IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005275 |

Ameliorating Accuracy Using XGBoost

R. Sudha Kishore[1], Pachabotla Anuhya[2], Nallagorla Yamini[3], Pentela Mrudula[4]

[1] Associate Professor, Department of IT, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Andhra Pradesh, India

[2,3,4] UG Student, Department of IT, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India

ABSTRACT: Machine learning uses programmed algorithms that receive and analyze input data to predict output values in a specific range. To classify a data we use legacy techniques like Naive Bayesian, Decision Tree Algorithms which give certain results. In order to get accurate results while classifying the datasets we use XGBoost algorithm on both binary data(Titanic datasets) and multinary data(Glass Type Detection data sets) . The term "Boosting" refers to a family of an algorithms which converts weak learners to strong learner. XGBoost(eXtreme Gradient Boosting) is a boosting algorithm that has recently been dominating applied machine learning for structured or tabular data. XGBoost is an implementation of gradient boosting decision trees. The goals of implementing XGBoost algorithm is to enhance the Execution Speed and Model Performance.

KEYWORDS: -XGBoost, Titanic Dataset, Glass Type Detection Dataset

I. INTRODUCTION

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. Machine learning is a branch of artificial intelligence that aims at solving real life engineering problems. It provides the opportunity to learn without being explicitly programmed and it is based on the concept of learning from data. It is so much ubiquitously used dozen a times a day that we may not even know it. Few decades ago, analysis of data for effective and efficient decision making in our organizations was through the use of Mathematical and Statistical methods where tools such as charts, regression

method, etc. were used for decision making. It was however, observed that the estimated amount of information in the world doubles every twenty (20) months leaving the size and number of the databases increasing even faster (Ann, 1997) and because of this large data size, extraction of useful and meaningful information has proven so herculean and challenging as the primitive tools can no longer analyse these datasets. The large data made readily available in our information industry are not useful until being worked upon and converted into useful and meaningful information.

II.OBJECTIVES

The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data. The problem of predicting students' performance and failure rate among the students is one of the famous applications of EDM. Using data mining in higher education is a recent research field; a lot of work has been done recently on the application of machine learning to educational databases to solve educational challenges that has to do with prediction of student's performance.

III.MOTIVATION

In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data. Actually we have so many classifiers to classify the datasets Like K-nearest neighbor classifier, random forest classifier, decision trees, etc. They won't giveaaccurate results while classifying the datasets. So, we researched about the classifying techniques which gives the accurate results. We found the boosting algorithms. In machine learning we have 3 different boosting algorithms named as Adaptive Boosting Algorithm, Gradient Boosting Algorithm and Extreme Gradient Boosting Algorithm. In these 3 Algorithms Extreme Gradient Boosting Algorithm is very fast, secure, high accuracy and also execution speed is better than other classifiers. So, I have used Extreme

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005275 |

Gradient Boosting Algorithm to classify the datasets.

IV.LITERATURE SURVEY

Machine learning is a broad term encompassing a number of methods that allow the investigator to learn from the data. These methods may permit large real-world databases to be more rapidly translated to applications to inform patient-provider decision making.
Machine learning is a broad term encompassing a number of methods that allow the investigator to learn from the data. These methods may permit large real-world databases to be more rapidly translated to applications to inform patient-provider decision making.
A total of 34 publications from January 2014 to September 2020 were identified and evaluated for this review. There were diverse methods, statistical packages and approaches used across identified studies. The most common methods included decision tree and random forest approaches. Most studies applied internal validation but only two conducted external validation. Most studies utilized one algorithm, and only eight studies applied multiple machine learning algorithms to the data. Seven items on the Luo checklist failed to be met by more than 50% of published studies.

