EasyChair Preprint

№ 9194

Predicting Students Academic Performance Using Machine Learning Techniques

Amal Shaker and Abdelmoniem Helmy

October 31, 2022

Predicting Students Academic Performance using Machine learning Techniques

[1]Eng. Amal Shaker & [2]Dr. AbdelMoniem Helmy
[1]Lecturer, Sudan University of Science and Technology
[2]Assistant Professor, Faculty of Graduate Studies for Statistical Research, Cairo university
Cairo, Egypt
Email of the Corresponding Author: amal4tec@yahoo.com

***Abstract*—Educational institutes are concerned with determining the outputs of their educational programs, which are the knowledge and skills that their students acquire during their studies, and generally determine the educational level provided by these institutes. So, it was important to work on predicting academic performance by looking at the data and activities on student performance, attitudes, and interactions at school or university, then trying to predict whether a student would get a high, intermediate, or low score. The use of machine learning algorithms to predict students' outcomes based on their present behavior and performance has shown to be a useful tool for prophesying low, middle, and high levels results at various educational stages. Furthermore, it involves the need to achieve an accurate prediction of student's academic performance in the future based on their present behavior and performances is unquestionable. The use of a machine learning (ML) techniques is regarded as the most suitable approach for achieving this objective. Early prediction of student performance is very helpful in taking early action to improve learning outcomes. Predicting student performance from present academic data is one of the most useful applications at educational organizations in taking early action of improving learning outcome, thus is a valuable and good source of information as it can be used to improve student performance at the start of learning process, by having a high prediction accuracy of their performance. The paper attempts to identify best machine learning technique in predicting student academic performance demonstrate students' levels, through train several models for predicting whether a given student will have a high, medium, or low grades based on academic and behavior information. Therefore, it is crucial to evaluate the accuracy, precision and recall of machine learning models to determine which one best predicts students' performance. The study comes to the conclusion that it would be imperative to use a variety of machine learning techniques to effectively forecast student performance. It's critical to appropriately mattress machine learning models according to how accurate they can anticipate students' performance.**

Keywords—Machine learning Techniques; Academic Performance; Student Performance Prediction.

I. INTRODUCTION

Educational institutes are interested in specifying its educational program outcomes, which are the knowledge and skills gained by its students during their study, and they generally determine the educational level provided by these institutes. Therefore, it is essential that academic programs with particular qualification objectives develop the expected level of learning outcomes. Assessment of students' learning outcomes is a dynamic process that produces rich data, which helps stakeholders for decision making. Simultaneously, it is important for students, for parents and for employers. The Machine learning techniques could be used for extracting useful, valid patterns from learning outcomes data to contribute ensuring students maximize their academic output.

Predicting student success helps educational institutions to improve learning and teaching methods by defining educational goals that suit students and achieve the achievement of a variety of basic information [1]. The ability to predict in time the academic performance of students is very important in educational institutions[2]. Predicting student performance is a very important and sensitive area because it can help teachers identify students who need additional academic help[2]. Predicting students' academic performance helps teachers develop a good understanding of how well or poorly students are performing in their classes, so that teachers can take proactive measures to improve student learning, so accurate prediction of students' future performance based on their ongoing academic records is critical to effectively conducting needed pedagogical interventions to ensure that students complete the course on time and satisfactorily[3]. "[4]" traditional education curricula are usually unable to detect the large number of students at risk. "[5]" shows that data science algorithms play an important role in predicting students' academic performance. For this reason, data science techniques were adopted in this study to identify students' lagging by predicting their academic performance based on feature classes of demographic and academic characteristics and VLE interactions[3]. Data science technique specifically educational data mining was applied in this study to gain useful insights and

comprehensive analysis of the data set. Machine learning algorithms have been applied to extract hidden useful information from education data.

In this sense, educational organizations need to work on accurate prediction of students' future performance based on their ongoing academic records is essential for successfully carrying out necessary pedagogical interventions because it gives teachers a better understanding of how well or poorly students are performing in their classes, enabling teachers to take proactive measures to improve student learning.

The term machine learning is often referred to as an analytical process designed to discover patterns in data and relationships between data variables. Moreover, a key feature of machine learning is the ability to analyze complex nonlinear relationships, given that complex input variables are predicted. The concept of machine learning, how it actually works, and the algorithms applied in this project, will be explained extensively, and machine learning is very different from traditional computational approaches, where systems are explicitly programmed to compute or solve a problem [6]. [*1]The application of machine learning techniques to predict student performance, based on their background information and long-term performance, has proven to be a useful tool for predicting good and bad performance at various levels of education.[7].

II. RELATED WORKS

The concept of machine learning originated from this environment. Computers can analyze digital data to find patterns and laws in ways too complex for a human to do. The basic idea of machine learning is that a computer can automatically learn from experience[5]. Several studies have been conducted on using machine learning algorithms for early prediction of student performance and these studies fall into the field of Machine Learning Techniques which provides great value to educational organizations.

According to Hussain, Muhsin, Salal, Theodorou, Kurtoğlu and Hazarika, Machine learning is useful in monitoring and analyzing the learning process in schools, predicting learners' performance with required academic assistance, academic guidance and advice, examining the efficiency and effectiveness of learning methods, providing useful feedback to teachers and learners, and modifying learning environments in their favor about the students. The idea of looking for patterns and regularities in data is a fundamental concept in the field of pattern recognition and data classification. Machine learning often focuses on the development and application of computer algorithms in the field [5] Machine learning deals with the input data used to train a model as the model learns different patterns in the input data and uses that knowledge for unknown prediction results [8].

There are several tools that are used in EDM. Data manipulation and feature engineering tools include Microsoft Excel, EDM Workbench, Python and Jupyter notebook, and Structured Query Language (SQL). No one tool can be used for EDM as different tools suit different tasks [9]. A wide range of classification algorithms can be used to predict processes and performance, namely, random forest, support vector machines, AdaBoost, decision tree, Na¨ıve Bayes, and K-nearest neighbour[10]. Kumar et al.[11] In the context of the data extraction step, it is important to choose the correct method to appropriately handle the task. This is often done using machine learning methods.

