Analysis of Student Academic Performance using Regression Methods

S. Kranthi Reddy, S. Poojitha, G. Bhargavi, B. Harika

Abstract**:** ***In the educational industry, student's early performance prediction is important so that strategic intervention can be planned before students reach the final semester. With rapid change in the technology and the lot innovative software, it has become quite convenient to analyze the performance of the student. Machine Learning plays an important role in today's world and it helps the educational institutions to predict and make decisions related to student's performance. The scope of this paper is to predict the student marks through desktop application. In this project, the data of our institute students is taken and regression algorithms are applied to predict the academic status of the student.***

Keywords**:** ***Desktop application, Machine Learning, Regression algorithms, Student's performance***

I. INTRODUCTION

In this competitive world the students face the biggest challenge in terms of their academics. They face a lot of stress and pressure due to increase in cut-offs, parents expectations, etc. [7]. The amount of competition they face is extremely high when it comes to examinations and scores. It is important to predict the academic performance of the students to help them grow in their career.

The main goal of higher education institutions is to present quality education to its students [8]. One way to accomplish the higher level of quality in higher education scheme is by predicting student's academic performance and there by taking early actions to improve student's performance and teaching quality. Constant evaluation of students performance have to be done which helps the lecturer to take necessary actions like more attention to the particular student, teaching in a different way that student can grasp quickly, conducting exams, etc.

The importance of machine learning is rapidly increasing day to day. Machine learning is the next level of the technological advancement [9]. In its true sense, it has a great deal of importance in almost every sector of our lives, including the education system.

Machine learning algorithms are generally classified as unsupervised and supervised algorithms. Supervised algorithms require a supervisor who provides both input data and expected output to train the model. The data used is labeled data. Unsupervised Algorithms does not require any supervisor or expected output to train the model. Here the data is neither labelled nor classified. Hence the data is grouped according to their patterns, characteristics, etc.

Classification model is used to predict the target class for each instance in the data. A predictive model with a categorical target like "yes" or "no", " good" or "bad", "pass" or "fail", etc., uses classification algorithms. The basic type of classification problem is binary classification.

Regression model is used to investigate the relationship between two or more independent variable[x] and estimate one dependent variable[y] based on others. A predictive model with a numerical target uses regression algorithms. The basic type of regression model is linear regression model which finds the best fit straight line through the points.

II. LITERATURE SURVEY

Sadiq Hussian., [1] compared various data mining techniques for analyzing academic performance of students. They collected data of 300 students from three different colleges of Assam. They applied feature selection methods and figured out mainly 12 attributes influence the student performance. Classification methods like J48, PART, Random Forest and Bayes Network classifiers are used among which Random Forest gives maximum accuracy.

Anal Acharya., [2] developed a data model for predicting student results. For this purpose, data is collected from group of student's performance pursuing CSE in some UG colleges in Kolkata. Five classes of machine learning algorithms were applied and best results were observed with the decision tree class of algorithms.

Pedro Strecht et. al. [3] have predicted students' results and grades in their work. They used classification models for the student's results and regression model for the prediction of the grades. They carried out the experiments using 700 courses students' data who studied at the University of Porto. They used decision trees and support vector machine for classification while support vector machine, random forest were best suited for regression analysis. The classification model was able to extract useful patterns, but the model for regression was not able to beat a simple base line.

Anuradha., [4] made a study for analysis of final year results of UG Degree students using data mining techniques. They used decision tree algorithms C4.5, Bayesian classifiers,

Revised Version Manuscript Received on 22 March 2019.

S. Kranthi Reddy, Assistant Professor, Department of CSE, Vignan Institute of Technology and Science, Hyderabad, India.

S. Poojitha, B.Tech (CSE), Vignan Institute of Technology and Science, Hyderabad, India.

G. Bhargavi, B.Tech (CSE), Vignan Institute of Technology and Science, Hyderabad, India.

B. Harika, B.Tech (CSE), Vignan Institute of Technology and Science, Hyderabad, India.

KNN algorithms for classifying the performance and to develop a model of student's performance predictors. They collected data set from there of private colleges in Tamil Nadu state of India. Their range of population varies from 61.75%. The study can be further extended to measure the performance of other classification techniques with large sample data.

