# Applying Supervised Machine Learning in Stroke Dataset

Deepanshu Bhatt | [deepanshu.2006bhatt@gmail.com](mailto:deepanshu.2006bhatt@gmail.com) | [Linkedin](https://www.linkedin.com/in/deepanshu-bhatt-05295b359/) | [GitHub](https://github.com/dbosmrt) | [X](https://x.com/dbosmrt) |

## Abstract

Stroke is one of the major causes of death around the world. For this supervised machine learning project, Healthcare stroke prediction dataset was taken from Kaggle dataset, which contains 11 clinical features. The dataset was loaded and manipulated using python libraries like pandas and NumPy. Initial EDA (Exploratory data analysis) was conducted, following with outlier detection and removal, finally ending data preprocessing with Encoding and Transformation. SMOTE (Single Minority Over Sampling Technique) was used to deal with the imbalanced target values. Data visualization was done using Matplotlib, and Seaborn. The data was evaluated using various classification model such as Logistic Regression, Decision Tree Classifier, Random Forest Algorithm, Gradient Boost Classifier. The accuracy of the model was observed, and the project ended with plotting confusion matrix, ROC-AUC graphs, and Evaluating Precision, Recall and F1. The dataset was quite tricky to handle and highly imbalanced.

## Introduction

### Stroke (Cerebrovascular Accident)

Stroke refers to a medical condition where poor flow of blood causes cell death in the brain. Signs and symptoms observed in case of stroke can be inability to move or feel on side, dizziness loss of vision on either one side or dizziness [1].

There are two ways or types in which stroke may occur to individual. One is Ischemic which occurs due to lack of blood flow and the other is haemorrhage which is due to bleeding [1]. The diagnosis of stroke can be done by general physical examination which includes taking a medical history of patient, accompanied with several imaging techniques (MRI, CT-Scan) and a neurological examination [2]. Imaging data can also be used for building and training deep learning modes i.e. CNN (Convolutional Neural Networks). The underlying cause of stroke is seen to be chronic hypertension, high cholesterol, Diabetes mellitus, surgeries, diet and work life.

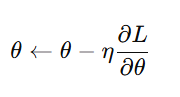
### Purpose of this Project

The efforts being putted in building this project is, with the intention of building a prediction model, which can predict the whether the person will have stroke or not. The projects only deal with the data preprocessing, training various classification models, testing these models, and getting performance matrices. The deployment of the project, is not the scope of this project, but in future may be developed.

### Supervised Machine Learning

Supervised Machine learning (SML), is a branch of machine learning which use labelled dataset, for the construction of predictive models. Here each input is associated with the known output or target variable [3]. Primary goal of a SML is to learn Mapping function **f :X→Y**, to predict accurate outputs for previously unseen inputs. This is achieved by identifying the statistical relationships and patterns between the dependent and independent variables.

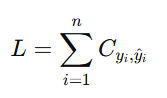
It is widely applied in domains such as medical diagnostics, fraud detection, natural language processing, and image recognition. It is widely applied in domains such as medical diagnostics, fraud detection, natural language processing, and image recognition, due to its ability to generalize from historical data for accurate predictions.

****Supervised machine learning workflow involves four main stages, follows as data collection, preprocessing, model training and evaluation. The problems can be either classified as Classification or Regression. Mathematically speaking SML learning entails, θ optimization to minimize a loss function L(y, y^​) using technique like gradient descent:

Here, neu is the learning rate.

Ther performance score is assessed on independent test set using metrics such as accuracy, F1 score, precision, recall and AUC-ROC curve.

### Cost-sensitive learning

It’s a machine learning approach incorporating varying costs of different error into the training process, instead of treating all misclassifications equally. In cases of medical diagnosis or fraud detection, it assigns higher penalties to more critical errors like False negatives. This method can be achieved by modification of loss function, re-weighting class, or using sampling strategies. It can be implemented by modifying the loss function with cost-matrix [4].

Here “j” represents the cost of the predicted class and “i” represents the true class. Cost sensitive method improves real world decision making where error consequences are unequal. enabling models to prioritize minimizing high-impact mistakes over overall accuracy, thereby aligning predictions with domain-specific risk and resource considerations.

## Methodology

### Dataset

We are using Stroke Dataset, which is real health data, specifically Electronic Health Records (EHR) of 5110 individuals, in Bangladesh having 11 medical attributes. This dataset originates from McKinsey & Company healthcare analytics dataset, which was initially used in a public analytics hackathon on several platforms.

