Predicting Stroke using Naive Bayes Classifier

Gaganpreet Kaur
Computer Science Department
Hood College
Frederick, MD
gk7@hood.edu

Abstract—Predicting one's chances of getting a stroke is both complex and time-consuming. It is an even more tedious task to predict with high accuracy for large numbers of individuals. This paper shares different data mining classification techniques used to predict the chance of getting a stroke. The classifier algorithms used to make predictions include K-Nearest Neighbors, Decision Tree, Random Forest, and Naive Bayes. This paper evaluates the results produced by these algorithms by looking at the confusion matrix and classification report. Further techniques can be explored, or existing models can be fine-tuned to achieve better results.

Index Terms—Algorithms, Data Mining, Data Visualization, Decision Tree, Classification, Confusion Matrix, Machine Learning, Logistic Regression, Naive Bayes, Stroke, K-nearest Neighbors

1 Introduction

Millions of people die every year from a stroke. Stroke is a life-threatening cardiovascular disease. People who survive a stroke may have to deal with a severe long-term disability because stroke affects a person's ability to walk, speak, or understand. According to the CDC, a center of disease control and prevention, there are more than 795,000 people who experience a stroke every year in the U.S. There are 15 million people who experience a stroke every year and 5.5 million people die of stroke every year [1]. Stroke is the second leading cause of death worldwide [3].

Stroke is caused by two main causes: a bursting of blood vessels (hemorrhagic stroke) or blocked arteries (ischemic stroke). There are various risk factors that increase the chance of a stroke including age, race, weight, exercise habits, genetics, hypertension, high cholesterol, diabetes, and previous history of stroke. It is an ongoing problem worldwide. However, stroke can be prevented, or one can reduce chances of getting a stroke by following a healthy lifestyle, visiting a doctor routinely, and monitoring one's health data such as blood pressure or cholesterol level. It will be beneficial to know one's chance of getting a stroke in advance. It will help save many lives or add additional years to one's life.

Data mining techniques can help predict a stroke from one's health data. The recent technological advances in processing power, storage, and speed have created more potential in analyzing, predicting, and uncovering hidden patterns from very large datasets than before. The data mining techniques help "provide the methodology and technology to transform these mounds of data into useful information for decision making [2]." It is now possible to process large datasets efficiently and analyze them using data mining techniques. Data mining has been used in many industries including healthcare, financial, and government. According to one article, an intelligent decision support system can be built with the use of data mining techniques and physicians' knowledge to predict stroke [4]. Data mining is one of many powerful approaches in solving real-world complex problems.

2 Methodology

Four implementation steps were taken as a part of the data mining technique. The first step in the process was to analyze the raw data and look for any data quality issues. The data was analyzed by looking for any missing or null values. The target predicate value of a stroke was also assessed for any bias. The second step was to prepare the data by doing the data scrubbing. The data visualizations were also produced using Python's plot library. The data visuals provided key information on how different categories have stroke ratios. For instance, one can see that the female had a higher stroke rate versus their male counterpart. The next step followed was to set up a model for classifier algorithms to use. Finally, the comparison of various algorithms was made to better understand the outcome of predictability. The decision tree was generated using the Decision Tree classifier algorithm to visually see the classification. The data was split into test and train datasets. Four different supervised machine learning algorithms were used independently to classify and predict stroke accuracy. These algorithms include the K-nearest Neighbors, Decision Tree, Random Forest, and Naive Bayes. The results were compared, and the data was further cleansed for better results.

3 Data Processing

Prediction algorithms require quality data to achieve a higher accuracy rate. The data quality can be measured by looking at its completeness, consistency, validity,

uniqueness, and accuracy. For example, inconsistent data or missing data requires extra work to prepare the data for processing. It will also impact the prediction accuracy rate. In this project, the patient data was taken from the Kaggle platform to analyze, model, and make predictions. The data contains patient health information like age, gender, and health history. The data consists of 5110 observations and 12 attributes. The 12 attributes and their description is listed below for reference. The partial data is provided in Fig. 1 to illustrate the values used for this project.

1. id: it is a unique identifier of a patient
2. gender: "Male", "Female" or "Other"
3. age: age of the patient
4. hypertension: 0 indicates that the patient doesn't have hypertension, and 1 indicates that the patient has hypertension
5. heart_disease: 0 indicates patient doesn't have any heart diseases, and 1 indicates that the patient has a heart disease
6. ever_married: "No" or "Yes"
7. work_type: "children", "Govt_job", "Never_worked", "Private" or "Self-employed"
8. Residence_type: "Rural" or "Urban"
9. avg_glucose_level: An average glucose level in blood
10. bmi: contains body mass index value of a patient
11. smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown".
12. stroke: 1 indicates the patient had a stroke or 0 if the patient never had a stroke.

