Student Performance Prediction Using Machine Learning Algorithm Logistic Regression And K-Nearest Neighbor Algorithm

V J Abinaya[1] , K Sindhu[2]
[1, 2] Dept of Communication Systems
[1, 2] Bethlahem Institute Of Engineering, Karungal, India

Abstract- *Education system is very old and due to a large population of students, there are some serious issues in analyzing and predicting students' performance. Every institution has its own set of standards for evaluating student success. In existing system there is no proper procedure for monitoring and analyzing a student's performance and progress. In this paper, focus is given on additional external factors like students status, parent occupation etc. That will be more effective in visualizing and analyzing student's performance. This project proposed two types of machine learning algorithms used like K-Nearest Neighbor (KNN) Algorithm and Logistic Regression algorithm. Performance analysis is also done for these two algorithms based on accuracy level of results as well as with some existing work.*

Keywords- KNN, Logistic Regression, Predication

I. INTRODUCTION

Education is an ongoing process and guidance is a key to raise the level of qualification.. In higher institutions, main objective is to improve quality of education which can be achieved by good prediction of student's success chances,[1]. This predication can be managed by identifying the source of problem which can be family issue, health issue etc,[2]. These factors plays a major role in student's academic performance. A very good student may not perform well in exams if he or she is suffering from these type of issues.
The objective of this paper is predict students performance based on several inputs including academic and non-academics factors. For this, two different machine learning algorithms K-Nearest Neighbor and Logistic Regression are used.

II. EXISTING SYSTEM

In existing system used various analytical methodologies to predict student achievement where most of the researchers have used grade points as the assessment process. There is no proper procedure for monitoring and analyzing a student's performance and progress. Difficult to determine the optimal prediction methodology for visualizing student academic growth and performance and affect students' academic performance and achievement. Main disadvantage is such case data points cannot be classified.Accuracy level is low.

III. PROPOSED SYSTEM

In this project proposed, some external factors along with academic grades are also considered for predicting student performance such as family background like mother job, student status, students results who receiving higher education etc.Using two different machine learning algorithms.KNN classifier and Logistic Regression are applied dataset. Performance analysis is also done for these two algorithms based on accuracy level of results as well as with some existing work. Advantages are easy to understand and realize, efficient, time complexity, good accuracy.Fig. 1 expalained about system architecture.

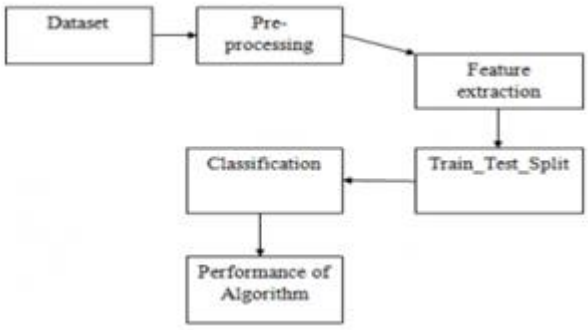

Fig.1. System Architecture

IV. METHODOLOGY

The methodology adopted in this paper comprises of four stages: Data Acquisition, Data Preprocessing and Feature Selection, Data Visualization,Classification and Prediction. This prediction class can be one of the following: higher chances of getting pass and lower chances of getting pass. Fig. 2 shows data flow diagram.

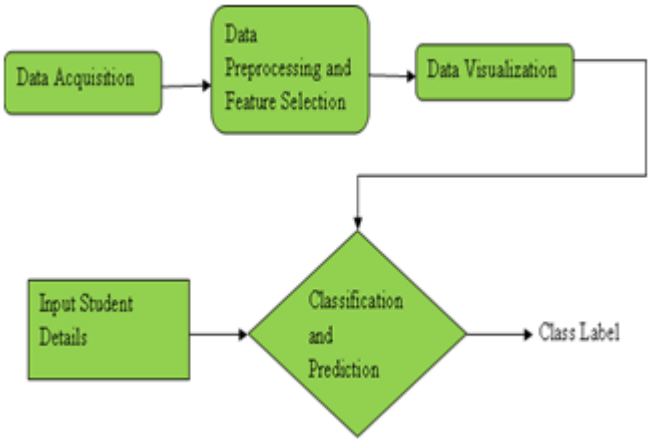

Fig.2.Data Flow diagram

A. Data Acquisition

This dataset containing data from different geographical areas. Fig. 3 shows datasets.

