**ML BASED PREDICTIVE ANALYTICS OF STUDENT ACADEMIC PERFORMANCE**

1. **ABSTRACT:**

**2) INTRODUCTION:**

Machine learning algorithms have proven to be a helpful tool in predicting students performance based on various factors. The variables used for prediction included internal exams, semester exams and attitude toward studying, including marks obtained in assignments and student performance. Social and school related features were also noted. Incremental approach in machine learning is important for real world predictions for reasons, the learning approaches must do some changes on the trained system so that unlearned knowledge can be proved useful. It is proven for the fact that student had whether passed in final exam or not, but there are also other relevant variables contributing for accurate predictions [1]. Several studies have been conducted to predict student academic performance using various features. Feature selection is an important aspect in making predictions. Many features have been used, ranging from personal, social, demographic, behavioral, and academic data. From those factors, the academic factors are still the most influential feature in determining academic performance [3]. In recent times, machine learning (ML) practices have been a big deal in various industries in the world including the educational frontier. The need to automate different tasks such as grading students, improving student retention, testing students, predicting student performance, as well as administrative tasks like material optimizations in the academic facet, has called for the application of machine learning techniques and methods [4]. The quality of an academic institution is depend on the performance of student and dropout rate between the enrolled students in a course and finally completed the course [8]. The sole focus of this field is automatic learning methods. Learning refers to modification or improvement of algorithm based on past “experiences” automatically without any external assistance from human[9] .

2.1 THEORY

2.1.1 Machine Learning and Data Analytics:

ML is considered as a branch of Artificial intelligence (AI). It aims to imitate the way human beings solve problems in the form of creating computational neural networks that mimics neural networks in the brain [6]. ML is considered as a branch of Artificial intelligence (AI). It aims to imitate the way human beings solve problems in the form of creating computational neural networks that mimics neural networks in the brain[9]. Moreover, a key feature of machine learning is the capacity to analyze complex non- linear relationships, given that complex input variables are expected. The application of machine learning techniques to predicting students’ performance, based on their background information and their in-term performance has proved to be a helpful tool for foreseeing poor and good performances in various levels of education [11]. Data analytics is the science of analyzing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.

2.2 Background Information

:

Prediction by analogy is so pervasive, that we normally don’t notice it. This prediction technique would prove advantageous for teachers in providing additional assistance to weak students and also promoting the dedicated ones. The data of students is collected and utilized to meet the needs of students. Other approaches fail to notice the students’ performance pattern over the course of passing semesters. Applying machine learning algorithms on datasets could prove beneficial to all participants in educational institutes. It has been proved that most accurate machine learning algorithm for such prediction[1] . According to the classification in ML is the process of determining which category an observation belongs to. In essence, it’s a process that establishes relationships between a dependent variable (is categorical in nature) and an independent variable (can be categorical or numerical in nature) [6]. A substantial amount of literature focuses on predicting student performance in solving problems or completing courses. Many machine learning techniques, such as decision trees, artificial neural networks, matrix factorization, collaborative filters and probabilistic graphical models have been applied to develop prediction algorithms. It is evident that different machine learning reveals different prediction accuracy of students’ performance. No clear model has been proved to be the best in accurately predicting students’ performance. It may also be important to consider student’s academic performance affected by many factors, like personal, socio-economic and other environmental variable when testing the accuracy of various machine learning models in

predicting student performance. Moreover, the various machine learning models did not identify

the best model in improving student’s outcome. It is not clear which model is best in predicting

performance and at the same time best in improving learning in among students [11].

2.3 Problem Statement:

To predict and identify the students’ academic performance to guide them to better results and provide quality education[1] . Determine to what extent student data and machine learning can be used to identify students that need help with their university academic performance [6].

2.4 Objective:

The main objective of higher education institutions is to provide quality education to its students. A good prediction of the students’ performance is helpful to identify the low performance students at the beginning. The intention behind this research is identification and extraction of knowledge for foreseeing poor and good performances[1].