The main difference between humans and computers for a long time has been that humans automatically tend to improve their way of tackling a problem. Humans learn from past mistakes and try to solve them by correcting them or finding new ways to tackle the problem. Conventional computer programs do not take into account the results of their tasks and are therefore unable to improve their behaviour. The field of machine learning addresses exactly this problem and involves creating computer programs that are able to learn and thus improve their performance by collecting more data and experience [12].

The ability to predict performance of students is very crucial in our present education system. However, it is not evident which machine learning model is best in predicting student performance and which one is best in improving learning outcome [13]. In 1967, the first pattern recognition program was able to detect patterns in data by comparing new data to known data and finding similarities between them. [12] Machine learning consists of three categories such as supervised learning, unsupervised learning, and reinforcement learning. This deals with supervised learning and we will discuss it in the next section, for now we can define supervised learning as the approach in which the model is trained using input and output labels.[14]

[*] https://mm.tt/map/2389082676?t=M07mPQk7Sl

In contrast, unsupervised learning is where the dataset contains input labels, (for example, the model is trained with unlabeled data), from which it learns different patterns and structures and is implemented in applications such as visual recognition, robotics, speech recognition, etc. that. on me. Reinforcement learning deals with learning how to achieve a complex goal by maximizing a specific dimension step by step[11].

III. EXPEREMANT

This study applies machine learning techniques to evaluate and predict academic performance of students, based on their performance and behavior according to the appropriate level of (high, medium, or low grade)., based on academic and socio information. In addition, use preset data of students' learning outcomes to predict their final achievement status. Furthermore, the study includes several Models to achieve a defined goal. The results of the system be tested and compared with manual work. The dataset includes students' attendance, parents' participation in the educational process, and parent response survey. Models be trained with four features containing students' academic, and socio information.

Supervised Learning: In simple terms, supervised learning means the tuning of model parameters using labeled data sets so that the tuned model parameters can work for larger and unseen data [15].

have n models y=fw1(x)y=fw1(x), y=fw2(x)y=fw2(x),…, y=fwn(x)y=fwn(x), and we select the best model y=fwi(x)y=fwi(x) through training and validation processes by using a labeled data set. The performance of the selected model is evaluated by testing it on another data set. Therefore, the objectives of the supervised learning can be divided into the following steps:

tuning model parameters, (2) generating algorithms for tuning, (3) improving the models to work with unseen data, and (4) applying efficient quantitative and qualitative measures for tuning.

Supervised learning is similar to the sampling techniques used in statistics to estimate population parameters from the random samples. However, the distinction is the adoption of supervised learning in the development of the machine learning algorithms. These objectives may be divided systematically into three algorithms, namely training, testing, and validation algorithms [16].

Training Algorithm: Using a labelled data set, the training algorithm must offer a methodical way to build model parameters and choose the best ones. The labelled data set in this instance is referred to as the training data set. The error between the predicted class labels and the actual class labels in the training data set is minimized as a result of choosing the optimum parameters. We require a reliable quantitative measure to accomplish this.

Testing Algorithm: The testing algorithm must also offer a methodical way to determine whether the model with tuned (optimal) parameters performs well with a different set of labelled data. The data set in this instance is referred to as the test data set. To demonstrate that the tuning settings are effective, the trained model must provide.

Supervised machine learning model is depicted graphically. Original preprocessed data sets containing known variables and targets are divided into training data and test data in supervised learning. (Above) The training data are used to train a learning algorithm in an attempt to develop an accurate predictive model during the training phase. The test data are then applied to the model to evaluate it, and the predictive accuracy is evaluated. (Below) After the model has been validated, new data is fed into it in an attempt to make new predictions.

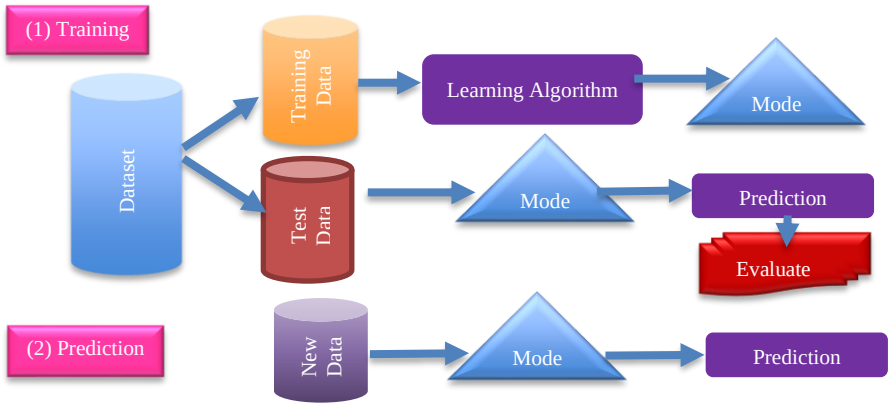

Fig. 1: Supervised machine learning Model

The accurate prediction of students' academic performance is of importance to institutions as it provides valuable information for decision making in the admission process and enhances educational services by allocating customized assistance according to the predicted performance. The purpose of this study is to build and train a model with high prediction ability trained by four recent heuristic algorithms inspired by the behaviors of students and parents, namely, LG, DT, KNN, and SVM algorithms. The study used previous exam results and other factors, for example academic background features such as educational stage, grade Level and section, also behavioral features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction as input variables, and predicted the student's expected performance. supervised machine learning algorithms were utilized to train the model for prediction. These algorithms optimized the classifier, since it is a 3-dimensional problem, our decision boundary will be a plane, and similarly, the complexity of the solution will increase with the rising number of features. The Model's results of all algorithms were then discussed and analyzed. It found that the Model which trained by the DT algorithm is record the highest Accuracy and Precision rates, measures of the rest of models have recorded results all of which are almost similar in terms of accuracy, so any of them could be used in the prediction of students' academic performance. This work is expected to be used to support student admission procedures and to strengthen the service system in educational organizations. In general, this paper will perform testing of four classification algorithms namely LG, DT, KNN and SVM and compare algorithms to evaluate the accuracy, precision and recall of machine learning models to determine which one best predicts students' performance measures.