This study originated from a systematic literature review that was conducted in MEDLINEand PsychInfo; a refreshed search was conducted in September 2020 to obtain newer publications. Eligible studies were those that analyzed prospective or retrospective observational data, reported quantitative results, and described statistical methods specifically applicable to patient-level decision making. Specifically, patient-level decision making referred to studies that provided data for or against a particular intervention at the patient level, so that the data could be used to inform decision making at the patient-provider level. Studies did not meet this criterion if only a population-based estimates, mortality risk predictors, or satisfaction with care were evaluated. Additionally, studies designed to improve diagnostic tools and those evaluating health care system quality indicators did not meet the patient-provider decision-making criterion. Eligible statistical methods for this study were limited to machine learning-based approaches. Eligibility was assessed by two reviewers and any discrepancies were discussed; a third reviewer was available to serve as a tie breaker in case of different opinions. The final set of eligible publications were then abstracted into a Microsoft Excel document.Study quality was evaluated using a modified Luo scale, which was developed specifically as a tool to standardize high-quality publication of machine learning models.

V. EXISTING SYSTEM

In the existing system, they have classified the datasets by using the machine learning legacy techniques like naïve Bayesian classifier, random forest classifier, k-nearest neighbour classifier, etc. By using those classifiers we will not get the accurate results. And the accuracy of those techniques are not better and execution speed is very low for the large datasets. Those classifiers need large amount of storage to store the datasets.

VI.PROPOSED MODEL

XGBoost (eXtreme Gradient Boosting) is a popular supervised-learning algorithm used for regression and classification on large datasets. It uses sequentially-built shallow decision trees to provide accurate results and a highly-scalable training method that avoids overfitting.XGboost makes use of a gradient descent algorithm which is the reason that it is called Gradient Boosting. The whole idea is to correct the previous mistake done by the model, learn from it and its next step improves the performance. The previous results are rectified and performance is enhanced.
So, we are using XGBoost algorithm to classify the datasets. It gives more accurate results, execution speed is fast and it needs less amount of space to store the datasets. It is faster than Gradient Boosting. It supports regularization. It is designed to handle missing data with its in-build features.