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

Fig. 1: This figure contains only 5 records about patient health data.

The data from Fig. 1 was cleansed to prepare for data classification. For example, the gender value for “Male” and “Female” was converted into binary values of 1 and 0. There was a single value of “Other” which was dropped. Also, the attribute value of id was removed from processing. In a real-world application, an id attribute to identify stroke candidates. The attribute stroke is being used as a predicted value. The null value check revealed 201 null values (see Fig. 2) for body mass index (BMI). The null values of BMI were filled with the mean value of BMI.

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
```

Fig. 2: This figure provides a count of null values. Note the bmi attribute contains 201 null values.

The scrubbed data was further analyzed to better analyze and understand the quality of data. The scrubbed data is listed in Fig. 3.

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	avg_glucose_level	bmi	age
0	1	67.0	0	1	1	2	1	228.69	36.600000	1	1	2.706375	1.001234e+00	1.051434
1	0	61.0	0	0	1	3	0	202.21	28.893237	2	1	2.121559	1.384666e-15	0.786070
2	1	80.0	0	1	1	2	0	105.92	32.500000	2	1	-0.005028	4.685773e-01	1.626390
3	0	49.0	0	0	1	2	1	171.23	34.400000	3	1	1.437358	7.154182e-01	0.255342
4	0	79.0	1	0	1	3	0	174.12	24.000000	2	1	1.501184	-6.357112e-01	1.582163

Fig. 3: This figure shows the converted data with numeric values.

4 Data Analysis And Visualization

To better analyze the data, different matrices and graphs were produced during this project. The data visualization depicted previous stroke information of a patient over different categories. In the dataset, there are 2994 females and 2115 males in the data set. The gender numbers can be seen from Fig. 5 on the top-left graph. In the dataset, there are a total of 249 people who had a stroke and 276 people who had some form of heart disease. These numbers were obtained by using the value_counts method. Here is an example that shows stroke statistics for the ever_married attribute.

```
data[(data.stroke==1)].ever_married.value_counts()
```

```
1    220
0     29
```

Fig. 4. Stoke data by maritial status.

It is interesting to note that people who were ever married have a higher risk of getting a stroke. The same analysis can be drawn from people who have a history of hypertension and heart disease. These groups of people have higher chances of getting a stroke.

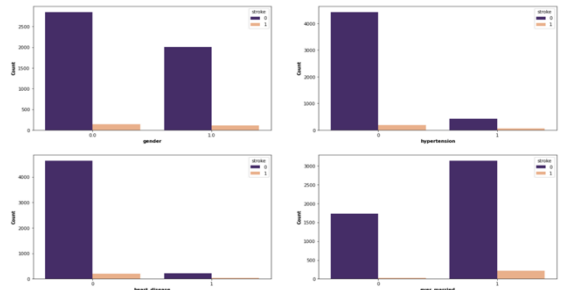

Fig. 5: Previous stroke information across different categories including gender, hypertension, heart disease and marital status

The following Fig. 6 shows strokes among males and females over different age ranges. Based on this figure, the graph depicts those females are more likely to get a stroke than males. Also, it shows that females are more likely to get a stroke earlier than males since there is no stroke history for males between the age of 30 and 40.

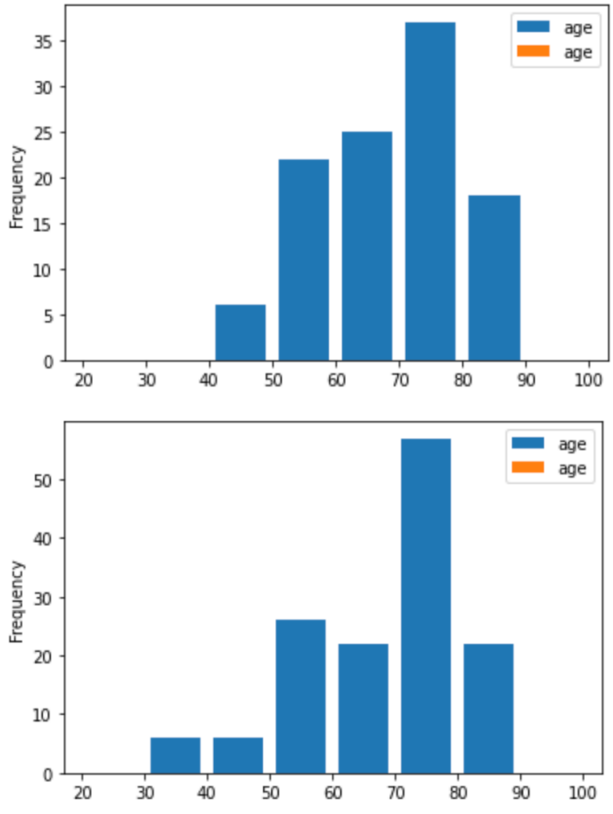

Fig. 6. Number of strokes among male (left) and females (right) over different age groups.

5 Evaluation And Results

The results were produced by using four different algorithms as discussed before. This paper evaluates the results by analyzing the confusion matrix and classification reports. The confusion matrix helps reveal how well the model performed. It shows four different measures including true positive, true negative, false positive, and false negative. Fig. 6 illustrates the confusion matrix layout. The false negative is called a type 2 error which should be looked at carefully.

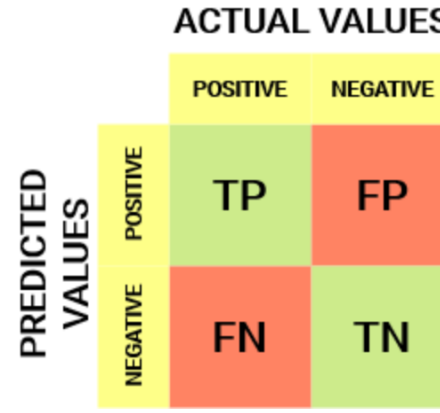

Fig. 7: Confusion Matrix shows actual and predictive values

The classification report shows the accuracy, precision, recall, and f1-score values. The results of these reports are listed below in Fig. 8 - Fig. 11. Based on the accuracy value from these reports, one can conclude that the Naive Bayes algorithm has the lowest accuracy of 85 percent, and all other algorithms have the highest accuracy. However, it is not the case here. The confusion matrix shows the opposite of that. The Naive Bayes report has the lowest false negative which means the actual value was positive and the algorithm predicted as negative. In addition, the precision number also supports the fact that Naive Bayes performed better than other algorithms.

K-nearest Neighbors:

