International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Comparison of Regression Algorithms for Predicting Students' Academic Performance

Mamta Saxena[#1], Sachin Gupta[#2]

[#]School of Engineering and Technology, MVN University Palwal, Haryana, India

17ca9001w@mvn.edu.in [2]sachin.gupta@mvn.edu.in

Abstract - In recent years, higher education has reached the peak level in growth. Many Universities, educational institutes and colleges are being set up at private level and public level for the growth of the education sector and for the well being of students [1]. Each of the institutions aims at providing higher knowledge to the students by grooming their faculties and implementing best teaching and educational practices. But still the problem persists as most of the students are at risk of unemployment due to poor academic performance. DM techniques can be used to analyse vast quantity of data that is readily available in bulk, extract usable knowledge, and help decision-making in order to forecast the academic success of students. [2].

The amount of data present in the education sector is examined, and by extracting the data from it, beneficial patterns are found. This is known as educational data mining. This paper discusses numerous regression techniques that can be used in educational institutions to forecast students' academic achievement. To forecast student performance, regression techniques such as Bayesian networks and decision trees can be used. Applications of the CART, J48, Random Forest, and ID3 decision tree algorithms are made to undergraduate student data obtained from private colleges in order to forecast how well they will perform on the final exam of the semester and determine whether or not they will be promoted for the following year. [3].

Index Terms—Educational Data Mining, Regression, Random forest, Student Academic Performance,

I. INTRODUCTION

In the present scenario of the education sector, the data mining theories and methods are gaining keen interests from researchers and educational leaders. In the real world, there is a large volume of data available from Universities, educational institutes and colleges and it is a challenging task to forecast student success in light of the rapid advancement of technology. Because of this, EDM becoming a popular new trend in the field of education. EDM is used in many different industries, including machine learning, statistics, medicine, and research.EDM is a new field of study that focuses on creating techniques for examining the distinctive and increasingly large- scale data collected from diverse colleges, universities, and educational institutions in order to comprehend people and the learning situations better.

Educational institutions typically gather enormous amounts of data about students, faculty, administration, and management of educational processes, but this data is useless because it is gathered but not appropriately utilised. Simple queries and conventional reports that can't be considerably used for decision-making are generated using this data. In the end, it leads to an increase in the volume and complexity of data that cannot be
effectively managed or turned into valuable patterns.

Finding patterns and information from massive data sets is a process known as data mining, sometimes referred to as Knowledge Discovery in Data (KDD). Data mining concepts and techniques can be used to a variety of fields, including marketing, healthcare, real estate, CRM, engineering, web mining, etc. [3] Education Data Mining is the name given to this method as it is currently being used in the educational field. To extract the valuable patterns from educational data, a variety of techniques like Bayesian Network, K-Nearest Neighbour, Naive Bayes, CART, and Decision Trees are used. By employing these methods,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

numerous information types can be retrieved to create classification andclustering algorithms, which can then be utilised to forecast students' academic success by adjusting various parameters.

The two criteria that can be set to forecast a student's academic achievement are MSP and ESP. Other factors, such as study time, internet access, alcohol consumption, and socio-economic status, can also be used to set standards for predicting students' performance and determine whether they will pass or fail the exam. If it is predicted that a student will fail the exam, additional effort can be made to boost his performance and lower his chances of failing.

II. RELATED WORKS

Several Studies have been conducted in educational Institutions for predicting the academic performance of students. It entails the analysis of numerous features and the sampling of data from various sources in order to forecast the grade of a student for various outcomes.

Researchers used data mining techniques like ID3, C4.5, and Bagging to separate the knowledge discovery from the student data base. They discovered that when there is noise, ID3 does not support cutting and does not produce accurate results. Additionally, the C4.5 technique eliminates the partial perspective of information gain when an attribute has a large number of possible values by building a tree using the gain ratio. Indicating that the model is successful in identifying the The outcome reveals that for ID3 and C4.5 decision trees, the classifier's accuracy in predicting the model's real positive rate for the FAIL class is 0.84.

Student's academic achievement have been tested by using decision tree algorithms like ID3, C4.5, and Classification and Regression Technique, or CART. By employing an information gain metric to divide an attribute, the ID3 Algorithm operates on this premise. Only categorical attributes are accepted. C4.5, which allows for the construction of decision trees using both category and nonstop attributes, is a superior interpretation of ID3. They evaluate depending on their accuracy and processing speed, decision tree algorithms are efficient.