VII. METHODOLOGY

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005275 |

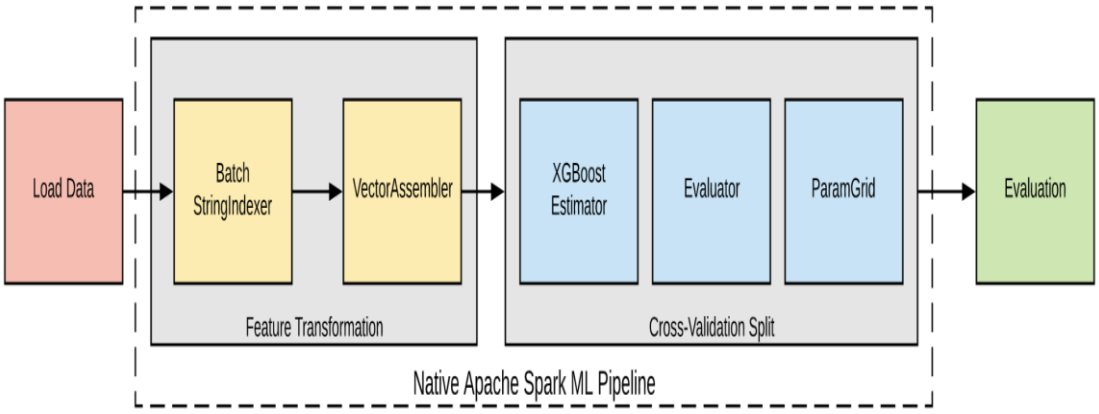


Fig 1: System Architecture

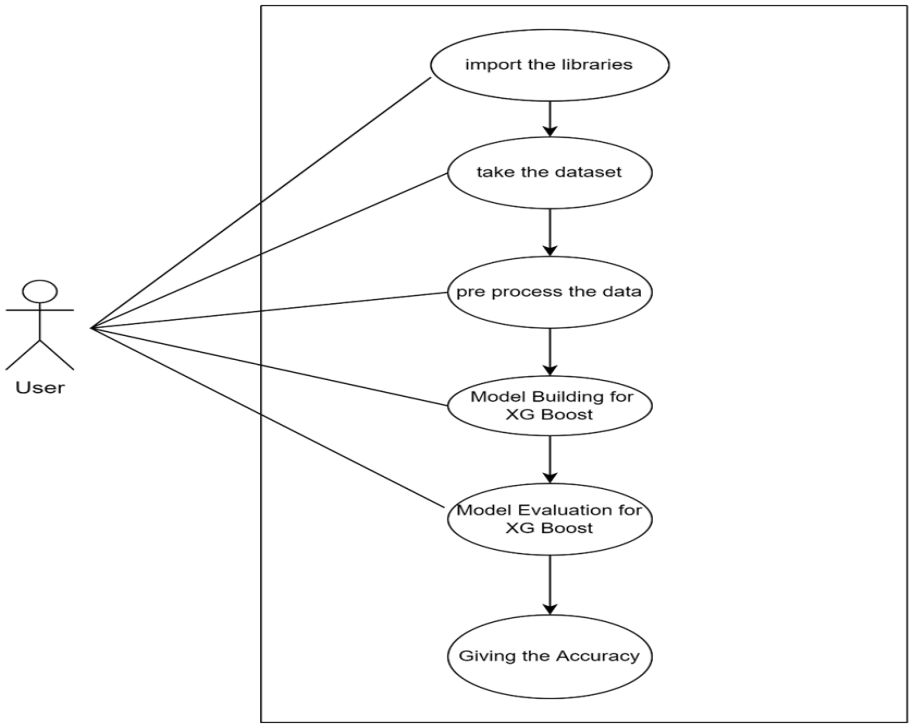


Fig 2: Use case Diagram

A. DATASET

A data set is an ordered collection of data. A dataset is a set of numbers or values that pertain to a specific topic. The numerical data set is a data set, where the data are expressed in numbers rather than natural language. A data set that has two variables is called a Bivariate data set.

B. DATA PREPARATION

Articulate the problem early. Establish data collection mechanisms. Check your data quality. Format data to make it consistent. Reduce data. Complete data cleaning. Create new features out of existing ones.

C. TRAINING ANDTESTING

Train/Test is **a method to measure the accuracy of your model**. It is called Train/Test because you split the data set

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005275 |

into two sets: a training set and a testing set. 80% for training, and 20% for testing. You train the model using the training set. You test the model using the testing set.

D. RESULT

In Classification, the output variable must be a discrete value. The task of the regression algorithm is to map the input value (x) with the continuous output variable(y). The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).