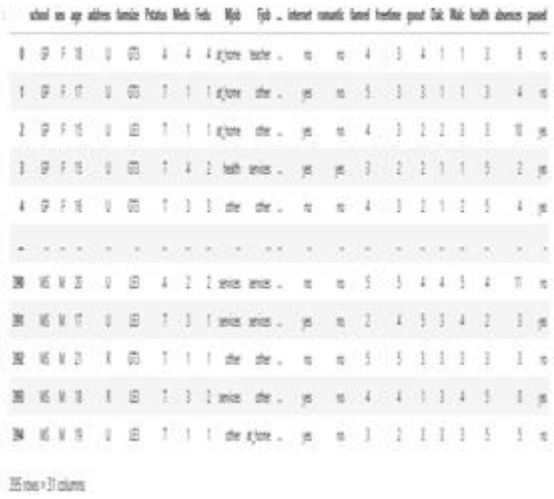

Fig.3.Datasets

B. Data Preprocessing and Feature Selection

After taking data, pre-processing is done to check if whether it contains any missing values or not. and to find correlation between attributes.Fig.4 shows data preprocessing.

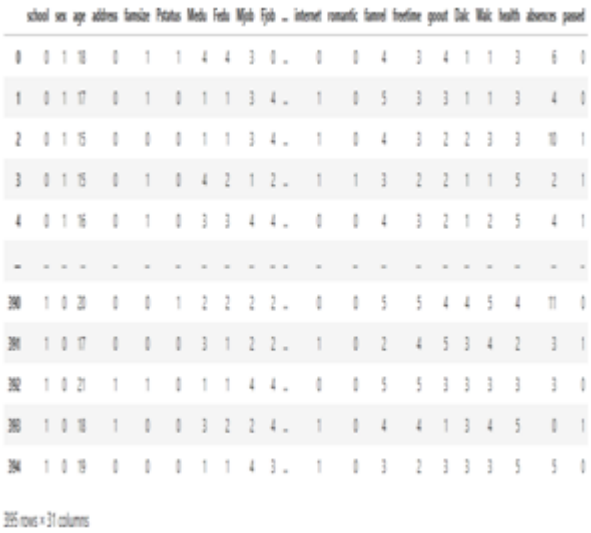

Fig.4.Data Preprocessing

C. Data Visualization

Visualization techniques are used to understand the patterns and analysis the data to understand the relation and dependencies between the data attributes. In this study, the categorical label values are converted into numerical values. Fig. 5 shows data visualization., Fig.6 shows describing students mother job

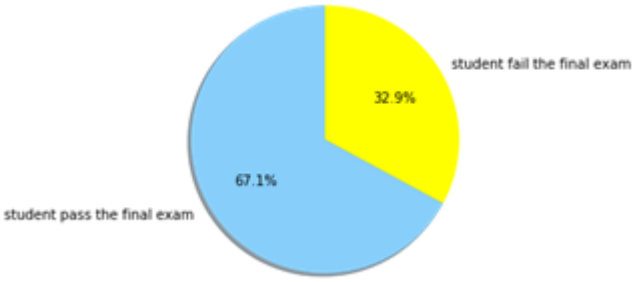

Fig.5.Data Visualization

D. Classification and Prediction

Two supervised machine learning algorithms KNN and Logistic Regression were applied for classification and results are obtained.

