MACHINE LEARNING : THE ULTIMATE GUIDE TO IMBALANCED DATA

D.Jeyanthi[1], Dr.A.Shanthasheela[2],

[1]Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodaikanal, Tamil Nadu.

[2]Assistant Professor, Department of Computer Science, M.V.Muthiah Governement Arts College(W), Dindigul, Tamil Nadu.

Abstract - In machine learning applications, data is a crucial element that must be balanced in order to extract features. Thus, creating ideal training data is the main machine learning challenge. In order to evaluate the effectiveness of machine learning algorithms like the k-nearest classifier, XG Boost classifier, CatBoost classifier, DecisionTree classifie, and GradientBoost classifier, this study experiments with various stroke data samples. XG Boost outperformed other algorithms with datasets that were sampled differently,achieving accuracy levels above 91 percent.

Keywords - *Machine Learning, K-Nearest Neighbor classifier, XG Boost Classifier, CatBoost classifier, DecisionTree classifier and GradientBoost classifier.*

1. INTRODUCTION

A stroke is a life-threatening medical condition brought on by a blood vessel blockage or leak that causes the brain to malfunction. An ischemic stroke is brought on by a blockage in a blood vessel,while a hemorrhagic stroke is brought on by an artery rupture. Stroke prevention is vital because it is the second leading cause of death and long-term disability after heart disease. In India, a stroke claims a life every forty seconds. Against a population of one million,an average of 194 to 215 new cases have been reported. The timely prediction will aid in appropriate care and preserve human life. Machine learning is essential to the medical field because it can identify the precise location of a disease, predict it before it manifest, and assist in treatment.

In this paper, top machine learning algorithms are compared using various sampled dataset attributes, and their respective performance are analyzed using evaluation metrics.

2. LITERATURE REVIEW:

The effectiveness of machine learning(ML) in predicting stroke disease has been examined in multiple studies. Typically, stroke prediction uses both text and image data.

Six machine learning algorithms were assessed by [2](Sailasya & Aruna Kumari, n.d.), including Naive Bayes classification(NBC), K-Nearest Neighbor classification(KNN), Support Vector Machine (SVM), Decision tree Classification (DTC), Naive Regression (LR) and Random Forest Classification (RFC). Of these, Naive Bayes Classification (82 percent) yielded the best accuracy. The web application was created in order to receive input. The data set was imbalanced, so the authors under-sampled it into 249 rows from 5110 rows, which is the minimum amount of data needed to run the machine learning algorihtms.

[3] (Emon et al., 2020) suggested a weighted voting classifier to predict the risk of stroke, and various machine learning classifier were compared with it. The weighted voting

classifier(WVC) was defined using the seven major attributes-hypertension, heart disease, age, BMI, glucose level, previous stroke status, and smoking status-among the dataset's twelve attributes, The study came to the conclusion that WVC is the most effective classifier available.

[4](Gazi Üniversitesi et al., n.d.) categorized the kind of stroke by applying eight distinct machine learning algorithms. For this diagnosis, the CT scan image data was used. Grey Level Co-occurrence Matrix(GLCM) used to exteact the features from the image. When compared to KNN, SVM, NBC, DT, SGD, LR, and Deep learning (DL), Random Forest classification yielded the best accuracy (95.97%). Using 10-fold cross-validation, the performances were examined.

[5](Teoh, 2018) adhered to regularization terms in the standard cross entropy loss-function to prevent inaccurate predictions resulting from unbalanced data in the recurrent Neural Network model. Eight thousand electronic health records of stroke victims were used for this purose. The Receiver Operating Characteristic was used to assess the model's performance.

3. FEATURE DESCRIPTION AND ANALYSIS OF INFLUENCING FACTOR:

This study compares several machine learning algorithms using various attributesof a data set: K-Nearest Neighbor, XG Boost Classifier, CatBoost classifier, DecisionTree calssifier, and GradientBoost classifier. The data set utilized in this study obtained from Kaggle and is not balance. Of the 5110 patients, only 249 have a positive stroke test result. This will cause errors in the prediction of strokes. Thus, pre processing is required to ensure that data is ready for analysis.