VIII. RESULTS & DISCUSSION

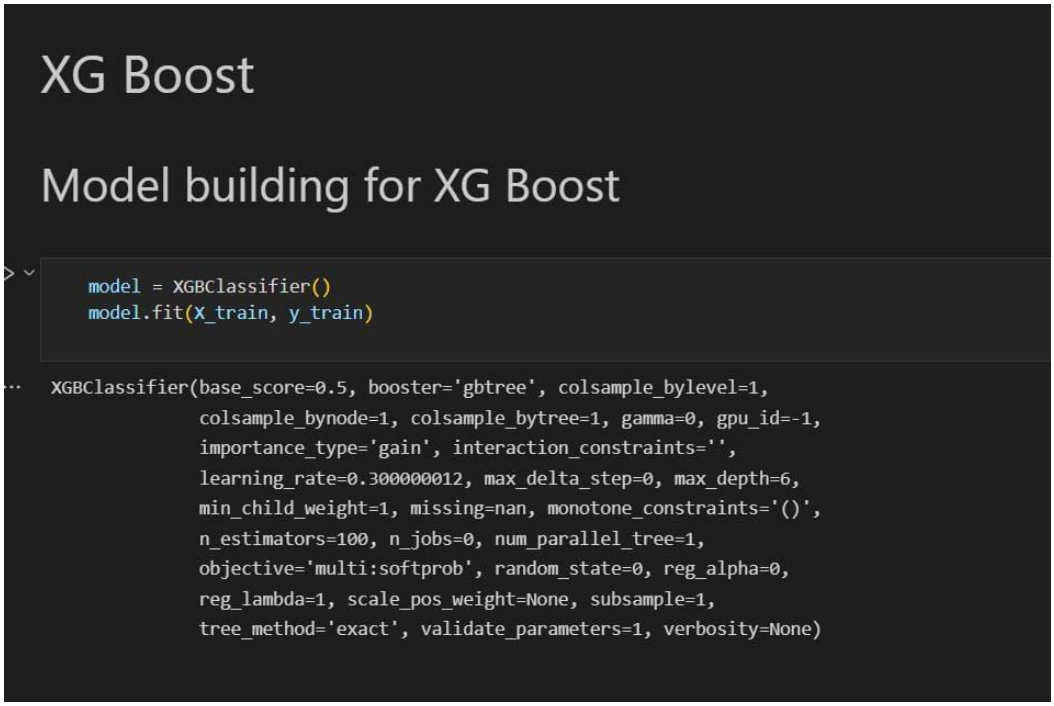
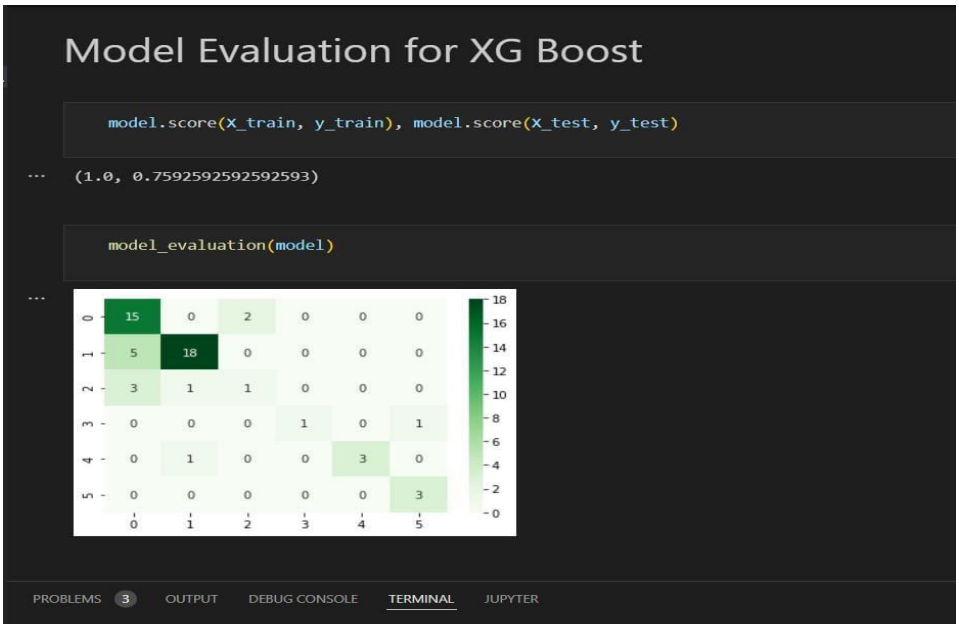


Fig 3: Model building for XG Boost Algorithm



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005275 |

IX. CONCLUSION

In this paper, a survey on Ameliorating accuracy using XGBoost is presented and various techniques have been studied and analysed for the same. In classification process, classifier plays a crucial part in which datasets are classified by using the training dataset and test dataset. A wide variety of approaches, algorithms, statistical software, and validation strategies were employed in the application of machine learning methods to inform patient-provider decision making. There is a need to ensure that multiple machine learning approaches are used, the model selection strategy is clearly defined, and both internal and external validation are necessary to be sure that decisions for patient care are being made with the highest quality evidence. Future work should routinely employ ensemble methods incorporating multiple machine learning algorithms.

REFERENCES

[1] Abdulsalam S.O., Adewole, K. S., Akintola, A. G. and Hambali, M. A. (2014). Data Mining in Market Basket Transaction: An Association Rule Mining Approach. International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA, 7(10), pp.15 – 20.

[2] Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting :ShubharthiDey, Yash Kumar, SnehanshuSaha, SuryodayBasak

[3] All Machine Learning Models - "https://towardsdatascience.com/all-machine-learning- models- explained-in-6-minutes-9fe30ff6776a"

[4] Scikit Learn - "https://en.wikipedia.org/wiki/Scikit-learn"

[5] An interpretable boosting model to predict side effects of analgesics for osteoarthritis: Liangliang Liu, Ying Yu, Zhihui Fei, Min Li, Fang-Xiang Wu, Hong-Dong Li, Yi Pan and Jianxin Wang

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details