1. **LITERATURE REVIEW:**

**4) PROPOSED WORK:**

4.1 Research Goal:

To understand the problem more formally, a literature review has been performed to study the existing work and related studies. There are many researchers who analyzed the performance of students. Most of them worked on the effects of variables like students’ attendance, personal efforts, time for studies etc. on their performance. Some of them also considered the current CGPA of students but none of them has taken student’s grade in individual courses as an attribute for the study good prediction of the students’ performance is helpful to identify the low performance students at the beginning of the learning process. This process will be done by using Machine Learning and Data Science tools and techniques. The intention behind the model is identification and extraction of potentially valuable knowledge for foreseeing poor and good performances in academic exams [1] .

4.2 Data Collection:

The data is collected from the Computer Science department.

4.3 Data Preparation:

The data acquired [1] contains redundancy and useless information that is not needed for the predictions and modelling. Furthermore, the structure of the data is also required to change because it can’t be used as an input for classification models.

4.3.1 Data Integration:

Combining [1] the data was required for Neural Network and SVM Classifier. Microsoft Excel and its tools were used to simply integrate two datasets into a single dataset.

4.3.2 Data Transformation:

For the modelling process,[1] each student must be in a single row with all

its courses in the same record and in the data acquired each student, course pair (i.e., Registration) is a single separate record. This transformation is done using PowerPivot, an external library of Microsoft Excel for data analysis, modelling, cleaning and transformation. The Tables - 1 & 2 will further illustrate the process of transformation.

Table 1- Data before Transformation

|  |  |  |
| --- | --- | --- |
| Roll no | Course Title | Grade |
| 1 | Computer Programming | 3.33 |
| 2 | Data Structures | 2.33 |
| 1 | Discrete Structures | 2.00 |
| 1 | Data Structures | 2.67 |
| 2 | Computer Programming | 3.33 |

Table 2- Data after Transformation

|  |  |  |  |
| --- | --- | --- | --- |
| Roll no | Computer Programming | Data Structures | Discrete Structures |
| 1 | 3.33 | 2.67 | 2.00 |
| 2 | 3.33 | 2.33 |  |

After this transformation, the appropriate course columns are selected for the modelling process.

4.3.3 Data Cleaning:

The data [1] we now have after transformation has many missing values and some other errors. Given below is the checklist of the whole cleansing process step by step.

Missing Values are found in the data set after the data transformation.

The causes of these students with the missing grade in a course are:

1. They are from the other Degree Program (i.e., Electrical Engineering)

2. They haven’t completed their degree or they are university dropouts.

3. The course selected is renamed or is discarded from the list of offered courses in the later semesters.

For the case 3, above the observations were merged with the renamed courses. For Case 1 and 2 the observations are removed, if they haven’t completed the chain of courses or if they are from another degree program.

4.3.4 Exploratory Data Analysis

This phase [1] is about exploring data and the goal is not just to clean the data but to discover

the irregularity or exceptions that were missed before. Graphs, Plots and Charts are used to explore data and are combined so that they can provide even more insights.

4.4 FLOW CHART

4.5 MODULES & EVALUTION:

4.5.1 Model Selection:

The methods [1] are selected on the basis that they classify the students into four categories; depending on the Students academic performances.

The models selected are:

1. REGRESSION:

Regression analysis [9] is part of predictive analytics and exploits the co-relation between **dependent** (target) and **independent variables**. The notable regression model is Logistic Regression.

1. Naïve Bayes:

In machine learning, [1] Naïve Bayes is a classifier based on probabilities, applying Bayes' theorem with strong (naïve) independence assumptions between the features. Naïve Bayes Classifier is selected because the future grades of the particular course chain are required to be predicted if the details of previous courses/subjects are provided. Naive Bayes is one of the inductive learning algorithms that is effective and efficient for machine learning [11].