In an effort to research studies progress has been made in studying student trends and behavior towards education and examining the behavioral patterns of students in a cross section. The researcher investigated the possibility of using data mining to evaluate UG final-year students in the educational system and presented their analysis of the findings using WEKA tool. The DBSCAN method is used to cluster the pupils and the ZeroR algorithm to classify them in order to boost student performance.

The rule mining method of evaluating student performance was also covered the outcomes of a tree-based categorization and reveals that it is difficult to predict and depend on the technical skill of the decision maker, neither of them employed associationrules in their research articles.

A new recommended system for forecasting was created for calculating students' graduation grades based on their admission results information. The suggested approach uses a top-down greedy search to analyse each attribute while using the ID3 algorithm to separate the data and build the decision tree. This study findings help management staff and academic planners improve student performance as a whole, lowering failure rates across most academic institutions.

A similar study was carried out by employing educational data processing to enhance the performance of 50 students from the Hindustan College of Arts and Science in Coimbatore, India. The performance of the student may be predictedusing decision tree classification on eight different criteria, which revealed that the categories of exam, seminar, attendance, and lab practical's are necessary. This forecast will enable the assessor to give students more attention and boost their confidence in their studies.

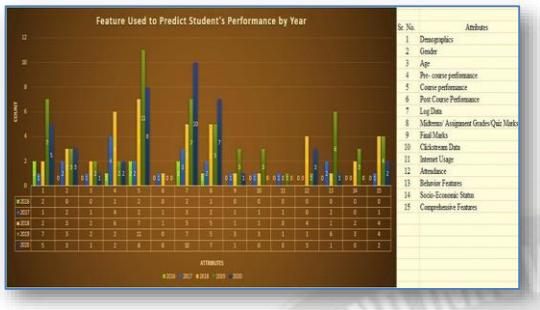
Classification and regression have been used extensively in studies to forecast student achievement. The decision tree, Naive Bayes, Support vector machine, logistic regression, and linear regression models were all employed by the researchers. The student's academic success can be predicted at the beginning of the programme, midway during the semester, or at the end of the semester.

Typically, two variables help predict how well scholars would succeed. The first are the important factors that go into prediction, as shown in figure, and the second is the prediction approach that goes

into analysing learner performance, as shown in figure 1.1.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

III. RESEARCH METHODOLOGY USED



This paper discussed various Regression Algorithms and their performances are compared by using machine learning. Regression is a supervised learning technique that helps in finding the relation between various variables and helps in predicting the values between the output variable and one or more predictor variables. Regression is mostly used to predict values other than those that can be forecasted and to identify the causal connections between variables. Analysing the relation between dependent variable and independent variable is made easier by the regression algorithm. Additionally, data trend analysis is performed using it. It aids in the prediction of continuous and real values. Regression makes it simple to identify the most crucial component, the least crucial element, and how each factor affects the others.

REGRESSION TECHNIQUES USED FOR PREDICTION OF STUDENT PERFORMANCE

There are various types of regression that can be applied to data sets to predict the academic performance of students. Regression can be linear, lasso, Random Forest, Support Vector, Decision tree, Polynomial and Logistic.

Linear Regression: One of the simplest regressions that may be used on continuous variables for predictive analytics is linear regression. It is used to solve the regression problem in machine learning. If one input variable is taken for study then it is known as Simple Linear Regression and if more than one input variable is taken for analysis then it is known as Multiple Linear Regression. Mathematically it can be represented as:

$$Y=ax+b;$$

Where Y is dependent variable and x is independent variable.

SVM: Classification and regression problems can be solved using the supervised machine learning technique known as SVM. Support Vector Regression is the term used when SVM is combined with regression. Working with continuous variables is the SVM algorithm. As many data points as possible must be highlighted within the boundary lines for SVR to be effective, and the hyperplane must contain as many data points as feasible to yield the best-fit line.

Decision Tree: Decision trees are used in data mining and machine learning. Both classification and regression problems can be solved using the decision tree method of supervised learning.. It can be applied to tackle category and numerical data-based challenges. To create a structure that resembles a tree, utilise a decision tree. Decision trees are created using an algorithmic method to determine the conditions under which data sets can be divided.