IV. DATASET USED

A. *Dataset Description:*

A dataset created from a higher education institution (acquired from several disjoint databases) related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. The dataset includes information known at the time of student enrollment (academic path, demographics, and social-economic factors) and the students' academic performance at the end of the first and second semesters. The data is used to build classification models to predict students' dropout and academic success. The problem is formulated as a three-category classification task (dropout, enrolled, and graduate) at the end of the normal duration of the course. Feature information: detailed information about features used can be find in the following figure:

```
In [10]: # Show null counts and data types
         data.info()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 480 entries, 0 to 479
         Data columns (total 17 columns):
          #   Column                    Non-Null Count  Dtype
         ---  ------                    --------------  -----
          0   gender                    480 non-null    object
          1   NationalITy               480 non-null    object
          2   PlaceofBirth              480 non-null    object
          3   StageID                   480 non-null    object
          4   GradeID                   480 non-null    object
          5   SectionID                 480 non-null    object
          6   Topic                     480 non-null    object
          7   Semester                  480 non-null    object
          8   Relation                  480 non-null    object
          9   raisedhands               480 non-null    int64
          10  VisITedResources          480 non-null    int64
          11  AnnouncementsView         480 non-null    int64
          12  Discussion                480 non-null    int64
          13  ParentAnsweringSurvey     480 non-null    object
          14  ParentschoolSatisfaction  480 non-null    object
          15  StudentAbsenceDays        480 non-null    object
          16  Class                     480 non-null    object
         dtypes: int64(4), object(13)
         memory usage: 63.9+ KB
```

Fig. 2: Dataset Feature Information

Data Set Characteristics: Multivariate, Number of Instances: 480, Area: E-learning, Education, Predictive models, Educational Data Mining, Attribute Characteristics: Integer/Categorical, Number of Attributes: 16, Associated Tasks: Classification, Missing Values? No, File formats: xAPI-Edu-Data.csv

B. *Understand Dataset by Information:*

This is an educational data set which is collected from learning management system (LMS) called Kalboard 360. Kalboard 360 is a multi-agent LMS, which has been designed to facilitate learning through the use of leading-edge technology. Such system provides users with a synchronous access to educational resources from any device with

Internet connection. The dataset consists of 480 student records and 16 features. The features are classified into three major categories:

(1) Demographic features such as gender and nationality, (2) Academic background features such as educational stage, grade Level and section, and (3) Behavioral features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction.

The dataset consists of 305 males and 175 females. The students come from different origins such as 179 students are from Kuwait, 172 students are from Jordan, 28 students from Palestine, 22 students are from Iraq, 17 students from Lebanon, 12 students from Tunis, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from USA, Iran and Libya, 4 students from Morocco and one student from Venezuela.

The dataset is collected through two educational semesters: 245 student records are collected during the first semester and 235 student records are collected during the second semester.

The data set includes also the school attendance feature such as the students are classified into two categories based on their absence days: 191 students exceed 7 absence days and 289 students their absence days under 7.

This dataset includes also a new category of features; this feature is parent parturition in the educational process. Parent participation feature have two sub features: Parent Answering Survey and Parent School Satisfaction. There are 270 of the parents answered survey and 210 are not, 292 of the parents are satisfied from the school and 188 are not.

Students Numerical Classification: The students are classified into three numerical intervals based on their total grade/mark; it displays in the following table:

TABLE 1: STUDENTS NUMERICAL CLASSIFICATION

No.	Level	interval includes values
1.	Low	from 0 to 69
2.	Middle	from 70 to 89
3.	High	from 90-100

Attributes: detailed information about collected attributes can be find in the following table:

TABLE 2: DATASET ATTRIBUTES

No.	Feature	Type	Description	Attributes
4.	Gender	nominal	student's gender	'Male' or 'Female'
5.	Nationality	nominal	- student's nationality	Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia
6.	Place of birth	nominal	student's Place of birth	Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan',' Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia'
7.	Educational Stages	nominal	educational level student belongs	lowerlevel','MiddleSchool','HighSchool'
8.	Grade Levels	nominal	grade student belongs	'G-01', 'G-02', 'G-03', 'G-04', 'G-05', 'G-06', 'G-07', 'G-08', 'G-09', 'G-10', 'G-11', 'G-12
9.	Section ID	nominal	classroom student belongs	'A','B','C'
10.	Topic	nominal	course topic	' English',' Spanish', 'French',' Arabic',' IT',' Math',' Chemistry', 'Biology', 'Science',' History',' Quran',' Geology'
11.	Semester	nominal	school year semester	' First',' Second'
12.	Parent	nominal	responsible for student	'mom','father'
13.	Raised hand	numeric	how many times the student raises his/her hand on classroom	0-100
14.	Visited resources	numeric	how many times the student visits a course content	0-100

15.	Viewing announcements	numeric	how many times the student checks the new announcements	0-100
16.	Discussion groups	numeric	how many times the student participate on discussion groups	0-100
17.	Parent Answering Survey	nominal	parent answered the surveys which are provided from school or not	'Yes','No'
18.	15 Parent School Satisfaction	nominal	the Degree of parent satisfaction from school	'Yes','No'
19.	Student Absence Days	nominal	the number of absence days for each student	above-7, under-7

C. Understand Dataset with Visualization:

Understanding the dataset is necessary to get the best results from machine learning algorithms. The usage of data visualization is the quickest technique to understand more about datasets. The author of this thesis utilized the Pandas module to demonstrate how to precisely visualize machine learning data in Python.