The importance and influence of 12 attributes in predicting stroke disease has been discribed below:

- ✓ Identification number of patient doesn't influence in the stroke prediction. So that can be removed from the stroke prediction algorithm.
- ✓ Similarly, biological factor gender influences in some way to predict the stroke disease.
- ✓ In this data set, the age attribute has some outliers. The average age of stroke positive rate is 67.
- ✓ Hypertension is one of the most influence factor in stroke disease prediction.
- ✓ Likewise, Heart disease is a main factor of stroke prediction.
- ✓ Maritial staus of the patient has high influence on stroke data.
- ✓ Work type of the patient- private, government, self employment plays a vital role in stroke prediction. Here, private employees have high stroke rate than other employees.
- ✓ Residential type is an influential factor of the stroke data.Here, urban people have high stroke rate than rural people.
- ✓ Average glucose level is major agent in stroke identification.
- ✓ Body-Mass-Index mainly influenced in stroke prediction. In this dataset, 201 rows are not available.
- ✓ Smoking is medium factor in stroke prediction.

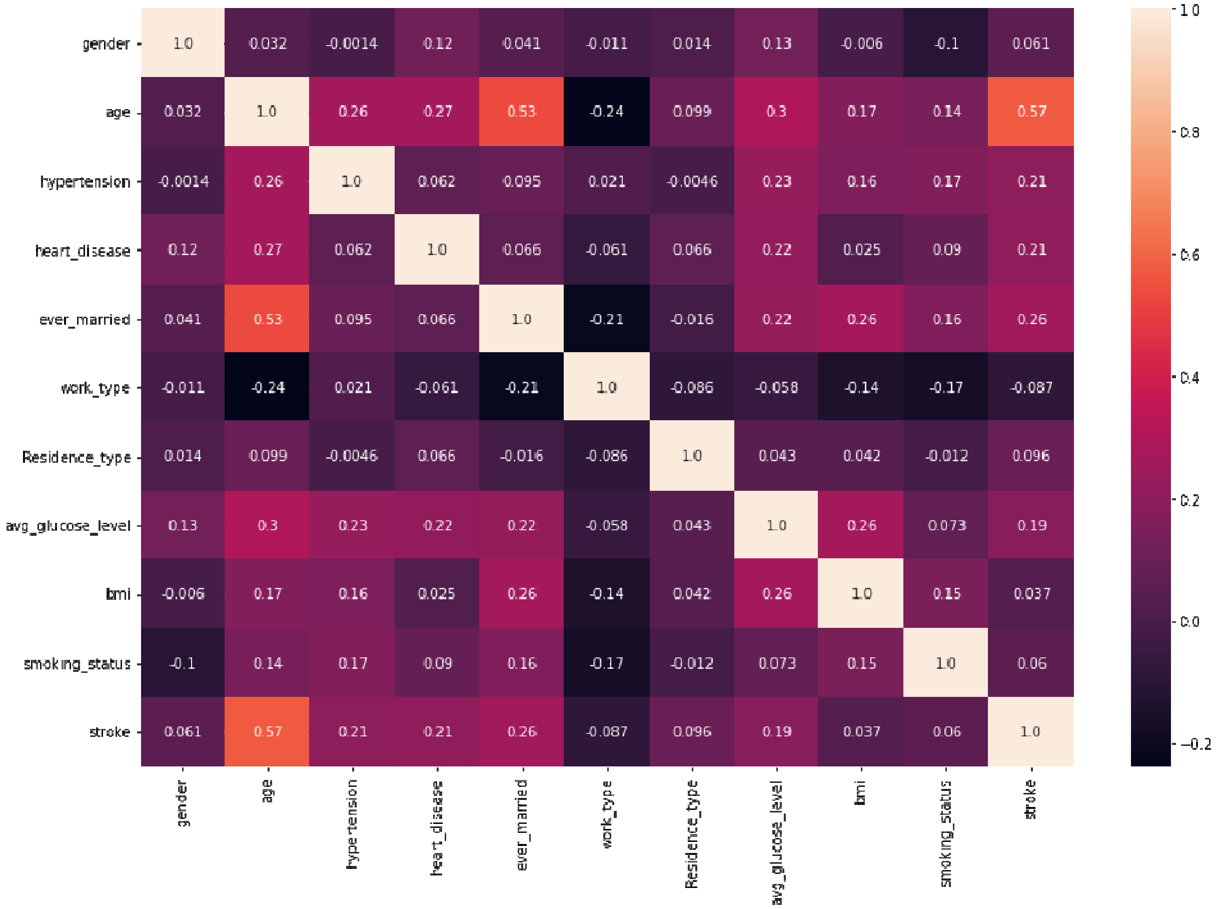

Figure 2: High impact fields in the examined dataset.