Surbhi Agarwal., [5] Uses student's historical record which includes their living habits, backgrounds and so forth are utilized as dataset. The performances predicted using four algorithms which are decision tree, naive bayes, random forest and rule indicator. These analyzed results are explicitly used to predict the upcoming grades of the students which affect the academic performance of students. The prediction of final semester grades of student is generated before the actual final semester exam held. This can help the education institutes to minimize the dropping of the overall results by giving extra attention to the weak students.

Ahmad.,[10] presented a framework to predict the performance of first year bachelor students in Computer Science Course. The data set contains 6 years data from July 2006-2007 upto July 2011-2012. The data was collecting from various features of students records, Navie Bayes Classifier is used to predict the class label as a categorical value. The rule based classifier has obtained the best accuracy of 71.3%.The success level of first year student was predicted. Submit your manuscript electronically for review.

A. Proposed Work

The proposed work tries to predict the performance of student based on previous year's performance. Here data is collected from Vignan Institute of Technology and Science. The dataset contains marks of student belonging to various streams such as CSE, ECE, MECH, CIVIL, etc. The course curriculum is divided into seven semesters as mentioned in Table I. The Performance of the student in every semester is evaluated in terms of percentage. We consider the seven semesters as seven attributes. The dataset contains 1452 instances.

Table I. Dataset Description

Attribute	Type	Description
I	Numeric	First Semester Percentage
II.I	Numeric	Second Semester Percentage
II.II	Numeric	Third Semester Percentage
III.I	Numeric	Fourth Semester Percentage
III.II	Numeric	Fifth Semester Percentage
IV.I	Numeric	Sixth Semester Percentage
IV.II	Numeric	Seventh Semester Percentage

III. APPLIED ALGORITHMS

Our work mainly consists of two steps:

1. Model building
2. Desktop Application

A. Model Building

The dataset consists of seven numerical attributes. Since the output variable is continuous, so the model was build using regression algorithms. Regression technique is used to predict dependent (continuous) variable given a set of independent (continuous or categorical) variables. Regression models gives best results when the dataset follows the below assumptions:

- There should exist linear and additive relationship between dependent and independent variables.
- There should be no correlation between independent variables. Violation of this assumption will lead to Multicollinearity.
- The error terms should have constant variance. Violation leads to Heteroskedestacity.
- The error terms should not be correlated. Presence of correlation is known as autocorrelation.
- The dependent variable and the error terms must have the normal distribution.

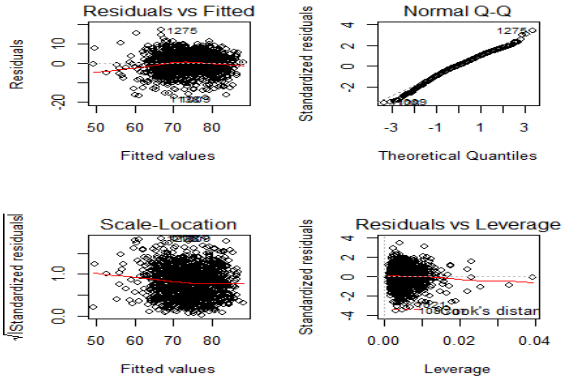


Fig I: Plots representing assumptions of regression model

B. Data Preprocessing

Data Preprocessing is a technique used to convert raw data into clean data. It mainly involves handling of missing values, selection of significant attributes, handling outliers, etc. We can handle missing values by deleting rows, replacing them with mean/median/mode, assigning a unique category or using algorithms that support missing values. Selection of significant attributes is to use the attributes that have most influence on output variable. Initially our data set contains nine attributes with 1452 instances and after performing data preprocessing we have seven attributes with 1296 instances.