Missing values: Missing values are a common issue in many real-world datasets. Missing values can skew the results of machine learning models and/or reduce the model's accuracy. An overview of the history of the development of decision tree induction algorithms is followed by a review of techniques for dealing with missing attribute values in the operation of these methods. missing values is a waste of data which can result from case wise deletion of missing values in statistical algorithms is discussed and alternatives proposed.[17]

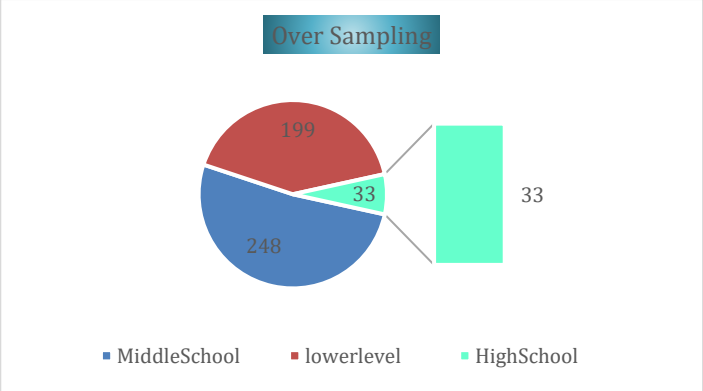

Fig. 3: Dataset Distribution of Stage ID

Author can find that Missing data: is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset.

Data Balance: Although the data is balanced, the author encountered a problem in distributing the data, which maybe it requires for data augmentation. The author faced a problem of distribution records among the three educational levels to which a student belongs is skewed, with one minority level, "HighSchool," standing out data within the dataset used in the current research, where the graph down showed the third value is significantly less than in the other values as show in the figure.

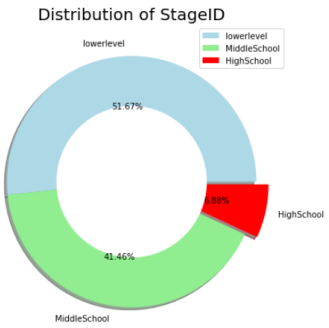

Fig. 4: Distribute Data Before Augmentation:

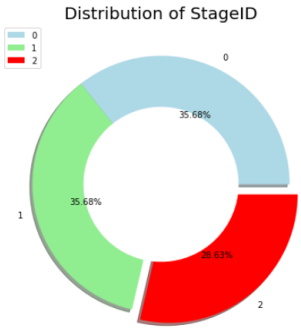

Fig. 5: Distribute Data After Augmentation

D. Data Sampling Techniques

To overcome the class imbalance problem, sampling strategies are frequently used, such as removing some data from the majority class (under-sampling), duplicating data from the minority class (over-sampling), or adding artificially generated data to the minority class. The disadvantage of undersampling techniques is that they reduce the size of the data set; the disadvantage of oversampling by data duplication approaches is that they add no new information to the models. In imbalanced classification, data augmentation approaches based on the synthesis of new data from the minority class have produced very good results. This bias in the training dataset can influence many of machine learning algorithms, causing some to completely disregard the minority class. This is a problem because minority predictions are typically the most important. One method for dealing with class imbalance is to randomly resample the training dataset. To randomly resample an imbalanced dataset, the two main approaches are to delete examples from the majority class, known as undersampling, and to duplicate examples from the minority class, known as oversampling.

Resampling involves creating a new transformed version of the training dataset in which the selected examples have a different class distribution. This is a simple and effective strategy for imbalanced classification problems. Applying re-sampling strategies to obtain a more balanced data distribution is an effective solution to the imbalance problem[18]. In this study used random oversampling for imbalanced classification strategies for data augmentation

E. Outliers:

Knowing how to find outliers in a dataset helps a lot to better understand your data. There are numerous definitions of outliers in the statistical and machine learning literatures. In the statistical literature, a commonly used definition is that outliers are a minority of observations in dataset that have different patterns from that of the majority of observations in the dataset. The assumption here is that there is a core of at least 50%of observations in a dataset that are homogeneous (that is, represented by a common pattern) and the remaining observations (hopefully few) have patterns that are inconsistent with this common pattern. For example, the points in the upper-left hand corner in the scatter plot in Figure 1 are obvious outliers.[19]

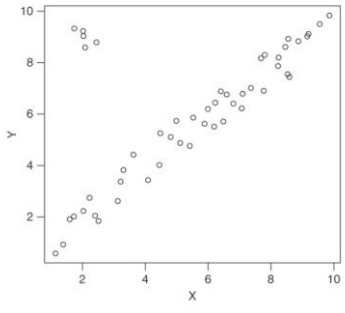

Fig. 6: Detect of outliers

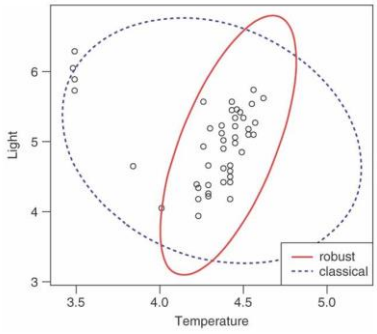

Fig. 7: Detect of outliers

In statistics, an outlier is an element that is out of the pattern characteristic of a particular set or combination. In all studies in statistics, mathematicians have accomplished algorithms capable of mitigating the impact of anomalous values, or canceling them, and even deleting them, using solid statistics methods. However, sometimes their presence is useful to know the behavior of a structure, or a system.

Density Plots: A fast way to get an idea of the distribution of each attribute is to look at histograms. Histograms group data into bins and provide you a count of the number of observations in each bin. From the shape of the bins you can quickly get a feeling for whether an attribute is Gaussian, skewed or even has an exponential distribution. It can also help you see possible outliers.
Density Plots: Density plots are another way of getting a quick idea of the distribution of each attribute. The plots look like an abstracted histogram with a smooth curve drawn through the top of each bin.
Univariate Plots: Accordingly, can see the distribution for each attribute is clear as shows in the following figure:

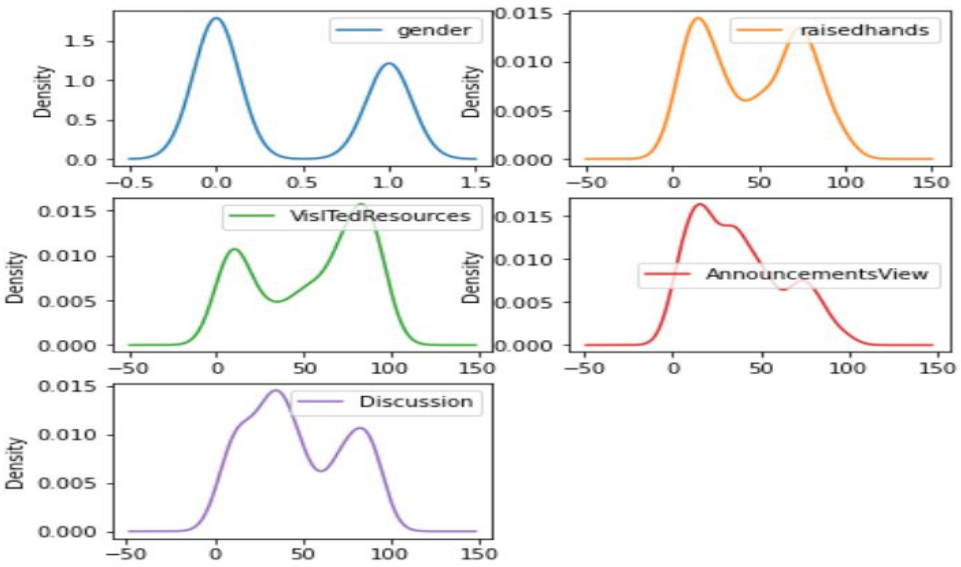

Fig. 8: Density plots of each attribute

For single outliers in normal theory fixed effects models a mean slippage model is commonly used. An alternative is to model the outlier as arising from an unknown observation with inflated variance. Maximum likelihood estimates for the position of the outlier under the two models need not agree. This paper considers maximizing a restricted part of the likelihood to estimate the variance parameters and characterizes these estimates in terms of standard least squares parameters. It is shown that the residual variance and outlier position are the same under both models.[20]

Summary: authors find that Outliers are a data point that is too high or too low relative to the nearest data point and the rest of the adjacent companion values in a data graph or dataset you are working with. Outliers can also severely affect the quality of the assumed statistical model, even to the extent of causing opposite conclusions.

V. EVALUATE MACHINE LEARNING MODEL

A. Measures

Through the training, validation, and testing processes, we have seen two measures—the quantitative measure and the qualitative measure—for fine-tuning the model parameters to achieve optimal values. Be aware that the supervised learning algorithms are supported by two crucial pillars: measures and metrics. A comparison value that is computed during training is referred to as a quantitative measure, and if it is computed during testing or validation, it is referred to as a qualitative measure. The qualitative indicators used in the testing and validation processes are referred to as metrics, whilst the quantitative measures utilised in the training algorithms are simply referred to as measures. These two categories of measures diagramed in the following figure.

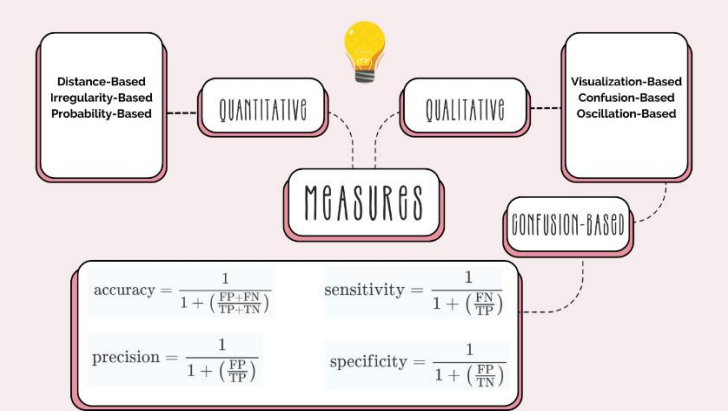

Fig. 9: the quantitative measure and the qualitative measure

B. Generating Confusion Matrix and Classification Report:

There are a lot of ways to measure the performance of your rating model but none have stood the test of time like the confusion matrix. It helps us evaluate how our model is performing, where something went wrong and gives us guidance to correct our route.

The Performance Evaluation of the Proposed Model: The proposed model is evaluated using several standard evaluation measures such as precisions, recall, accuracy, and F-score. Precision is the ratio of correctly predicted students' level rate in total students' levels samples. This means that the model correctly predicted the label for a proportion of the records in the dataset. This means that, out of all the levels rates given to students in the dataset, how many were correctly identified? The F-score is a measure of how well a model is performing. It takes into account both the precision and recall of the model.

Confusion Matrix Definition: Confusion Matrix, also known as an Error Matrix, is a table that describes the performance of a classification model. the confusion matrix function is also available in the metrics module and this case outputs a two-by-two matrix. The size is two-by-two because this is a binary classification problem, so if there were 3 possible response classes this would be a 3 by 3 matrix.

Other Definition: A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making. For a binary classification problem, we would have a 2 x 2 matrix as shown in the figure below with four values:

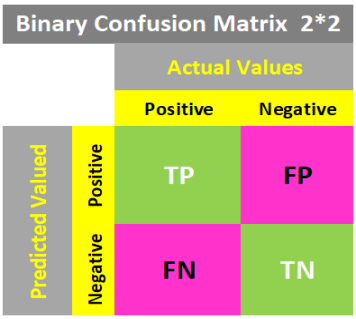

Fig. 10:Binary Classification Matrix

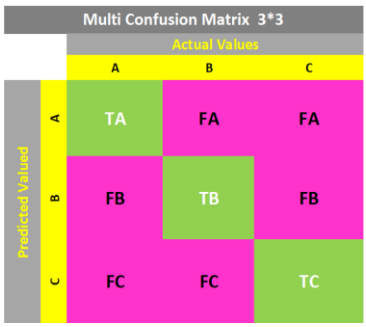

Fig. 11:Multi-Class Classification Matrix

1)The target variable has two values: Positive or Negative. 2)The columns represent the actual values of the target variable.3)The rows represent the predicted values of the target variable

Confusion Matrix for Multi-Class Classification: draw a confusion matrix for a multiclass problem where we have to predict whether a student get L, M or H as a class grad which represented by numeric array {0, 1, 2} The confusion matrix would be a 3 x 3 matrix like above:

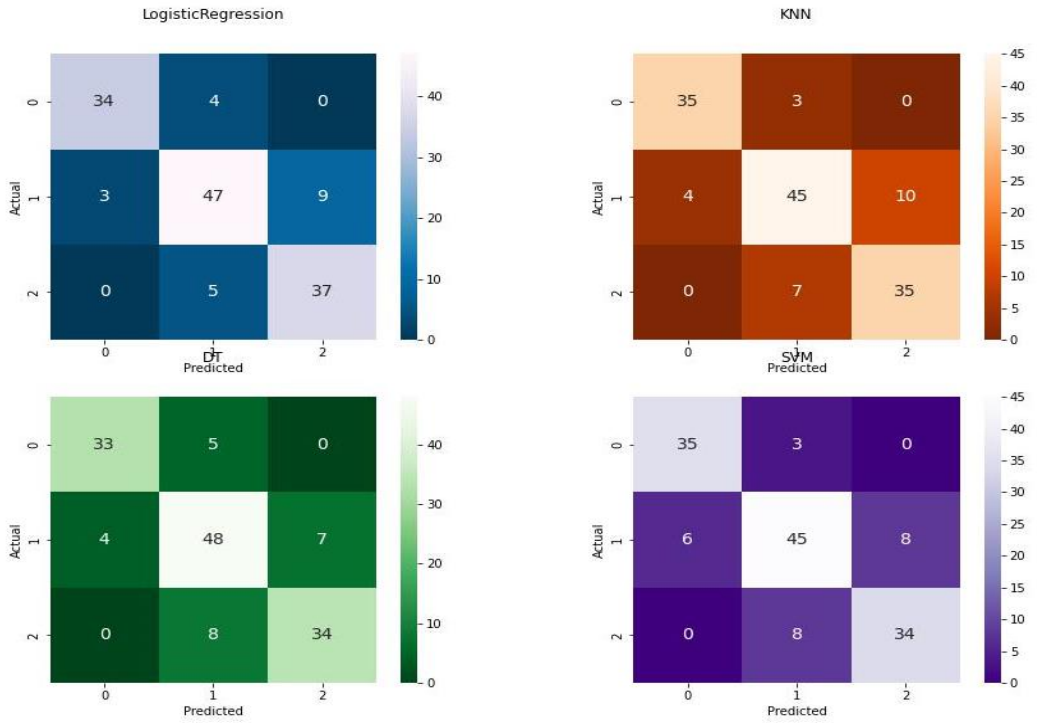

Fig. 12: Alx Net Confusion matrix

Our dataset is an example of an imbalanced dataset. There are 947 data points for the negative class and three data points for the positive class. This is how author calculate the accuracy:

C. Equations

$$Accuracy = \frac{Total\ correct}{Total\ observations} \tag{1}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$\boldsymbol{Accuracy = \frac{TA + TB + TC}{TA + FA + FA + FB + TB + FB + FC + FC + TC}} \tag{3}$$

$$\boldsymbol{LG.\ Accuracy = \frac{34 + 47 + 37}{34 + 4 + 0 + 3 + 47 + 9 + 0 + 5 + 37} = \frac{118}{139} = 85\%} \tag{4}$$

$$\boldsymbol{KNN.\ Accuracy = \frac{35 + 45 + 35}{35 + 3 + 0 + 4 + 45 + 10 + 0 + 7 + 35} = \frac{115}{139} = 83\%} \tag{5}$$

However, while working in an imbalanced domain accuracy is not an appropriate measure to evaluate model performance. For eg: A classifier which achieves an accuracy of 98 % with an event rate of 2 % is not accurate, if it classifies all instances as the majority class. And eliminates the 2 % minority class observations as noise.

Therefore, authors were used other metrics represents it is crucial to evaluate the accuracy, precision and recall of machine learning models to determine which one best predicts students' performance. The study comes to the conclusion that it would be imperative to use a variety of machine learning techniques to effectively forecast student performance. It's critical to appropriately mattress machine learning models according to how accurate they can anticipate students' performance.

