Prediction of Students Performance using Machine learning

J. Dhillapan¹*, N. Vijayalakshmi², S. Suriya³, Arockiya Christopher⁴

¹Professor & Head, Department of Computer Applications, SRM IST, Ramapuram Campus.
², ³Assistant Professor, Department of Computer Applications, SRM IST, Ramapuram Campus.
⁴Student, Department of Computer Applications, SRM IST, Ramapuram Campus.

*Email: hod.mca.rmp@srmist.edu.in

Abstract. An enormous measure of computerized information is being produced over a wide assortment in the field of data mining strategies. The creation of student achievement prediction models to predict student performance in academic institutions is a key area of the development of Education Data Mining. A prediction system has been proposed by using their 10th, 12th and previous semester marks. The study is evaluated using Binomial logical regression, Decision tree, and Entropy and KNN classifier. In order to attain their higher score, this framework would assist the student to recognize their final grade and improve their academic conduct.

1. Introduction
Due to the huge amount of data in educational databases, predicting the performance of students has become more difficult. The shortage of an established framework for evaluating and tracking the success of students also isn't currently being considered. There are two primary reasons why such kind of occurring. First, the research on existing methods of prediction is still insufficient to determine the most appropriate methods for predicting student performance in institutions. Second, is the absence of inquiry of the specific courses. [4-9]

The real goal is to have an overview of the systems of artificial intelligence that were used to predict academic learning. This research also focuses on how to classify the most relevant attributes in student data by using prediction algorithm. Using educational machine learning methods, we could potentially improve the performance and progress of students more efficiently in an efficient manner. Students, educator and academic institutions could benefit and also have an impact.

2. Related Work
An automated evaluation system has been proposed to evaluate student performance and to analyze the student achievement. Here the author uses tree algorithm for predicting student performance accurately [2-3]. In the proposed system Education Data Mining (EDM) is used for the classification. Clustering data mining technique is used for analyzing the large set of student database. This technique will speed up the searching process and the also yield the classification result more accurately.

A new Learning model has been proposed by using the student information from the college registration. The final dataset is provided as input so ML algorithms which can apply and predict student's academic performance.[11] They selected 13 algorithms from 5 categories of ML they are Naive Bayes, SVM, MLP, IBK, Rules and tree.
A comparative study on supervised learning for student prediction has been proposed. The author handles with 14 feature set for classification. The tools used for classification are: KNN, Decision tree, Navie Bayes.[13-15]

Psychometric analysis of the student behaviour has been proposed by using intellectual parameters of the student which affect their study. Various mining techniques are used to determine the educational data covering psychological factors. The accuracy rate of the previous study is 89% but by using the proposed system the accuracy rate has increased to 90%. Here the author used Radical Basis Function Kernel to produce higher accuracy.

3. System Architecture

![System Architecture Diagram]

**Figure 1.** System Architecture
The above figure clearly explains the system architecture of the proposed system. The steps involved in the proposed system are:

- Collecting data about users and the behavior.
- Transforming the data into features.
- Train the model.
- Deploy the trained model.

4. Methodology Used in the Proposed System

4.1 Decision Tree

Decision Tree classifier is the regression model which is represented in the form of tree structure. The purpose of Decision Tree classifier is to breakdown the dataset into smaller subset. The tree consists of decision nodes and leaf nodes.[19] In our proposed architecture the attribute which delivers maximum information will act as a decision node. The node which is present as the top most of the decision node acts as a predictor which is called as root node. The node which cannot be further divided is known as leaf node. The steps involved in the decision tree are specified below:

- Process 1: Start the root.
- Process 2: Perform the test.
- Process 3: Follow the edges corresponding to the outcome.
- Process 4: go to step 2 until reaches leaf node.
- Process 5: Predict the outcome associated with the leaf.

A. K-Nearest Neighbour

K-Nearest Neighbour is one of the basic and essential classification algorithms in machine learning. It is non-parametric and makes any underlying assumptions about the distribution of data. The steps involved the KNN is listed below:

- File the training data in a sample points array.
- The Euclidean distance measure.
- Make the least distance range available.

B. Linear Regression

Logical regression is the classification technique which handles with the threshold value. The following arguments are used for determining the threshold value:

- Low Precision / High Recall
- High Precision / Low Recall

Based on the number of categories the Logical regression can be classified as:

- Binomial
- Multinomial
- Ordinal

5. Working of the Proposed Model

The proposed system handles with the student individual mark that include 10th, 12th mark and there semester mark. The prediction of the system has the following task:

Case 1: The student who have secured below 50 percentage in their 10th and 12th.
Case 2: The student who have failed in internal.

Case 3: The student having less attendance percentage / irregular.

On the above three cases if any one achieved the student may not complete his/her degree successfully. By this scenario the system has been implemented.