Pros:

The following are some of the benefits of the Naive Bayes classifier:

* It is simple and easy to implement
* It doesn’t require as much training data
* It handles both continuous and discrete data
* It is highly scalable with the number of predictors and data points
* It is fast and can be used to make real-time predictions
* It is not sensitive to irrelevant features

Cons:

* If your test data set has a categorical variable of a category that wasn't present in the training data set, the Naive Bayes model will assign it zero probability and won't be able to make any predictions in this regard.

1. SVM:

Support Vector Machine [1] comprises of the concept of planes. A plane will have sets of different instances belonging to different classes. It has different classifying techniques. The simplest one is linear, which separates sets of instances into their groups with just a line. For more accurate classification, more complex structures are used. In our classification model, we have used different kernel functions; which are linear, polynomial and gaussian kernel.

In classification phase [11], increment of classes may cause the reduction of success rate in SVM. However, it can be used effectively for the problems.

Pros:

* It is well-suited for multi-class case, where the classification has to obtain the result with more than two classes.
* It works really well with a clear margin of separation. It is effective in high dimensional spaces.

Cons:

* It doesn't perform well when we have large data set because the required training time is higher.

1. KNN :

K-NN [11] is a kind of instance-based learning, or lazy learning, where the function is only approximated nearby and the entire calculation is delayed in anticipation of classification. The main problem of k-NN algorithm is that its accuracy can be strictly ruined by the existence of loud or inappropriate features.

Pros:

* There's no need to build a model or tune several parameters.

Cons:

* It is quite computationally inefficient and its difficult to pick the “correct” value of K.

4.6 MATERIALS AND METHODS:

The main goal [4] of this systematic literature review is to explore the applications of machine learning methods to educational data to predict or monitor in student academic performance.

It starts [6] by listing the scientific questions and how they were approached and solved.

4.6.1 Software Tools:

The software tools [1] used for this research consist of Microsoft Excel is used for data preparation . The models are created by using Python programming [3] language is used along with it’s Integrate Development Environment (IDLE) . Tools [4] like Numpy, Pandas, Seaborn, MATLAB and Python are used for machine learning as well as model deployment. The platform [1] used is Anaconda.

The machine [4] learning techniques used here is as follows:

* Supervised Algorithms - Regression (Logistic), Support Vector Machines, Naïve Bayes and

k- Nearest Neighbor.

4.6.2 Data set :

4.6.2.1 Description

The data [1] is collected from computer science department which consists of students’ transcript data that included their grades in all courses. The dataset consisted of students from the 3 academic years. After eliminating incomplete data, the samples consisted of 60 students.

Students’ final grades divided into four classes:

Table 3- Class Labels

|  |  |
| --- | --- |
| **CLASS** | **GRADE** |
| 1 | A+, A, A- |
| 2 | B+, B, B- |
| 3 | C+, C, C- |
| 4 | D+, D, D- |

Each row has the details of a student registration in a course.

The attributes of the data collected are the following:

* Student id : The students’ identification number in the university record.
* Course: It is the course code of the course, student is registered in.
* Attendance : The attendance of the student during academic.
* Assignment : The assignment test of the student in the course.
* Lab : The lab test of the student in the course.
* Project : The project or presentation of the student .
* Backlog : Whether the student has any backlog (i.e., Yes (1) No (0) )
* Final Semester result : Final semester result of the student.
* Final CGPA : The CGPA of the student in the course.
* Grade: The grade of the student in the course.

4.7 ALGORITHM :

4.7.1 Naïve Bayes

Result of applying naïve bayes algorithm for the past data set.,

4.7.2 KNN

Result of applying k-nearest neighbor algorithm for the past data set.,

4.7.3 SVM

Result of applying Support Vector Machine algorithm for the past data set.,

* 1. **RESULT:**

After applying various models to the dataset, different accuracy is obtained for each model. Table IV shows the accuracy comparison for all models in predicting academic.

Table 4:

|  |  |  |
| --- | --- | --- |
| **S.no .** | **Method** | **Accuracy** |
| 1 | Naïve Bayes |  |
| **2** | Knn |  |
| **3** | SVM |  |

**6) CONCLUSION:**

**7) REFERNCE:**