Student Performance Prediction using Technology of Machine Learning (FEAT-HUNCH)

Rahul Sharma[*1] Satyam Kumar Maurya [*2], Kaushal Kishor [#3],
[*1] *Student ABES Institute of Technology, Ghaziabad, U.P, India*

[*2] *Student ABES Institute of Technology, Ghaziabad, U.P, India*

[#3] *IT Department, ABES Institute of Technology, Ghaziabad, U.P, Indi*

***[1]rahulkaushik9491@gmail.com, [*2]satyamsam154@gmail.com*, [#3]kaushal.kishor@abesit.in*,**

ABSTRACT

In the given paper, the main focus of this report is education. Student performance prediction is our main target. Various factors have been taken into account to create a model used for student performance prediction. This helps to analyze the student's study environment so that his success rate increases in the field of studies. Our project makes use of various effective machine learning algorithms for creating the predictive model. Mainly, it is based on linear regression, decision trees, naïve Bayes classification, K nearest neighbours (KNN), and some improvements carried out through feature engineering that modifies the data to make it easier in understanding for ML. Data sets containing students' information are arranged in a tabular format. The row represents the name of the student, while each column contains different details about the student such as his background of the family, sex, any information about medical reports, and age. An additional column contains the variable of success rate that the algorithm is trying to predict. The final report is evaluated through these algorithms in which a function outputs whether the student can be successful or not. "FEAT HUNCH – STUDENT PERFORMANCE PREDICTOR in ML" aims at connecting all the students and teachers in an institute.

Keywords— *Student performance prediction, Naïve Bayes, ANN: Logistic Regression, Decision Tree*

1. INTRODUCTION

Education could be an important issue for achieving financial progress. The Student's scholastic performance focuses on different aspects, creating analysis little bit difficult. In upcoming years, there has been a rise within the percentage in rate of interest and concern over individuals within the use of data mining for analyzing academic qualities [12]. Data processing depicts growing and upcoming areas of researches in education and it has separate discrete needs that some fields lack. During this project, the performance analyses of scholar's are mentioned. The goal at of this project is providing students' performance using given strategies through different algorithms. [7].

A lot of studies in this field are that investigate the ways for applying techniques associated with machine learning in educational fields. It focuses on identifying high-risk students and also student performance. [9].

The concept of ML is that a system is predicting from knowledge. ML functionalities vary; its common function is the same as all its applications. The computer analyzes large amounts of data, and finds the definitions and information encrypted in data. [8] These definitions and rules are calculus in nature, and can be feasibly used for defining and processing by computer.

This report estimates the efficiency of the different algorithms and methods involved in ML. The linear regressions model, classification of Bayes', and tree structure like decision trees are mainly implemented for this paper because there all other many relevant algorithms used for designing predictive models. A method known as feature engineering is evaluated upon its improvements and then it suggests the data which can be edited and designed in such a way that it becomes easier in understanding for ML. In this technique (machine learning) category rate is in many cases a decisive property expected by designation. This efficiency standard is considered less frequent, but arises from the specification that a classifier should use only the appropriate time and memory for training and application provided by Gaga in the year 1996.

Data processing depicts growing and upcoming fields of researches in education and it has separate other needs that other field's do not possess. During this paper review, the performance is analyzed for scholars are mentioned.

2. LITERATURE REVIEW:

The study conducted by S. Kotsiantis [10] for prediction of dropouts in distance learning using machine learning is or of the first studies to examine the learning technology. The main powerful donation through this study was a primary and rounded the track for many such researches. While this ML technique has already been imposed on other settings, it was first time being implied on academic environment.

Hardwar and Pal conducted a survey across India in which students were shown to be the most influential factor which was most important. This performance took place in Faizabad. In the life of a student, he used the Bayesian classification to study it [11].In the paper "Data Mining Approach for predicting performance of a student" published by E. Oamanbegovic ct al. they implemented 3 machine learning algorithms with supervision for prediction of success in a course. It was found out that Naïve Bayes' classified performed way better in decision tree prediction and other methods of neural network. Snehal Kekane published "Automatic predicting performance of a

student and Automatic Student Performance Analysis and Monitoring" that proposed a system to display results of the students' performance through the user and helps in releasing staff pressure.

In 2015 the paper titled 'Student Performance Analysis System' disseminated by Chew Li 8A Et Al employed some characteristics in the upcoming given framework during its phase of implementation. Further performance in user interface predicted. There were some of them who ensured that the goal is achieved. In the year 2013,V. Ramesh published in this paper he analyzed the student perform testimonial and data mining point of view there he exhibited a survey cum experimental method used to create a database using the primary as well as secondary data source. It has been proved here that multilayer perturbation is considered the best performing qualifier for predicting student performance.