6. Result and Discussion

The Figure 2 presents the login page of the proposed system. The outcome of the prediction is shown below:

- Using Binomial Logistic Regression

Estimated regression coefficients: 
\[
[-5.93085499 3.65187957 -4.19083559 -0.34248963 4.48217135 \\
-0.04329762 4.79142356 6.05719461 6.51001631 -0.6430665 ]
\]

No. of iterations: 1735

Confusion Matrix: 
\[
[[7 4] \\
[1 158]]
\]

Accuracy : 97.05882352941177

Report : precision recall f1-score support
Predicted Value : will complete

- **Using Decision Tree**

Results Using Gini Index:

Predicted values:
\[
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 1. 1. 1.
  1. 1. 1. 1. 1. 1. 1. 1. 1.]
\]

Confusion Matrix: 
\[
\begin{bmatrix}
0 & 3 \\
1 & 30
\end{bmatrix}
\]

Accuracy : 88.23529411764706

Report :
| precision | recall | f1-score | support |
|-----------|--------|----------|---------|
| 0.0       | 0.00   | 0.00     | 3       |
| 1.0       | 0.91   | 0.97     | 31      |

accuracy 0.88 34
macro avg 0.45 0.48 0.47 34
weighted avg 0.83 0.88 0.85 34

- **Using Entropy**

Predicted values:
\[
[1. 1. 1. 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
  1. 1. 1. 1. 1. 1. 1. 1. 0. 1.]
\]

Confusion Matrix: 
\[
\begin{bmatrix}
1 & 2 \\
1 & 30
\end{bmatrix}
\]

Accuracy : 91.17647058823529

Report :
| precision | recall | f1-score | support |
|-----------|--------|----------|---------|
| 0.0       | 0.50   | 0.33     | 3       |
| 1.0       | 0.94   | 0.97     | 31      |

accuracy 0.91 34
macro avg 0.72 0.65 0.68 34
weighted avg 0.90 0.91 0.90 34

Predicted Value : will complete
• **Using K-Near Neighbor**

Using K-Near Neighbor:

```
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
```

Confusion Matrix:

```
[1 2]
[1 30]
```

Accuracy: 93.7167508223259

Report:

| precision | recall | f1-score | support |
|-----------|--------|----------|---------|
| 0.0      | 0.00   | 0.00     | 3       |
| 1.0      | 0.94   | 0.97     | 31      |

Accuracy: 0.93

Support: 34

Predicted Value: will complete

The accuracy of the proposed system using the machine learning technique are specified in the below table.

**Table 1. Accuracy of the Proposed System using Different Technique.**

| Technique               | Accuracy |
|-------------------------|----------|
| Binomial Logical regression | 97.05    |
| Decision Tree           | 88.23    |
| Entropy                 | 91.19    |
| K-NN                    | 93.71    |

**Figure 3. Accuracy of the student prediction system**
The figure 3 shows the accuracy level of the student prediction system. The dataset which is gathered are analyzed by using four techniques. The accuracy of binomial logical regression is 97.05%, Decision Tree 88.23%, Entropy 91.19% and K-NN 93.71%.

7. Conclusion
The paper focuses on the student academic growth analysis using machine learning techniques. For analysis Binomial logical regression, Decision tree, Entropy and KNN classifier are used. This process can help the instructor to decide easily about performance of the students and schedule better method for improving their academics. In future additional features are added to our dataset to acquire better accuracy.

Reference
[1] N.V.Krishna Rao, Dr.N.Mangathayaru, Dr.M.Sreenivasa Rao,” Evolution and Prediction of Radical MultiDimensional E-Learning System with Cluster based Data Mining Techniques”, International Conference on Trends in Electronics and Informatics,2017, PP.701-707.
[2] Pushpa S.K, Manjunath T.N, “Class result prediction using machine learning”, International Conference On Smart Technology for Smart Nation,2017,pp1208-1212.
[3] Micheal Bowles, Machine Learning in Python: Essential Techniques for Predictive Analysis. John Wiley & Sons, Inc. 2015.
[4] Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. Project Adam: Building an efficient and scalable deep learning training system. In 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14), pages 571–582, 2014. www.usenix.org/system/files/conference/osdi14/osdi14paper-chilimbi.pdf.
[5] Jack Clark. Google turning its lucrative web search over to AI machines, 2015. www.bloomberg.com/news/articles/2015-10-26/google-turning-its-lucrative-web-search-over-to-aimachines.
[6] J. Xu, K. H. Moon, and M. Van Der Schaar, “A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs,” IEEE J. Sel. Top. Signal Process., vol. 11, no. 5, pp. 742–753, 2017.
[7] A. M. Shahiri, W. Husain, and N. A. Rashid, “A Review on Predicting Student’s Performance Using Data Mining Techniques,” in Procedia Computer Science, 2015.
[8] Y. Meier, J. Xu, O. Atan, and M. Van Der Schaar, “Predicting grades,” IEEE Trans. Signal Process. vol. 64, no. 4, pp. 959–972, 2016.
[9] P. Guleria, N. Thakur, and M. Sood, “Predicting student performance using decision tree classifiers and information gain,” Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2014, pp. 126–129, 2015.
[10] K. P. Shaleena and S. Paul, “Data mining techniques for predicting student performance,” in ICETECH 2015 - 2015 IEEE International Conference on Engineering and Technology, 2015, no. March, pp. 0–2.
[11] Z. Alharbi, J. . Cornford, L. . Dolder, and B. . De La Iglesia, “Using data mining techniques to predict students at risk of poor performance,” Proc. 2016 SAI Comput. Conf. Series, pp. 523–531, 2016.
15] I. E. Livieris, K. Drakopoulou, and P. Pintelas, “Predicting students’ performance using artificial neural networks,” Proc. 8th Pan-Hellenic Conf. “Information Commun. Technol. Educ.,” pp. 28–30, 2012.

16] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, “Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm,” Procedia Technol., 2016.

17] P. Mohd Arsad, N. Buniyamin, and J. L. Ab Manan, “Neural Network and Linear Regression methods for prediction of students’ academic achievement,” IEEE Glob. Eng. Educ. Conf. EDUCON, no. April, pp. 916–921, 2014.

18] F. Sarker, T. Tiropanis, and H. C. Davis, “Linked data, data mining and external open data for better prediction of at-risk students,” in Proceedings - 2014 International Conference on Control, Decision and Information Technologies, CoDIT 2014, 2014.