C. Specifying the selected algorithms

There are various regression methods: Linear regression, Polynomial, Logistic, Quantile, Ridge, Lasso, Elastic Net, Principle Component, Partial Least Square, Support Vector, Ordinal, Poisson, Negative Binominal, Quasi-Poisson, Cox regression, CART. Specified algorithms were used for predicting the student academic performance

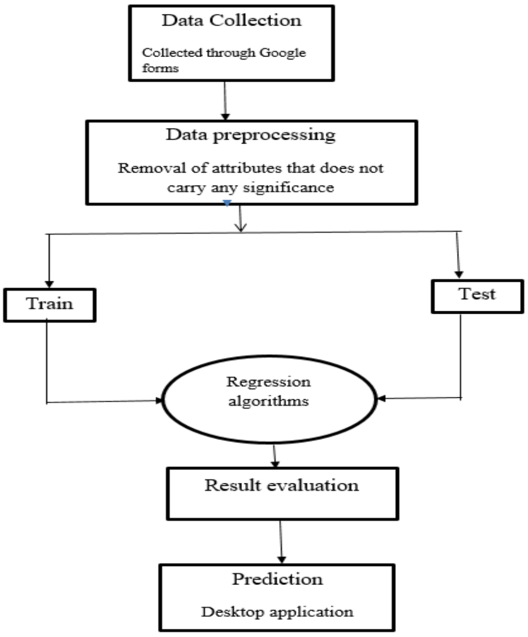


Fig II: Proposed Framework

D. Desktop Application

The application has been developed in Spyder IDE coded in Python 3.7.1. The GUI has been designed using 'tkinter' interface. Taking the six semester marks from this GUI application and passing them to python program that predicts the percentage of seventh semester of the student, then by passing the prediction to the GUI, so that it can be printed out to the user as shown in Fig IV.

IV. EXPERIMENTS & RESULTS

The data is ready for the experiments. Python has developed many modules which help in implementing machine learning algorithms. These modules include numpy, scipy and scikit-learn modules. The data set is splitted into training and testing data (70:30). Linear Regression and Gradient Boosting algorithms are used to build the model. The training set is used to train the model, while testing set is used to test the correctness of the model.

CART: A Classification and Regression tree is introduced by Leo Breiman to indicate decision tree algorithms that can be used for predictive modeling problems.

XGBoost: XGBoost is an ensemble learning method. Ensemble learning offers a systematic solution to combine the predictive power of machine learners. XGBoost was created by Tianqi Chen. It is used for supervised machine learning problems. XGBoost belongs to a family of boosting algorithms that convert weak learners into strong learners. It is an optimized distributed gradient boosting library.

Ridge Regression: Ridge regression is a technique used when the independent variables are highly correlated. Ridge regression reduces the size of coefficients by using l2 norm thereby reducing high complexity in the model. It helps us in over fitting and dealing with outliers.

Lasso Regression: Lasso (Least Absolute Shrinkage and Selection Operator) is similar to ridge regression. This is a regularization method and uses l1 regularization. If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero.

A. Result Evaluation

There are certain metrics used to evaluate the regression model:

1. R square - explains the percentage of variance explained by covariates in the model. It ranges between 0 and 1.
2. Adjusted R^2 – It does not increase unless the newly added variable is truly useful.
3. F Statistics – Evaluates the overall significance of the model. Higher the F Statistics, better the model.
4. RMSE/MSE/MAE – Error metric is the important evaluation metric. Since all these are errors, lower the value, better the model.

Table II. Comparison between applied algorithms

Model	Training Set		Testing Set	
	RMSE	R^2	RMSE	R^2
Linear	5.0761	0.5669	4.8376	0.5974
Gradient Boosting	4.8051	0.6103	4.7983	0.6028
XGBoost	2.3405	0.91	6.1927	0.34
Ridge	5.0761	0.5669	4.8376	0.5974
Lasso	5.0775	0.5667	4.8433	0.5965

If the RMSE values for train and test data are similar, then we can say that the built model is good. From the Table II, we can say that Gradient Boosting Algorithm best fits the given data.

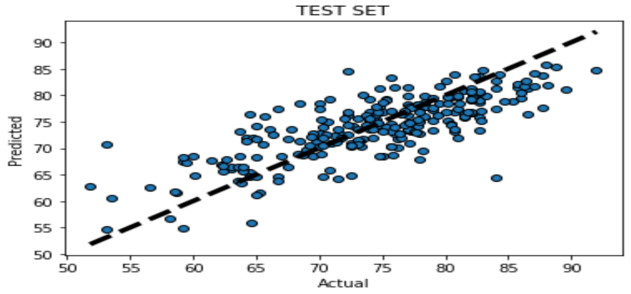


Fig III (a)