Our research work makes the following contribution: This paper compares the accuracy of machine learning algorithms in stroke prediction by using various attributes in the data set. The preprocessed 5110 rows with mean valued bmi, the preprocessed 5110 rows with median bmi, and the 498 data of 249 stroke positive and 249 stroke negative cases with mean bmi are the under-sampled data sets. The precision, recall, accuracy, F1 and support scores, ROC curve, and accuracy metric were used to compare the accuracy of the four models mentioned above.

4. MODEL ANALYSIS:

	[6]K-Nearest Neighbor Classifier	**[7]Decision Tree Classifier**	**[8]Gradient Boost Classifier**	**[9]XG Boost Classifier**	**[10]Cat Boost Classifier:**
Application	Simple ML algorithm used for classification and regression.	Super vised classification algorithm follows tree-structure.	One of the boosting algorithm well suited for complex and large dataset.	It is designed to address regression and classification issues, particularly with larger datasets.	It provides fast categorization without adjusting the features or requiring GPU support.
Role	The error value will be reduced by perfect K-value	It functions by representing decisions as a branch, data features as internal nodes and the final output as leaf node	It has a fixed base estimator value whereas the AdaBoost-base estimator value will change based on the input needs.	Runs on several GPUs over multiple networks, can accomplish this.It has the highest accuracy score out of all algorithms.	This model of supervised machine learning works effectively with noisy and heterogeneous data. In contrast to larger data, Cat Boost achieved higher accuracy in balanced data in this comparison analysis.

Table1 : Comparative analysis of Machine learning models

5. RESULT AND DISCUSSION:

This study used the machine learning algorithms KNN, Decision Tree Classifier, Gradient Boost Classifier, XG Boost Classifier, and Cat Boost Classifier to examine the effects of various sampled datasets on the categorization of stroke illness. The study's dataset has 5110 rows and 12 characteristics. Out of the twelve characteristics, age, heart disease, high blood pressure, employment type, smoking status, average blood sugar level, and body mass index have the greatest influence on stroke risk.

Preprocessing has been done on the huge and unbalanced data so that machine learning algorithms may use it. Firstly, stroke prediction does not require the patient's unique id field. As a result, the dataset no longer contains it. The BMI value, which has 201 null values, is an important determinant of stroke risk. It must therefore be filled in using the mean or median value. As a result, the dataset was divided into four samplings, and five distinct machine learning techniques were used to examine each under-sampled set of data. The outcome demonstrated that in all four samples, XG Boost Classifier had the highest average score.

