# Predicting Stroke using Naive Bayes Classifier

Gaganpreet Kaur

Hood College, Frederick , Maryland

## 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.

## Data Description

The patient data was taken from the Kaggle platform to analyze, model, and make predictions. The data contains patient health information like age, gender, marital status, smoking habits, 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 |

***Figure 1.*** *This figure contains 5 records about patient health data.The last column, stroke, is the predicted output and other columns are predictors.*

## Data Mining Steps and Data Modeling

*Data was cleansed as part of preparing the data. Data contained missing values of BMI which were replaced with mean of BMI. Some attributes contained string values like gender, ever_married, etc. These were converted into binary values for processing and running the models. Different visualizations were also produced to analyze the data.*

*In order to model the data set, I split the data into training and testing to datasets. The train data set contains data with stroke value equals 1. I used four different data mining classification algorithms* ***K-nearest Neighbors, Decision Tree, Random Forest, and Naive Bayes*** *for making the prediction.*

## Results

***1) 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
```

***2) Random Forest***

```
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
```

***3) Decision Tree***

```
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
```

***4) Naive Bayes***

```
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
```

*Naive Bayes outperformed all other algorithms since it has the highest precision and the confusion matrix reveals lowest type two errors. NB has 85 percent accuracy of predicting a stroke*

## Future Works

*Further tuning can be done to achieve better accuracy. It may be combination more train data set, combining algorithms, and doing additional scrubbing.*

*Stroke is 2nd leading cause of death worldwide. Data mining techniques can help make intelligent decisions and eventually save many lives. The application of data mining in other areas can also help with making predictions medicine effectiveness, role of genetics in birth defects, and in predicting heart attacks.*

*For more detail about this project and source code, please visit https://github.com/KellyK81/data-mining*

## Contact

**Gaganpreet Kaur**

**Hood College, Department of Computer Science**

**Frederick , MD**

**gk7@hood.edu**

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

[3] Fedesoriano, "Stroke prediction dataset," Kaggle, 26-Jan-2021. [Online]. Available: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. [Accessed: 15-Dec-2021].

[4] L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi, and N. Toghianfar, "Prediction and control of stroke by Data Mining," International journal of preventive medicine, May-2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3678226/. [Accessed: 15-Dec-2021].