Naïve Bayes: Using the Bayes Theorem as its foundation, the Nave Bayes algorithm is one of the supervised learning algorithms. It is primarily utilised for classification issues and is useful for making rapid predictions when combined with machine learning models. One of the probabilistic classifiers, it predicts the likelihood of an object. Spam filtration, Sentimental analysis, and article classification are a few applications for the Naive Bayes algorithm.

Random Forest: By combining the characteristics of several decision trees with the Bootstrap and Aggregation technique, also referred to as bagging, RF is another potent algorithmic tool that can perform both regression and classification. This is done so that numerous decision trees can be combined to get the final result rather of just relying on a single decision tree. Its name, Random Forest, refers to the fact that it is a forest made up of several trees. The term "random" denotes the ability to generate several decision trees at random.

Neural Network: Neural networks, also referred to as artificial neural networks (ANNs) or simulated neural networks (SNNs), are the foundation of deep learning techniques. For neural networks to grow and improve their accuracy over time, training data is necessary. These learning algorithms allow us to

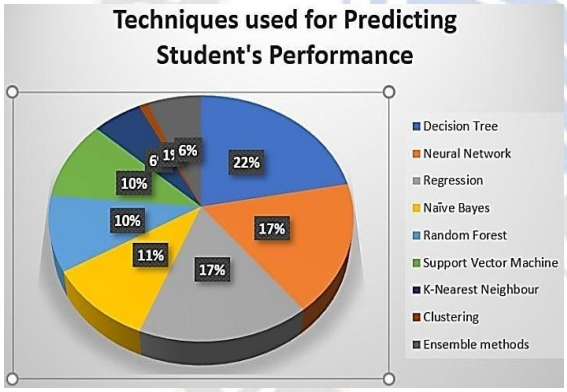
swiftly classify and cluster data, but when they are calibrated for accuracy, they transform into powerful tools for computer science and artificial intelligence. Tasks in speech

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

recognition or picture recognition can be done in minutes as opposed to hours when compared to manual identification by human experts.One of the most well-known neural networks is used in Google's search algorithm.

Clustering: The unlabeled dataset is classed using a machine learning technique called clustering or cluster analysis. This phrase means "a method of clustering the data points into different groups, each grouping similar data points." The things that might be connected are still grouped together, with little to no overlap with other objects. This is done by examining patterns in the unlabeled dataset, such as shape, size, colour, behaviour, etc., and classifying the data based on whether or not these patterns are present. It employs an unsupervised learning methodology with an unlabeled dataset, which means that the system receives no supervision.

A cluster-ID is assigned to each cluster or group following the use of this clustering technique, which can be used by ML systems.



IV. DATASET

The "Student Grade Prediction" Dataset is where the information for this study was found. As a result, it is an open-source dataset that can be used for academic and research purposes via the Kaggle Dataset repository. The dataset's main source is JBKnowledge Park, Faridabad. The dataset was compiled via reports and covers numerous factors such as student grades, demographics, and social features.

Table 1: Table Showing Dataset of Students

Attributes	Description
gender	Gender whether Male or Female;
section	section either A or B or C;
course	Course either CSE or EE or ECE;
socialnetworkingstatus	Status of Social networking;
parentseducation	Level of Parent's education;
lunch	Type of lunch
testpreparationcourse	Test preparation taken
istyearscore	Ist year marks
iindyearscore	IInd year marks
iiirdyearscore	IIIrd year marks
finalyearscore	Final year marks
status	Status