a) Applying K-Nearest Neighbor (KNN) Algorithm

KNN is a supervised machine learning algorithm which is used for the problem based on classification and regression. In the proposed model, KNN algorithm is applied to classify the students in PASS or FAIL category. This method stores the data and whenever new data point comes, it classifies the new data point based on similar features for that it selects the K value and find nearest Euclidean Distance of K- neighbors.

b) Applying Logistic Regression algorithm

Logistic Regression algorithm is also used in the proposed work for classification. It is basically used for predicting categorical dependent variables using a given set of independent variables. Output of Logistic Regression is discrete value, and it gives the probabilistic value between 0 and 1. Fig.7 shows describing students status.

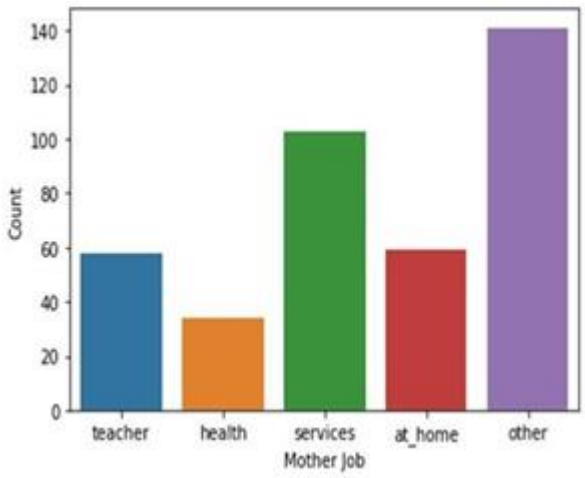

Fig.6. Describing Students Mother Job

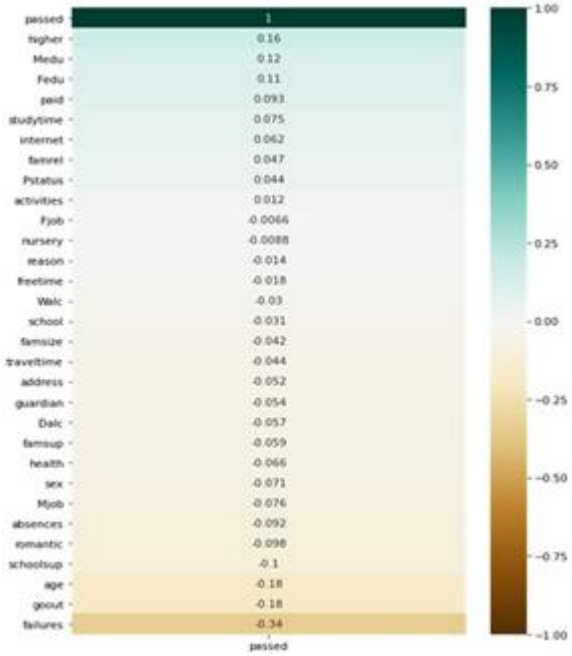

Fig.7. Describing Students Status

V. RESULTS AND DISCUSSION

In this section, analysed that performed to Logistic Regression algorithm better as compared KNN classifier. By applying Logistic Regression, 80.97% of accuracy was achieved, whereas by applying K-Nearest Neighbors on dataset 78.15% of accuracy was achieved. Fig. 8 shows KNN accuracy,Fig. 9 shows Logistic Regression and Fig.10 shows the comparison of both models.

```
*Accuracy is: 78.15126050420169
*f1 score is:  0.7102996254681648
random_state is  71027464
```

```
*the ROC curve:
```