VI. RESULTS AND DISCUSSION

A. *Accuracy Comparisons for All Algorithms*

TABLE 3: ACCURACY COMPARISONS

Algo.	LG	DT	KNN	SVM
Accuracy	81.295	84.173	82.734	82.014

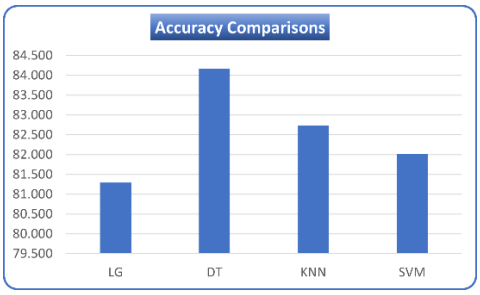

Chart 1:All Accuracy Comparisons

B. *Precision Comparisons for All Algorithms*

TABLE 4:PRECISION COMPARISONS

Algo.	LG	DT	KNN	SVM
Precision	81.741	85.686	83.131	82.218

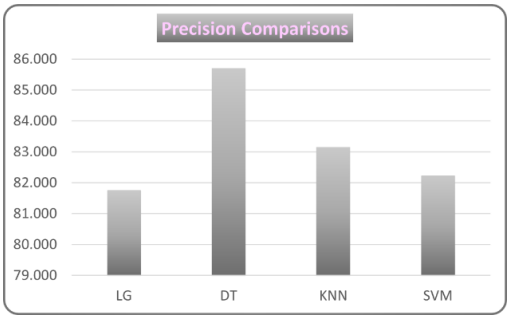

Chart 2:All Precision Comparisons

C. *Recall Comparisons for All Algorithms*

TABLE 5:RECALL COMPARISONS

Algo.	LG	DT	KNN	SVM
Recall	82.269	83.908	83.963	83.177

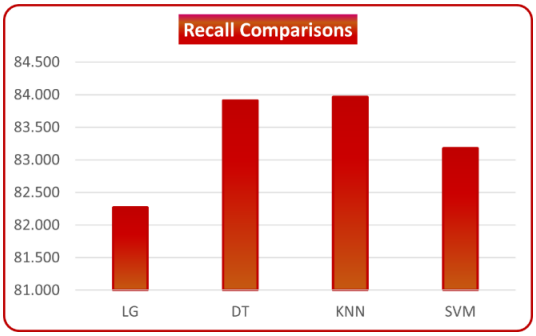

Chart 3:All Recall Comparisons

D. F1 Score Comparisons for All Algorithms

TABLE 6: F1 SCORE COMPARISONS

Algo.	LG	DT	KNN	SVM
F1 Score	82.269	83.908	83.963	83.170

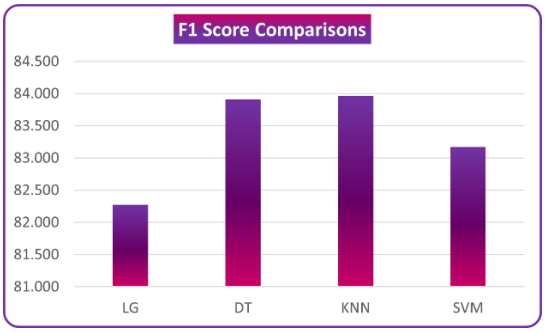

Chart 4: All F1 Score Comparisons

E. AUC Comparisons for All Algorithms

Receiver Operating Characteristic and Area Under Curve

TABLE 7: AUC COMPARISONS

Algo.	LG	DT	KNN	SVM
AUC	93.600	87.800	92.500	91.400

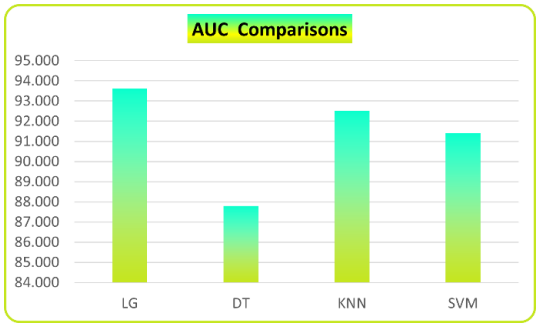

Chart 5: All AUC Comparisons

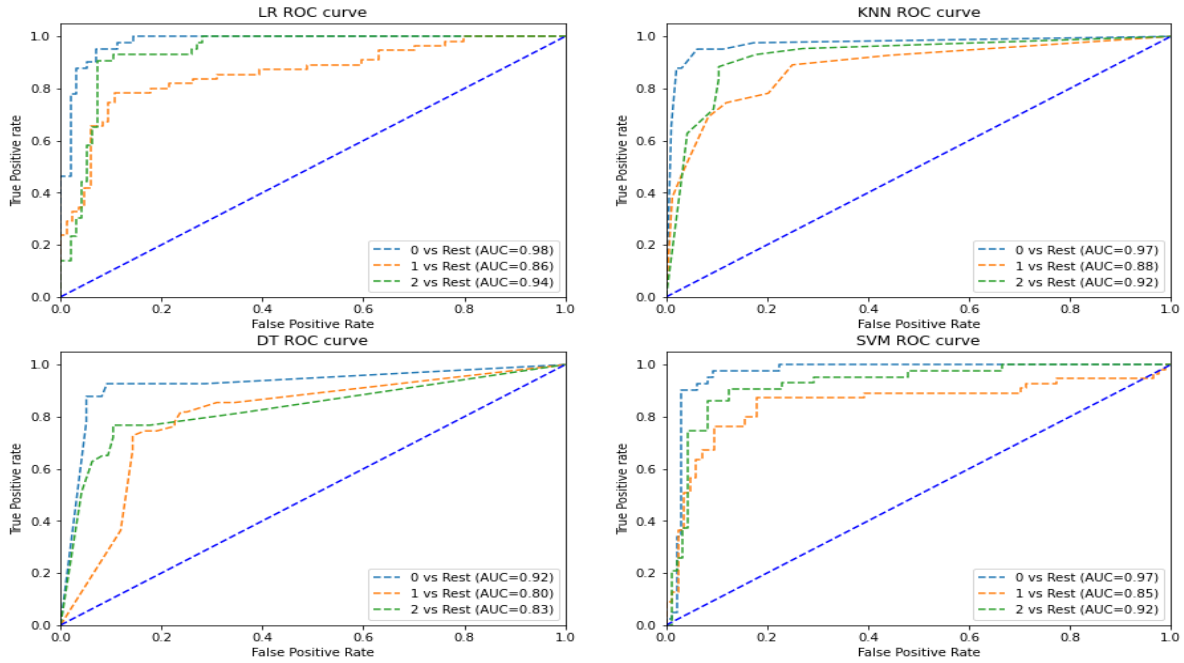

Fig. 13: AUC for All Models

F. Summary: Model Results Comparisons

TABLE 8:MODEL RESULTS COMPARISONS

Algo.	Accuracy	Precision	Recall	F1 Score	AUC
LG	81.295	81.741	82.269	82.269	93.600
DT	84.173	85.686	83.908	83.908	87.800
KNN	82.734	83.131	83.963	83.963	92.500
SVM	82.014	82.218	83.177	83.170	91.400

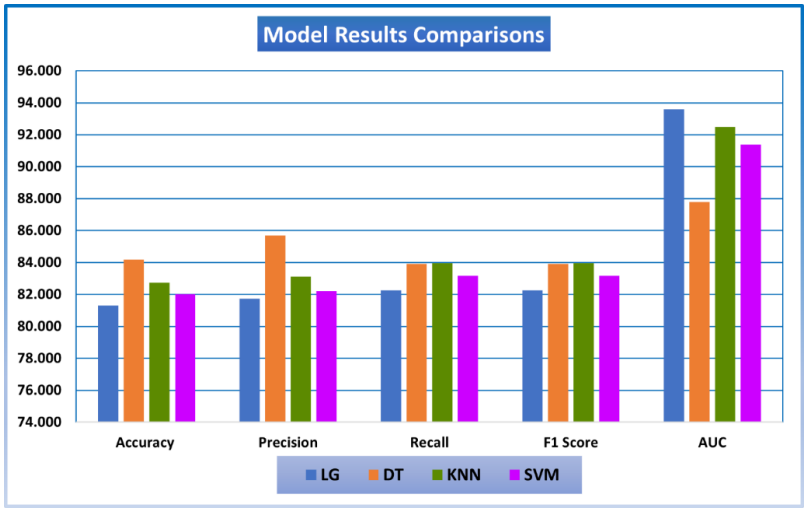

Chart 6:all Models Results Comparisons

VII. CONCLUSION

Of all the Algorithms applied in this study, authors saw that Machine learning techniques can use classroom engagement data, hands-raising, feedback from both teacher and parents, and attendance or frequent absenteeism status data to provide real-time feedback to students and teachers about individual student motivation, cognitive load, learning strategies, and public participation.