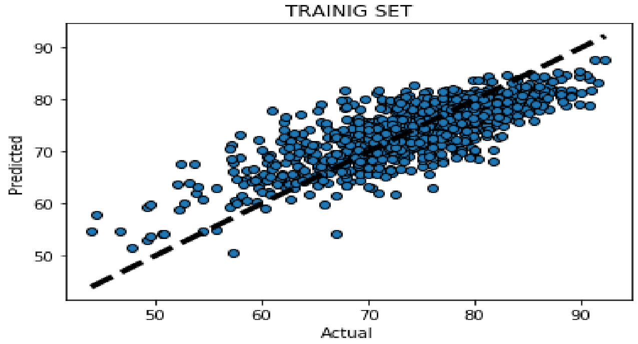
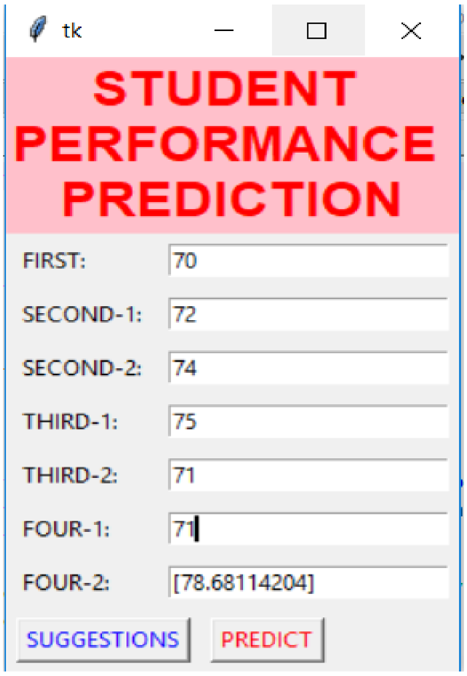


Fig III (b)

Fig III. Goodness of fit with predictions being visualized by the line.



STUDENT PERFORMANCE PREDICTION

FIRST: 70

SECOND-1: 72

SECOND-2: 74

THIRD-1: 75

THIRD-2: 71

FOUR-1: 71

FOUR-2: [78.68114204]

Fig IV. Desktop Application

V. CONCLUSION

The proposed work entitled “ANALYSIS OF STUDENT ACADEMIC PERFORMANCE USING REGRESSION METHODS” is a Python tool based application. This provides the facility of predicting the performance of the student early before they reach their end semester. It helps the student’s to improve their performance. CART and Gradient boosting regression algorithms provide the accuracy of 60%. The proposed work provides a desktop application for predicting the end semester percentage.

REFERENCES

1. Sadiq Hussain; Educational Data Mining and Analysis of Students Academic Performance using WEKA; Indonesian Journal of Electrical Engineering and Computer Science, Volume 9, No. 2, February 2018.
2. Anal Acharya and Devadatta Sinha; Early Prediction of Student Performance using Machine Learning techniques; International Journal of computer Applications(0975-8887) volume 107, No. 1, Dec 2014.
3. Strecht, P. et al; A Comparative study of Classification and Regression Algorithms for modelling Students Academic Performance; Processing’s of the 8th International Conference on educational data mining; 2015, p.3
4. Anuradha, C. and T. Velmurugan; A Comparative Analysis on the evaluation of classification Algorithms in the prediction of students performance; Indian Journal of Science and Technology, 2015, 805, p.12.
5. Surbhi Agarwal; Using Data Mining classifier for predicting student’s performance in UG Level; International Journal of computer Applications(0975-8887) volume 172, No. 8, August 2017.
6. Ahmad. F, N.H. Ismail; the Prediction of Student’s Academic Performance using classification data mining techniques; Applied Mathematical Sciences, 2015, p.12.
7. Felicia Nazareth, What are the challenges you face as a student and how to overcome them? Available: http://www.alignthoughts.com/what-are-the-challenges-you-face-as-a-student/
8. Will McGuinness, The Benefits and the Limitations of Machine Learning in Education. Available: https://www.gettingsmart.com/2018/02/the-benefits-and-the-limitations-of-machine-learning-in-education/
9. A.I. and Machine Learning and its impact on Education Technology. Available: https://theknowledgereview.com/machine-learning-impact-education-technology/