The research conduct by Erkan Er [6] upon was based on S. Kotsiantis along with other stuff studies. They conclude that the Naïve Bayes base algorithm actually performs better than any other machine learning algorithm. However, this research helped to evaluate time learning characteristics that prevent the process of learning of ML and therefore it was recommended to keep out of the research completely. It was then evaluated that demography of students was lower in characteristics than the previous attending rate and homework rate of students to forecast rates in earlier stages.

3.EXISTING METHOLOGY/ARCHITECTURE

3.1 Naïve Bayes Algorithm

The Classification method used is one of the most regularly studied problems by the researchers based on ML and data mining. It involves predicting values of attributes based on other attribute values. Different classification methods are used.

Bayes classification is the algorithm in which the Bayes law of conditional probability is based on the algorithm. The Bayes rule is a technique for estimating the probability of a feature in which a set of data is shown as input, the **Bayes'** rule or the **Bayes'** theorem-

Another recently made method of classification is artificial neural network. These networks are designed after the human nervous system and are not affected by any other algorithm, so more powerful is affected.

The usage is complex, but they can understand non-linear patterns in dataset.

However this classified data in binary group seems inadequate. The main goal of this study is not only to improve the performance of the students but also to find the students who are at risk; it can be very useful to classify the students according to the performance at different levels. Thus training can arrange efficient result for every student. That is our objective.

Naive Bayesian classification is based on Bayes' in which habitual self-sufficiency occurs among future speakers. A Naive Bayes model is easily constructed with no complex hierarchical parameter decision making it particularly useful for very large data sets. in the face of its simplicity, the Navy Bayes Classifier frequently performs amazingly well and is generally used because it frequently outperforms more worldly-wise classification methods. Bayes theorem provides a method of calculating the following probability, P (c|x), from P(c), P(x), and P (x|c) [3].Provides a way to compute Navy Bayes classifier assuming that the effect of the value of one predictor on a given class is independent of the values of other future speakers. This assumption is called class conditional independence. [9]

Bayes' Theorem-:

Bayes theorem is a mathematical equation that is found in the probability that any one event occurs before it occurs. Mathematical equation of Bayes' Theorem is:

$$P(A/B) = [P(B/A)\,P(A)] / [P(B)]$$

In which A and B make for occurrences

P (B)! = 0.

Actually, we tried to search for prospect of occurrence A given the occurrence B is true. The case B is also defined as evidence.

P [A] is the priori of A (the prior prospect, i.e. Prospect of event before authentication is found). The authentication is a random figure of a sample which is unknown like here it is case B

P (A|B) is a posteriori prospect of B, that is, prospect of occurrence after authentication is evaluated.

3.2 Decision Tree:

The decision tree is a flowchart-like structure. Just as a tree has leaves, roots and branches, similarly it has leaf nodes and branches. The topmost node in the decision tree is called the root node, with each leaf node representing a class. A decision tree is used for decision making. Most of the decision trees are binary. That is, a node has only two children.

Pseudo code of algorithm used in decision tree: -

- The most important feature of data set is employed as the root of the tree.
- The sets used in training are now classified in into subsets. They are designed in such a way that the attribute contains same data as each subset has.
- Finally step 1 and step 2 are repeated on every subset again till the leaf nodes are found in branch of tree.

3.3Neural Network:

The artificial neural network is mostly called just a neural network. One is the model that is inspired by biological neural networks. A network of systems consists of a purse of man-made construction and it modifies the information using a connection-oriented approach in a communication way. Various inputs are transferred through the network and their specific output is taken out after passing the hidden layer as shown in figure.