V. RESULTS

The following results are obtained by applyingoperations on the dataset

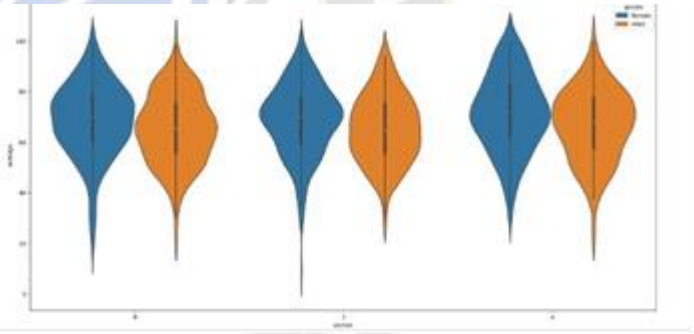


Fig 1 represents distribution of Average Score bySection and Gender

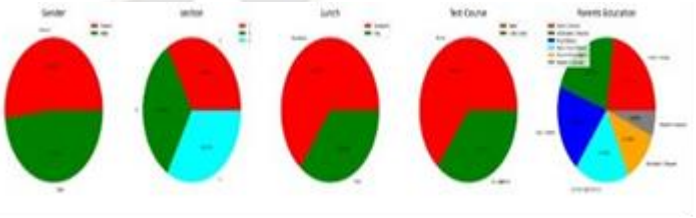
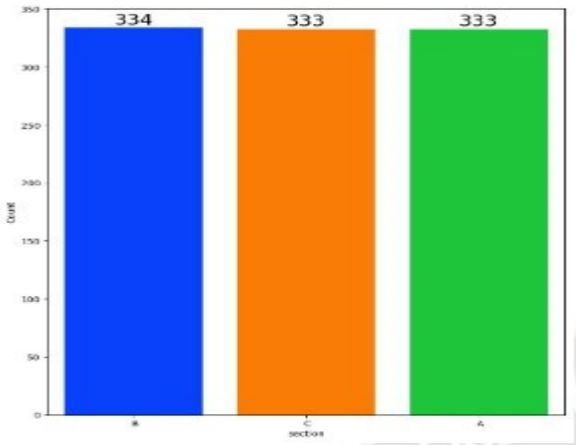


Fig 2 shows the distribution of various attribute ofdatasets

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*



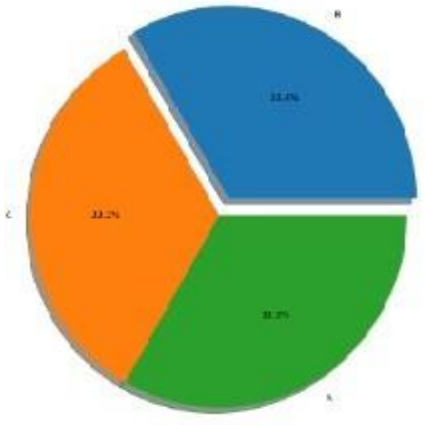


Fig 3 represents No. of students Gender wise in Different Sections in the form of chart and Graph

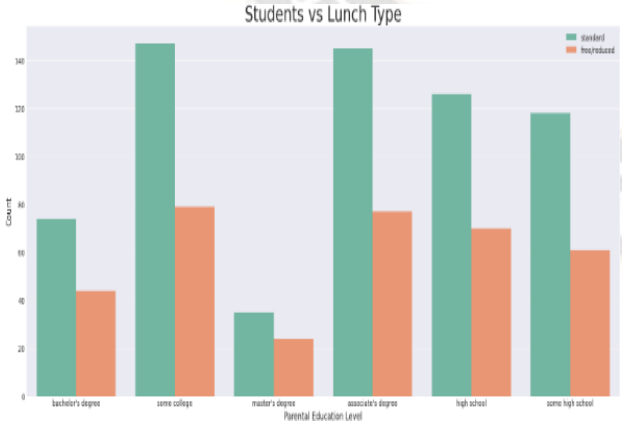


Fig 4 represents a plotted graph between Parental level of education and student vs lunch type

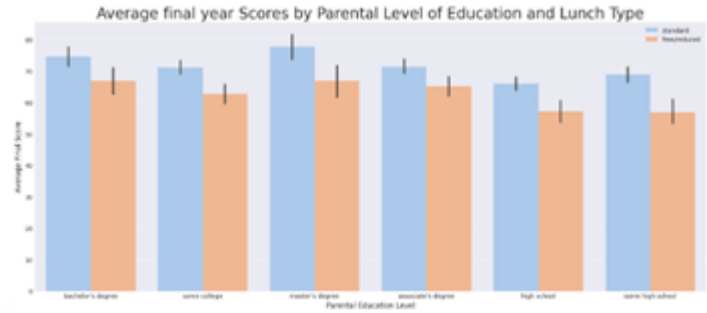


Fig 5 represents bar plot between parental level of education and final year score