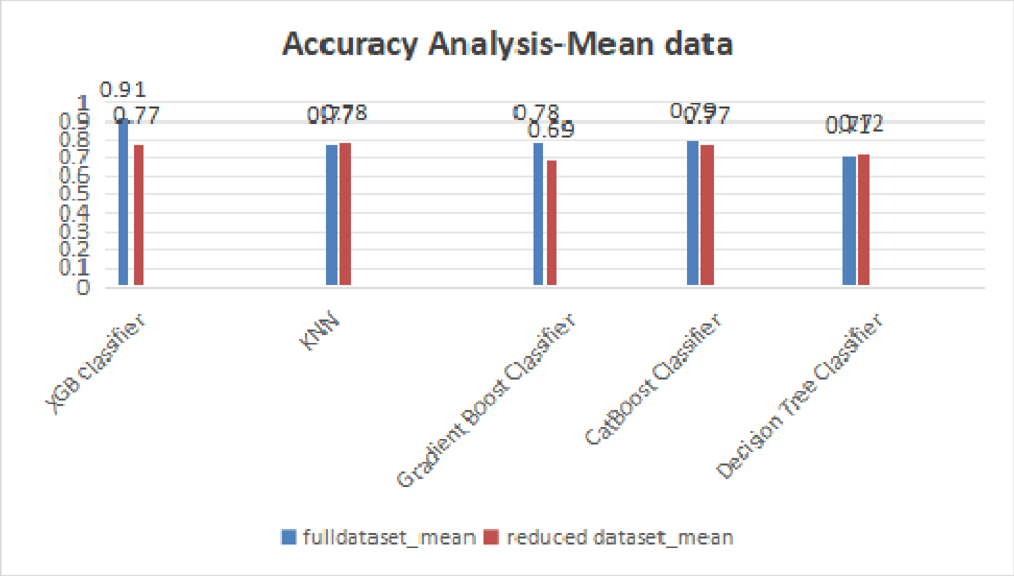

Figure 2: Accuracy Analysis of ML Models

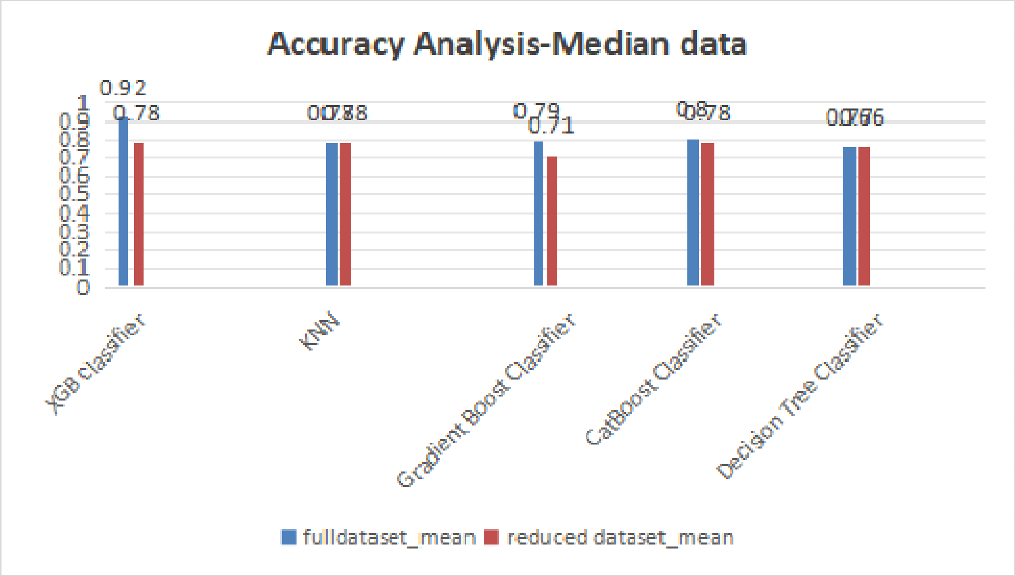

Figure 3 : Accuracy Analysis of ML Models

The models' performance was evaluated [12] using the ROC, accuracy score, precision score, recall score, and f1-score. The outcome shown that: 1) The machine learning models outperformed the smaller data set in the larger one.
2) The prediction rate for the BMI median field was greater than the mean value.
3) Among all machine learning models, the XG Boost Classification method has the greatest accuracy score.

5. CONCLUSION:

Timely classification of strokes can assist doctors in ensuring the appropriate treatment is administered. Machine learning is essential in the healthcare industry for analyzing a patient's historical data in drder to make predictions for the future. This study examines how the data set and efficient of machine learning models affect diverse under-sampled-data. The XG Boost classification algorithm outperformed KNN, Gradient Boost classification, Cat Boost classification, and Decision Tree classification with an impressive accuracy of 92%.

References:

1. Rajya sabha television report : https://www.youtube.com/watch?v=2F_JgLZSOWg&t=198s
2. Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Mamun, M. S. al, & Kaiser, M. S. (2020). Performance Analysis of Machine Learning Approaches in Stroke Prediction. *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, 1464–1469. https://doi.org/10.1109/ICECA49313.2020.9297525
3. Gazi Üniversitesi, Aksaray Üniversitesi, University of Buner, International Islamic University (Islāmābād, P., Institute of Electrical and Electronics Engineers. Turkey Section, & Institute of Electrical and Electronics Engineers. (n.d.). *2nd International Conference on Electrical, Communication, and Computer Engineering (ICECCE 2020) : 12th-13th June 2020, Istanbul, Turkey*.
4. Sailasya, G., & Aruna Kumari, G. L. (n.d.). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 12, Issue 6). www.ijacsa.thesai.org
5. Teoh, D. (2018). Towards stroke prediction using electronic health records 08 Information and Computing Sciences 0806 Information Systems. *BMC Medical Informatics and Decision Making*, *18*(1). https://doi.org/10.1186/s12911-018-0702-y

6. Documentation : https://www.ibm.com/in-en/topics/knn
7. Documentation:https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96

8. Documentation: https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/
9. https://practicaldatascience.co.uk/machine-learning/how-to-create-a-classification-model-using-xgboost
10. Documentation: https://catboost.ai/#:~:text=CatBoost%20is%20an%20algorithm%20for,CERN%2C%20Cloudflare%2C%20Careem%20taxi.
11. https://neptune.ai/blog/how-to-compare-machine-learning-models-and-algorithms
12. https://www.ritchieng.com/machine-learning-evaluate-classification-model/