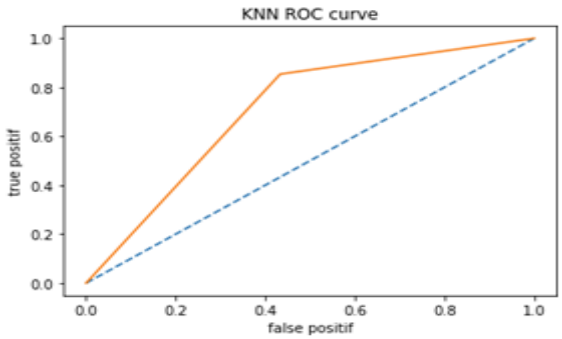

Fig. 8. KNN Accuracy

```
*Accuracy is: 80.67226890756302
*f1 score is:  0.7408389357068459
```

```
*the ROC curve:
```

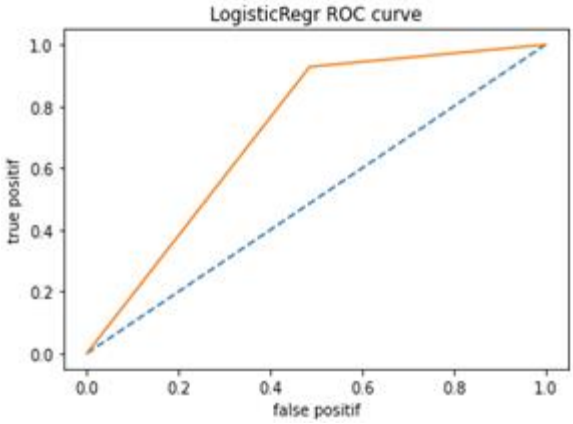

Fig .9. Logistic Regression Accuracy

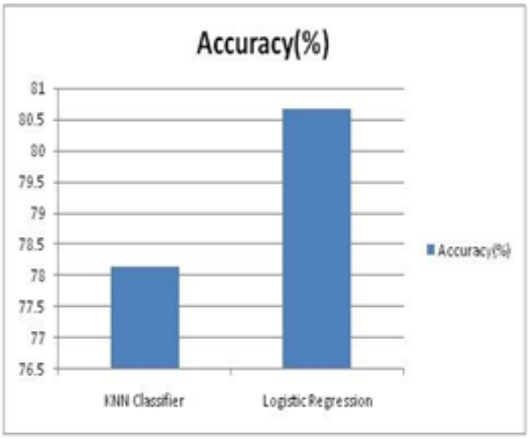

Fig.10. Bar Chart For Comparison-Based Evaluation

VI. CONCLUSIONS

Proposed model is also compared with some existing models to show the efficiency of the proposed work. This comparison proves that proposed model can be used efficiently to examines student performance by extracting more insights from the data. As well as to start introducing some beneficial monitoring systems that educationalists can use to precisely assign grades and assist them in improving their students' performance.

REFERENCES

[1] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana and E. Wani, "Prediction of student academic performance by an application of data mining techniques," in International Conference on Management and Artificial Intelligence IPEDR, vol 6, no 1, 2011.

[2] H. Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq, H. Badukhen, S. Alrashed, J. Alhiyafi and S. O. Olatunji, "Student performance prediction using support vector machine and k-nearest neighbor," in 30th Canadian conference on electrical and computer,IEEE, pp: 1-4, 2017

[3] B. Sekeroglu, K. Dimililer and K. Tuncal, "Student performance prediction and classification using machine learning algorithms," in 8th International Conference on Educational and Information Technology, pp. pp. 7-11, 2019.

[4] A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom and S. N. Liao, "Predicting academic performance: a systematic literature review," in 23rd annual ACM conference on innovation and technology in computer science education, pp: 175-199, 2018.

[5] L. Mortada, J. Bolbol and S. Kadry, "Factors Affecting Students' Performance a Case of Private Colleges in Lebanon," Journal of Mathematical and Statistical Analysis, vol. 1, no. 1, pp. 105-110, 2018

[6] A. M. Shahiri, W. Husain and N. A. Rashid, "A Review on predicting student's performance using data mining techniques," in Procedia Computer Science 72 , pp: 414-422., 2015.