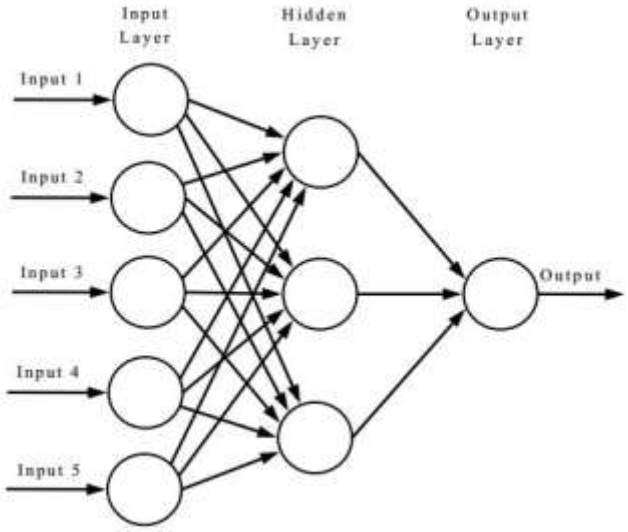


Fig1: Neural Network representation with Hidden element

3.4 Existing methodology feature:

Bhardwaj and Pal [11] selected 300 student through 5 different college that applied in Computer Application bachelors course of Or. R. M. L. Awadh University, Faizabad, India for their study for student performance .By using Bayesian classification formula on 17 features given, it was resulted in such that the functions such as students' marks in higher secondary exam, staying place, teaching habits, parents qualifications, other activities of pupil, income of family and their status are related to students' performance in academy.

In the existing study, the values whose credibility values were higher than 0.70 were taken due to understandings and overpowering influencing values with big credibility values. This feature was used for projection of sample construction. The selection of variables and projection construction of model, the publisher has used MATLAB [9].

It was concluded the next high possibility for student good performance is their home place and the third is teaching. Like in UP native language is Hindi. Hence students are more comfortable in Hindi and less in English.

3.5 Existing methodology Singularity:

Erkan Er managed a study that [6] confirmed important for the uniqueness in projected application. His/her work finish that all exiting machine learning applications in academies were on prediction of dropping out rates for distance learning courses such that no applications that attempts prediction of student performance.

4. PROPOSED METHODOLOGY/ ARCHITECTURE

The diagram, shows in figure 1 the main steps and components of the proposed machine learning system.

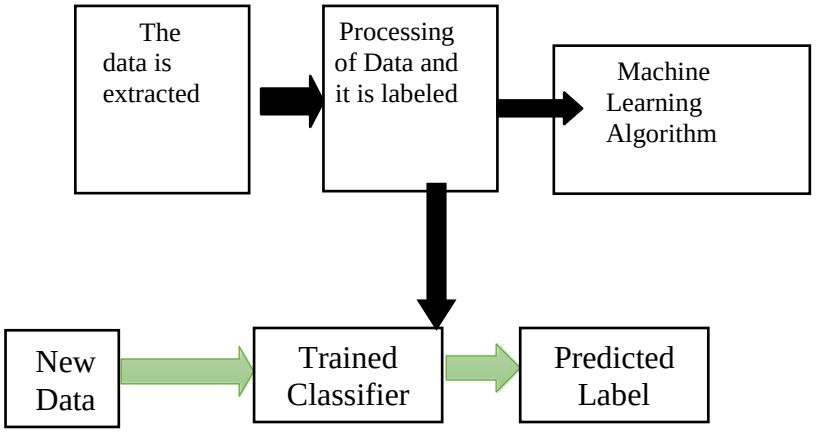


Fig2: Steps and Components of the Proposed System

- Firstly, the data sources provide the data for collection. The other data sources include surveys and the results of students.
- In the second phase, data is processed for getting the normal dataset and getting the rows labeled.
- The next step involves containing the previous step results and the training dataset to the machine learning algorithm.
- Now the ML Algorithm is ready, it designs a model based on the training data to verify the model through the given data used in test.
- Finally the ML algorithm has developed a trained model or a classifier that takes data as data row in input and label is predicted.

The performance is calculated based on student's outcomes of learning method. Different assessments form various other outcomes giving valuable info about the student method of learning and how much the teaching method is accurate. Still grading is much more important.
It is the standard method of predicting level of a performance. Grade is divided in 2 types like Grade point Average and cumulative Grade point Average. GPA is evaluated by how much student is earned grade points in given time period.

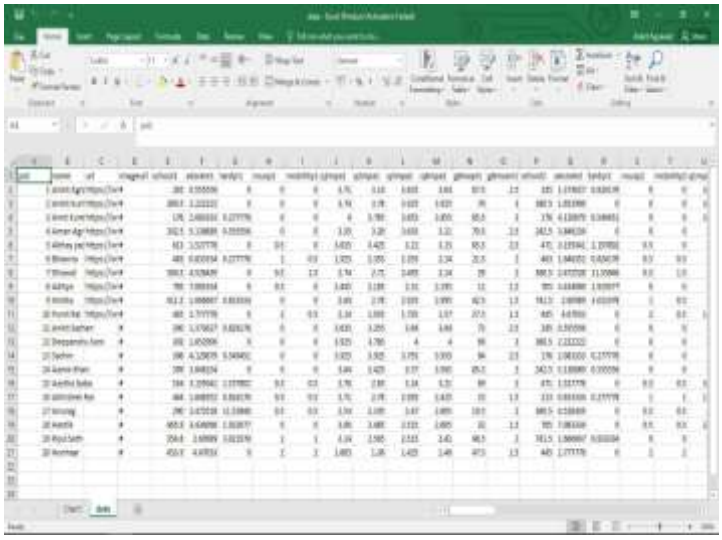
Student education is another recognized participative factor related to progress. Here, some basic things like gender, how many people are in the family, age, parents' cohabitation, how far the parents have been educated, and other things have been mentioned in this inquiry paper.

