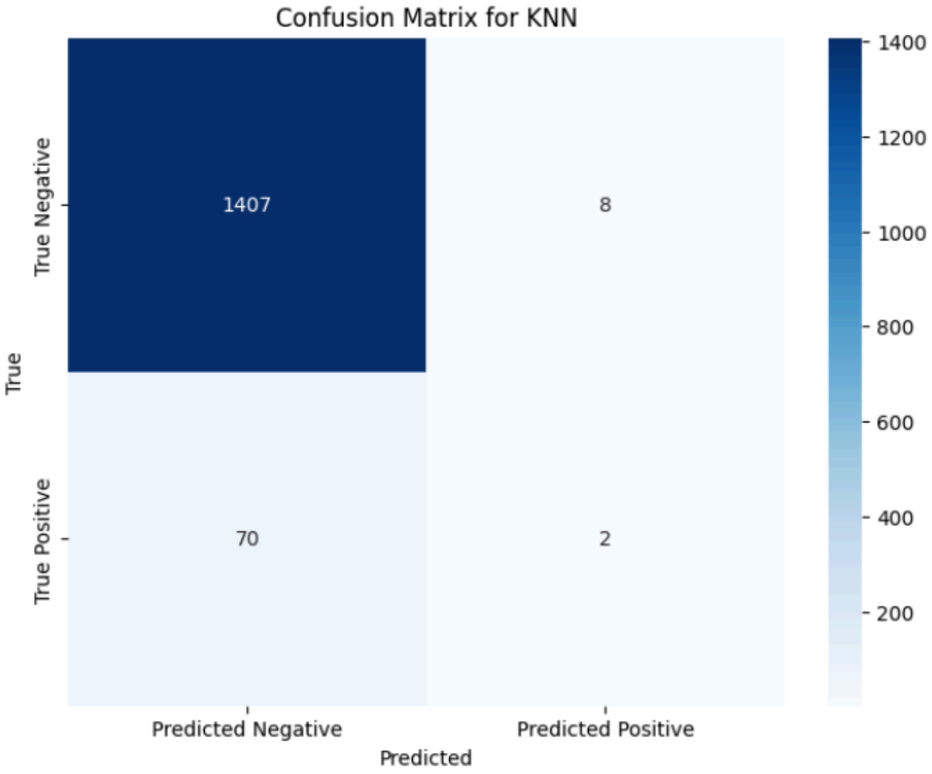
3. **Naive Bayes Model**

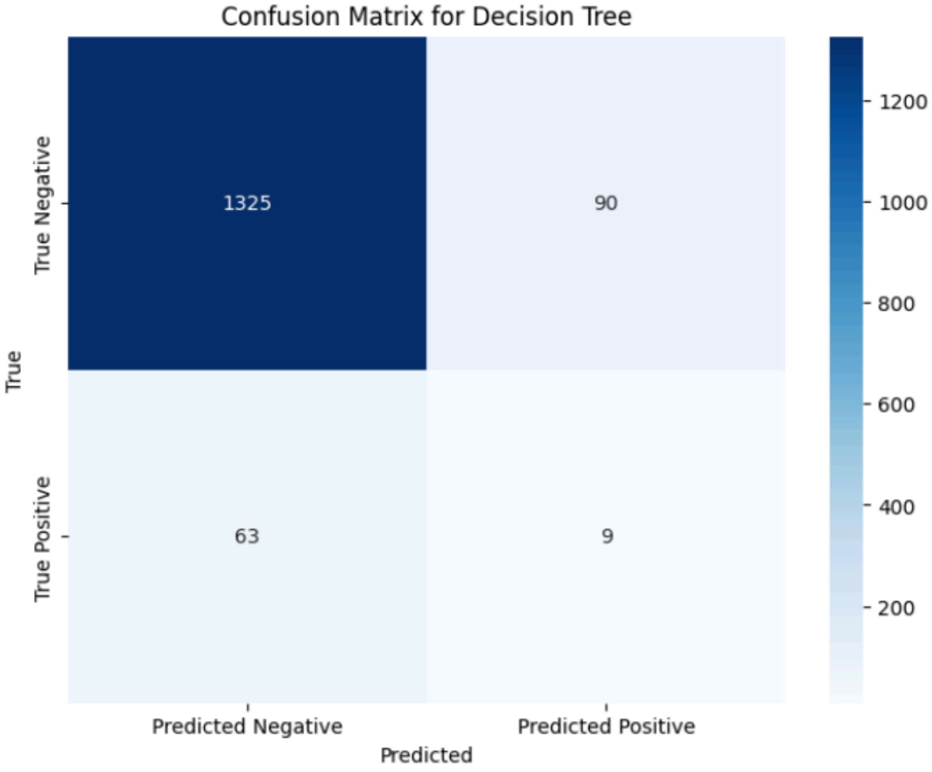
   Precision: 0.135678391959799

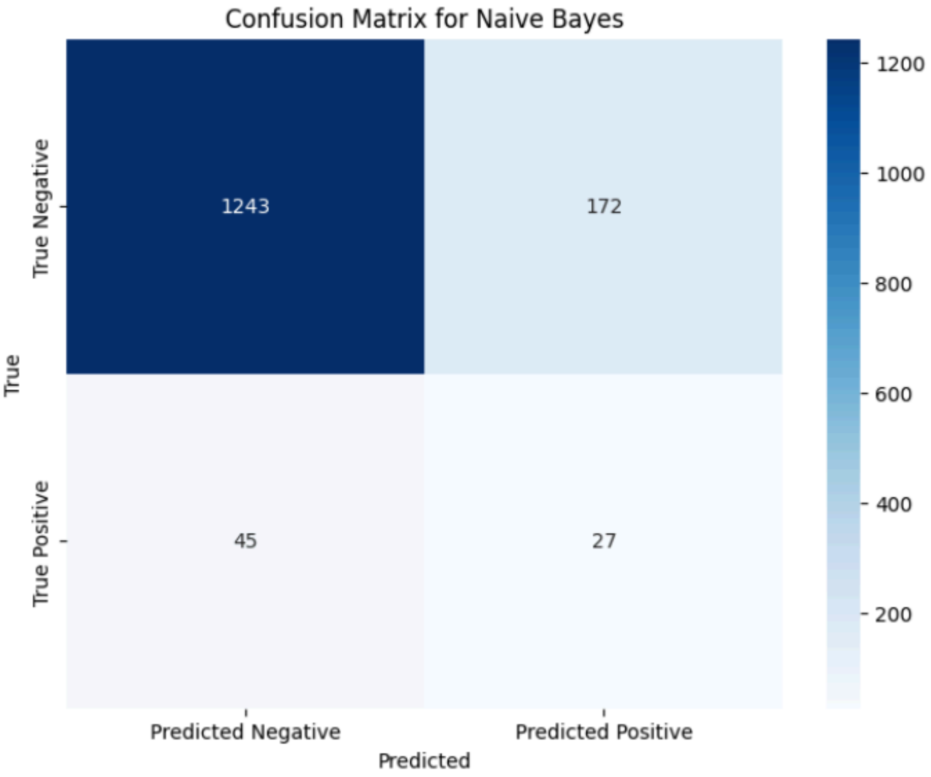
   Recall: 0.375

   F1 Score: 0.1992619926199262

Confusion Matrix For Every Model:







Conclusion

In this project we aimed to find the patients who suffered brain stroke. We used 3 classification models: KNN, Decision Tree and Naive bayes. Among them the KNN demonstrated highest accuracy followed by Decision Tree and Naive Bayes. This project can be used to detect brain stroke patients. But we also faced some issues. Because of the imbalanced dataset we had low precision and recall score. We also faced a slight issue at the time of encoding. Finally, This project helped us learn about Machine learning a lot