Future work: will include considering new and more datasets with higher number of instances and attributes, adding other techniques beside these, such as Functional neural networks associated with machine learning so that he can help help predict learners' individual learning paths and how their minds learn, allowing practitioners and teachers to implement effective learning tools at the individual student level.

ACKNOWLEDGMENT ***Not applicable.***

REFERENCES

[1] F. Ofori, E. Maina, and R. Gitonga, “Using Machine Learning Algorithms to Predict Students ’ Performance and Improve Learning Outcome : A Literature Based Review Francis Ofori , Dr . Elizaphan Maina and Dr . Rhoda Gitonga ISSN : 2617-3573 Using Machine Learning Algorithms to Predict Students,” *J. Inf. Technol.*, vol. 4, no. 1, pp. 33–55, 2020.

[2] M. Koutina and K. L. Kermanidis, “Predicting postgraduate students’ performance using machine learning techniques,” *IFIP Adv. Inf. Commun. Technol.*, vol. 364 AICT, no. PART 2, pp. 159–168, 2011, doi: 10.1007/978-3-642-23960-1_20.

[3] “Belachew & Gobena, 2017.” .

[4] H. Wang, Y. Ba, Q. Xing, and J. L. Du, “Diabetes mellitus and the risk of fractures at specific sites: A meta-analysis,” *BMJ Open*, vol. 9, no. 1, pp. 1–11, 2019, doi: 10.1136/bmjopen-2018-024067.

[5] M. Eaaswarkhanth, A. Al Madhoun, and F. Al-Mulla, “Could the D614G substitution in the SARS-CoV-2 spike (S) protein be associated with higher COVID-19 mortality?,” *Int. J. Infect. Dis.*, vol. 96, pp. 459–460, 2020, doi: 10.1016/j.ijid.2020.05.071.

[6] S. K. Yadav and S. Pal, “Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification,” vol. 2, no. 2, pp. 51–56, 2012, [Online]. Available: http://arxiv.org/abs/1203.3832.

[7] R. Tokunaga *et al.*, “CXCL9, CXCL10, CXCL11/CXCR3 axis for immune activation – A target for novel cancer therapy,” *Cancer Treat. Rev.*, vol. 63, pp. 40–47, 2018, doi: 10.1016/j.ctrv.2017.11.007.

[8] G. G. GORDON SAMMUT, “Points of View, Social Positioning and Intercultural Relations,” p. https://doi.org/10.1111/j.1468-5914.2009.00422.x.

[9] S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic, “Tools for Educational Data Mining: A Review,” *J. Educ. Behav. Stat.*, vol. 42, no. 1, pp. 85–106, 2017, doi: 10.3102/1076998616666808.

[10] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, “Educational data mining and analysis of students’ academic performance using WEKA,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 9, no. 2, pp. 447–459, 2018, doi: 10.11591/ijeecs.v9.i2.pp447-459.

[11] M. Kumar, S. Shambhu, and P. Aggarwal, “Recognition of slow learners using classification data mining techniques,” *Imp. J. …*, no. May, 2016, [Online]. Available: https://www.researchgate.net/profile/Mukesh_Kumar111/publication/316921996_Recognition_of_Slow_Learners_Using_Classification_Data_Mining_Techniques/links/591946f04585152e19a24b7b/Recognition-of-Slow-Learners-Using-Classification-Data-Mining-Techniques.pdf.

[12] B. P. Battula and R. Satya Prasad, “A novel network framework using similar-to-different learning strategy,” *AI Soc.*, vol. 30, no. 1, pp. 129–138, 2015, doi: 10.1007/s00146-013-0499-2.

[13] G. Kumar, “Creativity functioning in relation to personality, value-orientation and achievement motivation." Indian,” no. 110–115, p. Indian Educational Review.

[14] A. V. Francesco Camastra, *Machine Learning for Audio, Image and Video Analysis: Theory and Applications*. 2015.

[15] D. Berthelot, N. Carlini, I. Goodfellow, A. Oliver, N. Papernot, and C. Raffel, “MixMatch: A holistic approach to semi-supervised learning,” *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, pp. 1–11, 2019.

[16] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, “Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy,” *Proc. - Int. Carnahan Conf. Secur. Technol.*, vol. 2019-Octob, no. October, 2019, doi: 10.1109/CCST.2019.8888419.

[17] W. Z. Liu, A. P. White, S. G. Thompson, and M. A. Bramer, “Techniques for dealing with missing values in classification,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1280, pp. 527–536, 1997, doi: 10.1007/bfb0052868.

[18] P. Branco, L. Torgo, and R. P. Ribeiro, “A Survey of Predictive Modelling under Imbalanced Distributions,” 2015.

[19] R. Thompson, “A Note on Restricted Maximum Likelihood Estimation with an Alternative Outlier Model,” *J. R. Stat. Soc. Ser. B*, vol. 47, no. 1, pp. 53–55, Aug. 1985, [Online]. Available: http://www.jstor.org/stable/2345543.

[20] R. Thompson, “A Note on Restricted Maximum Likelihood Estimation with an Alternative Outlier Model,” *J. R. Stat. Soc. Ser. B*, vol. 47, no. 1, pp. 53–55, 1985, doi: 10.1111/j.2517-6161.1985.tb01329.x.