```
Confusion Matrix:
1457    9
67      0

Classification Report:

              precision    recall  f1-score   support

           0       0.96      0.99      0.97      1466
           1       0.00      0.00      0.00        67

    accuracy                           0.95      1533
   macro avg       0.48      0.50      0.49      1533
weighted avg       0.91      0.95      0.93      1533
```

Fig. 8. Confusion matrix and classification report of K-nearest Neighbors.

```
Confusion Matrix:
1462    4
67      0

Classification Report:

              precision    recall  f1-score   support

           0       0.96      1.00      0.98      1466
           1       0.00      0.00      0.00        67

    accuracy                           0.95      1533
   macro avg       0.48      0.50      0.49      1533
weighted avg       0.91      0.95      0.93      1533
```

Fig. 9: Confusion matrix and classification report of Random Forest

```
Confusion Matrix:
1390    76
59      8

Classification Report:

              precision    recall  f1-score   support

           0       0.96      0.95      0.95      1466
           1       0.10      0.12      0.11        67

    accuracy                           0.91      1533
   macro avg       0.53      0.53      0.53      1533
weighted avg       0.92      0.91      0.92      1533
```

Fig. 10. Confusion matrix and classification report of Decision Tree

```
Confusion Matrix:
1275    191
33      34

Classification Report:

              precision    recall  f1-score   support

           0       0.97      0.87      0.92      1466
           1       0.15      0.51      0.23        67

    accuracy                           0.85      1533
   macro avg       0.56      0.69      0.58      1533
weighted avg       0.94      0.85      0.89      1533
```

Fig. 11. Confusion matrix and classification report of Naïve Bayes

6 Conclusion

Data mining techniques have been used in many areas to make predictions and informed decisions. These techniques work by learning the patterns and making intelligent decisions. In this project, the supervised algorithms were used to make inferences on getting a stroke. The data was analyzed, processed for better achieving accuracy. In this project, different supervised algorithms were used to make predictions. Based on the results, the Naive Bayes algorithm outperformed all other algorithms with an accuracy of 85 percent.

References

[1] "Stroke facts," *Centers for Disease Control and Prevention*, 25-May-2021. [Online]. Available: https://www.cdc.gov/stroke/facts.htm. [Accessed: 12-Dec-2021].

[2] HC, K. and G, T., 2021. *Data mining applications in healthcare.* [online] PubMed.

[3] Available at: <https://pubmed.ncbi.nlm.nih.gov/15869215/> [Accessed 12 Dec-2021].

[4] Fedesoriano, "Stroke prediction dataset," *Kaggle*, 26-Jan-2021. [Online]. Available:

[5] https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. [Accessed: 15-Dec-2021].

[6] L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash, R.

[7] Norouzi, and N. Toghianfar, "Prediction and control of stroke by Data Mining,"

[8] *International journal of preventive medicine*, May-2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3678226/. [Accessed: 15-Dec-2021].