4.1 Steps implemented in the project :here are 1000 students in this data. This dataset consists of age, gender, grade point, father's job, and home status, lack of study, romantic activities, alcohol consumption, rotation, extracurricular activities, number

of students, and three different semesters.

- The next stage is pre-processing in which any data with zero or zero values are released.
- In the third stage, after pre-processing the data is selected according to the requirement of the particular object.
- Then in this phase K-Nearest Neighbor (KNN) and Naive Bayesian classification Algorithm are used to training and testing the model respectively implemented in Python.
- And, in the last step, R language is use to graphically represents the GPA attached by a student and one after another among the above elements.

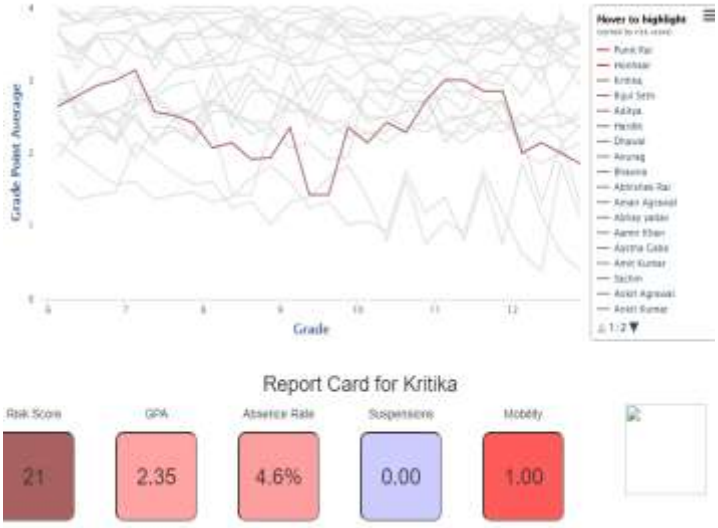
DATASET :



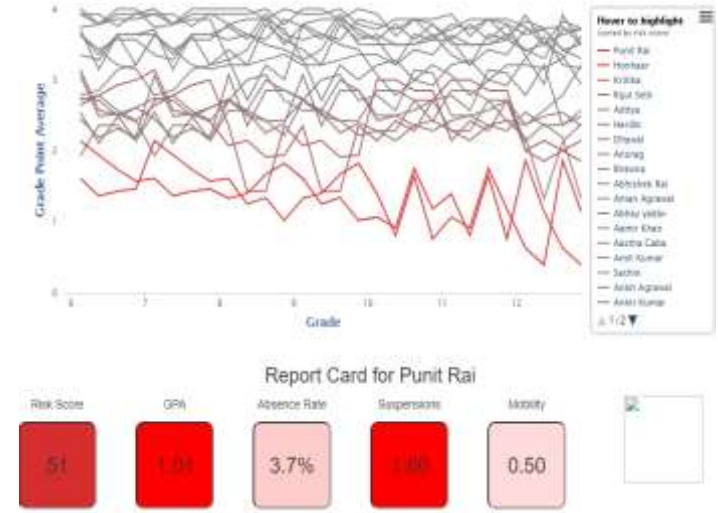
DASHBOARD 1:

Fig.4 Graph analysis of student performance using given Algorithms

DASHBOARD 2:



DASHBOARD 3:



5.RESULT:

The final result can be visualized in a graph. This report provides the information about the student's performance. The factors considered here are student's GPA, Risk Score, Absence rate, Suspensions and Mobility. Absence rates are the percentage of how a student is regularly attending a class. Suspension is also calculated for analyzing the discipline of a student and his behavior in the class. Mobility is a dependent factor for absence rate as students have to move from other cities have to move the institute during holidays and it can cause a spike in absence rates. For Example, a graph can be suggested that shows the performance of Ankit Agarwal based on his report card. Factors like GPA; GPA is evaluated after the number of points is collected of a student and its average is calculated for a given time period. Absence rate, Suspensions and Mobility all contribute as a dependent factor for calculating our final risk score for the student. The result of Ankit Agarwal is therefore calculated and final report is achieved as shown below:

Risk Score: 0
GPA: 3.85
Absence rate: 0.09%
Suspensions: 00
Mobility: 00

6. CONCLUSION:

For the benefit of the student and his/her performance, it is important to take note of the given stats and hence it will give us an idea of his/her future performance. The risk score is beneficial for a student to understand his overall capability to learn and implement new things. It majorly depends on the basic idea that how his GPA scores, Absence rate and other factors perform. Other factors like suspensions and mobility issues can significantly degrade a student's performance so they are kept in analysis. Overall analysis for Kritika shows that her report card doesn't fetch good result because of high risk score which are due to low GPA and high absence rates. Institutions can make use of this system to analyze the performance of their students over a long period. This can significantly help the institutes to attain stability across a large number of students in their performance.

7. REFERENCES:

[1.] Tyagi D., Sharma D., Singh R., Kishor K., "Real Time 'Driver Drowsiness' & Monitoring & Detection Techniques", International journal of Innovative Technology And Exploring Engineering, Vol. 9, Issue 8, ISSN 2278-3075, pp 280-284, June 2020

[2.] Jain A., Sarma Y., Dr. Kaushal Kishor, "Financial Supervision and Management System Using Ml Algorithm**",** Solid State Technology, Volume: 63 Issue: 6, PP 18974-18984, Publication Year: 2020'

[3.] Vairachilai S, Vamshidharreddy, (June 2020) "Student's Academic Performance Prediction Using Machine Learning Approach", IJAST, vol. 29, no. 9s, pp. 6731 – 6737.

[4.] R. Alshabandar, A. Hussain, R. Keight and W. Khan, (2020) "Students Performance Prediction in Online Courses Using Machine Learning Algorithms," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, pp. 1-7.

[5.] Roy C., Pandey M., Rautaray S.S. (2019). A Proposal for Optimization of Horizontal Scaling in Big Data Environment. In: Kolhe M., Trivedi M., Tiwari S., Singh V. (Eds) Advances in Data and Information Sciences. Lecture Notes in Networks and Systems, vol 38. Springer, Singapore.

[6.] Erkan Err, Roy, C., Pandey, M., & Rautaray, S. S. (2018). A Proposal for Optimization of Horizontal Scaling in Big Data Environment. In Advances in Data and Information Sciences (pp. 223-230). Springer, Singapore.

[7.] M. F. Sikder, M. J. Uddin and S. Halder, (2016). "Predicting students yearly performance using neural network: A case study of BSMRSTU," 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, Bangladesh, 2016, pp. 524-529

[8.] Shaukat, K., Nawaz, I., Aslam, S., Zaheer, S., & Shaukat, U.. (2016, December). Student's performance in the context of data mining. In 2016 19th International Multi- Topic Conference (INMIC) (pp.1-8). IEEE.

[9.] Fok, W. W., Chen, H., Yi, J., Li, S., Young, H. A., Ying, W., & Fang, L... (2014, September). Data mining application of decision trees for student profiling at the Open University of China. In 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications (pp. 732-738). IEEE.

[10.] Ramesh, V. A. M. A. N. A. N., Parkavi, P., & Ramar, K... (2013 February). Predicting student performance: a statistical and data mining approach. International journal of computer applications, 63. 975-8887. 10.5120/10489-5242.

[11.] B.K. Bharadwaj and S. Pal (2011, April)."Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140,.

[12.] Oyelade, O. J., Oladipupo, O. O., & Obagbuwa, I. C... (2010). Application of k means Clustering algorithm for prediction of Students Academic Performance. ArXiv preprint ArXiv: 1002.2425.

[13.] Nonis, S. A., & Hudson, G. I... (2006). Academic performance of college students: Influence of time spent studying and working. Journal of education for business, 81(3), 151-159.

[14.] Suryadarma, D., Suryahadi, A., Sumarto, S., & Rogers, F. H. (2006). Improving student performance in public primary schools in developing countries: Evidence from Indonesia. Education Economics, 14(4), 401-429.

[15.] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, (2003 September). "Preventing student dropout in distance learning systems using machine learning techniques," AI Techniques in Web-Based Educational Systems at Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems, pp. 3-5.

[16.] Chemers, M. M., Hu, L. T., & Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. Journal of Educational psychology, 93(1), 55.