### Performance Measurement

For training extremely imbalanced data, Using accuracy as an appropriate measure of performance. We use other metrics such as Confusion matrix, Precision , Recall and F1 Score, for evaluation of learning algorithm in imbalanced data. Accuracy cannot be considered alone, because a trivial classifier that predicts all cases as majority class can, still achieve high accuracy rate. This happens because the aim to reduce the overall error rate rather paying attention to the positive or minority class [5]. This requires other evaluation metrices mentioned above to get a more accurate understanding of our model prediction.

### Performance Result and Comparison

The data was trained and evaluated in 4 different classification models (Decision Tree classifiers, Logistics Regression, Random Forest Algorithm and Gradient Boost Classifier). The accuracy, Confusion Matrix, Classification Report (Precision, Recall, F1) were observed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Accuracy (%)** | **Random Forest** | **Decision Tree** | **Logistic Reg..** | **GB Classifier** |
| **Stroke Data** | 76.766 | 76.663 | 74.821 | 77.482 |

### Confusion Matrices

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive Class | Predicted Negative Class |
| Actual Positive Class | TP (30) | FP(22) |
| Actual Negative Class | FN (113) | TN (812) |

Table 1 Confusion Matrix with Logistic Regression

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive Class | Predicted Negative Class |
| Actual Positive Class | TP (43) | FP(9) |
| Actual Negative Class | FN (422) | TN (503) |

Table 2 Confusion Matrix with Decision Tree Classifier

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive Class | Predicted Negative Class |
| Actual Positive Class | TP (43) | FP(9) |
| Actual Negative Class | FN (422) | TN (503) |

Table 3 Confusion Matrix with Random Forest Classifier

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive Class | Predicted Negative Class |
| Actual Positive Class | TP (43) | FP(9) |
| Actual Negative Class | FN (426) | TN (499) |

Table 4 Confusion Matrix with Gradient Boosting ClassifierAUC – ROC Curve

### AUC- ROC Curves for different models:

Figure 2 Decision Tree Classifier

Figure 1 Logistic Regression

### Classification Report

Figure 3 Gradient Boosting Classifier

Figure 4 Random Forest Algorithm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Precision** | **Recall** | **F1 Score** | **Macro Recall** |
| **Decision Tree Classifier** | 0 | 0.98 | 0.54 | 0.70 | 0.69 |
| 1 | 0.09 | 0.83 | 0.17 |
|  | | | | | |
| **Random forest Algorithm** | 0 | 0.98 | 0.54 | 0.70 | 0.73 |
| 1 | 0.09 | 0.83 | 0.17 |
|  | | | | | |
| **Logistic Regression** | 0 | 0.97 | 0.88 | 0.92 | 0.69 |
| 1 | 0.21 | 0.58 | 0.31 |
|  | | | | | |
| **Gradient Boosting Classifier** | 0 | 0.98 | 0.54 | 0.70 | 0.69 |
| 1 | 0.09 | 0.83 | 0.17 |

Table 5 Classification Report on Different Models

## Inference and Conclusion

The dataset is highly imbalanced with non- stroke output being majority class. This can be easily seen in the confusion matrix.

### Accuracy

All of the classification models used shows similar accuracy scores (74-77%), with Gradient Boosting Classifier showing highest accuracy (77.48%), followed by Random Forest Algorithm (76.77%). But as the dataset is highly imbalanced the accuracy alone cannot be considered as a reliable indicator of performance.

### AUC-ROC

All of the ROC curve plotted for each algorithm shows AUC of 0.80 which is a strong indicator that the models can distinguish between stroke and non-stroke class.

### Classification Report

For precision, recall and F1 score we would mostly focus on Recall of Stroke which is 0.80, because in case of medical context missing stroke case (false negative), is considered way worse than False positives.

### Conclusion

Out of all classification models used, Gradient Boosting Classifier gives the best performance under cost sensitive objective, which accuracy of 77.48%, High recall of 0.83 and AUC of 0.80.

## References

[1]. National Heart, Lung, and Blood Institute, “What Is a Stroke?”, U.S. Department of Health & Human Services, archived by the Wayback Machine, Feb. 18, 2015. [Online]

[2]. Michael D Hill, Diagnostic Biomarkers for Stroke: A Stroke Neurologist’s Perspective, *Clinical Chemistry*, Volume 51, Issue 11, 1 November 2005, Pages 2001–2002.

[3].  Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) *Foundations of Machine Learning*, The MIT Press ISBN 9780262018258.

[4]. C. X. Ling and V. S. Sheng, “Cost-sensitive learning and the class imbalance problem,” in *Encyclopedia of Machine Learning*, C. Sammut, Ed. Springer, 2008.

[5]. C. Chen, A. Liaw, and L. Breiman, “Using Random Forest to Learn Imbalanced Data,” *Tech. Rep.* 666, Department of Statistics, University of California, Berkeley, CA, USA, July 2004.