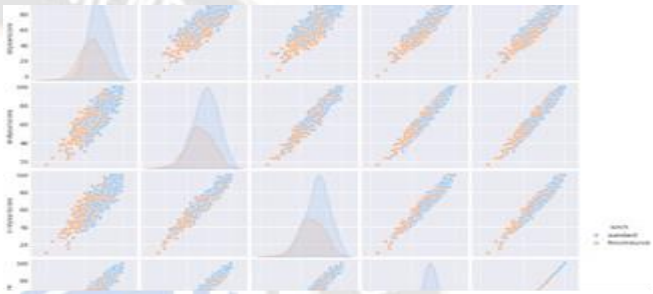


Fig 6 Pairplot of students performance data

lunch	testpreparationcourse	istyearscore	iindyearscore	iiirdyearscore	status	average
standard	none	72	72	74	ist	72.50
standard	completed	69	90	88	merit	82.25
standard	none	90	95	93	outstanding	92.50
ree/reduced	none	47	57	44	fail	49.25
standard	none	76	78	75	merit	76.25

Fig 7 showing records of correlated data attributes

Table 2 Comparison between different Regression Techniques

Regressio n Technique	RMSE	MS E	R2 Score
Linear Regression	0	0	1
Lasso Regression	1.05	0.84	0.99
Decision Tree	0.72	0.09	0.99
Random Forest	0.9	0.09	0.99

Classification Algorithms in the Prediction of Students Performance ", Indian Journal of Science and Technology(IJST).2015;8(15):01-12.

[3] Surjeet Kumar Yadav and Saurabh Pal. "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal (WCSIT).2012;2(2):51-56.

[4] Pal AK, Pal S. "Analysis and Mining ofEducational Data for Predicting the performance of Students", International Journal of ElectronicsCommunication and Computer Engineering(IJECCE).2013;4(5):1560–1565.

[5] Rathee A, Mathur RP. "Survey on Decision Tree Classification algorithm for the Evaluation of Student Performance", International Journal of computers & Technology(IJCT).2013;4(2):244-247.

[6] Aher SB, Lobo LMRJ. Data mining in educational system using Weka. IJCA Proceedings on International Conference on Emerging TechnologyTrends (ICETT). 2011; 3:20–5.

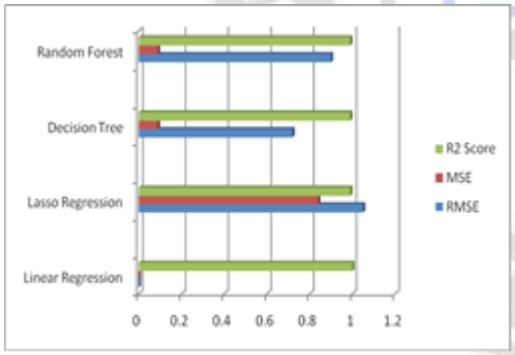


Chart 2 Representing R2 Score, MSE and RMSE Comparison of Regression Technique

CONCLUSION

We can see from the results above that Linear Regression is accurate and is better than Random Forest Algorithm, Lasso Regression and Decision Tree algorithm. Other factors like R2 Score, MSE and RMSE is also better than these algorithms. So it will be concluded that Linear Regression is better and it helps in predicting the students' academic performance effectively. Further implementation is performed by taking Linear Regression algorithm in the future work.

REFERENCES

1. A.K. Pal and S. Pal. "Data Mining Techniques in EDM for Predicting the Performance of Students", International Journal of Computer and Information Technology (IJCIT).2013;2(6): 1110-1116.
2. C. Anuradha and T. Velmurugan. "A Comparative Analysis on the Evaluation of Ogunde AO, Ajibade DA. A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm. Journal of Computer Science and Information Technology. 2014; 2(1):21–46.
3. Shanmuga PK. Improving the student's performance using Educational data mining. International Journal of Advanced Networking and Application. 2013; 4(4):1680–5.
4. Bag Ayan, A Comparative study of Regression algorithms for predicting graduate admission to a university.2020
5. Daud. Predicting Student Performance using Advanced Learning Analytics.Proceedings on International Conference on World Wide Web Companion.2017;415-21.
6. Shahiri & Husain,Rashid.A Review on Predicting Student's Performance Using Data Mining Techniques.2015;72:41 4-22.