International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 11 | Nov 2022 www.irjet.net p-ISSN: 2395-0072

Student Performance Predictor

Prasham Mehta, Nidhi Lade

Student, Dept. of Information Technology, Atharva College of Engineering
Student, Dept. of Information Technology, Atharva College of Engineering

---

Abstract - Today technology has occupied such a significant part of our lives that it is almost impossible to imagine life without it. A myriad of technological advancements occurs every day, so much so that it is difficult to track all of them. Also, today every educational institution handles and deals with large amounts of student data which can be advantageous for a number of reasons. One of the most important applications of such data is predicting student performance.

Integrating a Graphical User Interface (GUI) with Machine Learning algorithms is not an easy task. Hence, our student performance predictor application implements the same by combining Machine Learning algorithms along with GUI which makes it easy and efficient to use. As a result, this prediction can be helpful to both, the students and teachers. Predicting the performance of students is a challenging task, because of the large amounts of information stored in academic databases. With the help of Machine Learning (ML) and Artificial Intelligence (AI), this proposed system can help to predict as well as analyze students' performance by not only considering their academic details but also other factors like study time, failures or backlogs, and so on. For this, we have used several approaches or algorithms such as Logistic Regression, Support Vector Machine (SVM), Naive Bayes, and K-Neighbors Classifier respectively.

Through this task, we extract knowledge that describes students' overall performance. It helps in identifying the strengths and weaknesses of the students beforehand. And ultimately, this prediction can help students to enhance their performance in the future. In addition to this, our application also provides relations and patterns between the user input field in a detailed and graphical manner. The user can view these patterns and relations in the form of scatter plots, heat maps, box plots, histograms, dist-plot, violin-plot, etc.

Key Words: Prediction, Logistic Regression, Support Vector Machine, Machine Learning, Artificial Intelligence.

1. INTRODUCTION

[4] The data stored in educational databases are rapidly growing. So, there must be an easy and efficient way to handle this enormous data. Managing such large data and improving the academic performance of students is tough. Therefore, the main aim of this system is to predict the future performance of the student using certain data of the students such as previous semester marks, study time, etc. These predictions will not only help teachers to classify students based on predicted grades and suggest they focus on certain subjects, but will also help students improve their studying goals and improve their grades. We have evaluated the following algorithms by applying them to the dataset.

1.1 Supervised Learning

The algorithms used in this project are supervised learning algorithms. [1] Supervised Learning is the type of machine learning where machines are trained based on well "labelled" training data and based on this data, the machines predict the output. The labelled data means that some input data is already tagged with the correct output.

Starting with Logistic Regression, one of the most popular Machine Learning algorithms. Logistic Regression is used to estimate discrete values, usually binary values like 0 and 1 from a set of independent variables. Next is the Support Vector Machine, which is used for classification and regression problems. The goal of SVM is to generate the best line or decision boundary which can separate n-dimensionalspace into classes so that they can easily put the new data point in the correct category in the future. Further, we have the Naive Bayes, based on the Bayes theorem. Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is a method used to determine the probability of an event based on the occurrences of prior events. It is used to calculate conditional probability. Naïve Bayes Classifier is one of the simple and most effective algorithms used for classification which helps to build fast machine learning models that are capable of making quick predictions. And finally, we have the K-Neighbors Classifier is one of the simplest Machine Learning algorithms. It stores all available cases and classifies them by taking a majority vote of its k Neighbors.

1.2 Graphical User Interface (GUI)

Out of these algorithms, Logistic Regression (LR) is the one that we have implemented in our system as it provides the

best accuracy compared to other algorithms. Our student performance predictor is basically a web-based GUI application that lets students, as well as teachers, view and analyze the student's performance by simply entering details about the students such as their study time, previous grades, failures or backlogs, and many more. There are a total of three sections in the application. First is the home section which presents a brief description of all the above-mentioned algorithms along with the methodology. Second is the visualization section. Here, the user can view different things like the original dataset, the pre-processed data, and various features such as mean, standard deviation, and the minimum and maximum range in the dataset, followed by an intricate graphical visualization in the form of charts, plots and heat maps and finally the accuracy of the model. Finally, we have the prediction section where the user needs to enter their details and submit them. After the submission, the performance of the student will be displayed in the form of a grade. The student performance is predicted on the basis of five grades from 'A' to 'E' where 'A' indicates excellent and 'E' indicates poor.

The dataset used is taken from UCI Machine Learning Repository. This data is based on two Portuguese high schools. Student grades, demographic, social, and school-related features are the data attributes and these were collected by using school reports and questionnaires. Among all the algorithms mentioned above, Logistic Regression provided the best accuracy which is 93.16%.

2. ALGORITHMS

2.1 Logistic Regression

Logistic regression is one of the most widely used Machine Learning algorithms, which is the most important part of the Supervised Learning technique. Logistic regression is a type of regression analysis that is used to predict the probability of a binary outcome. The outcome is either a success or a failure. The logistic regression model is used to estimate the probability of success based on one or more independent variables. Logistic regression predicts the output of an explicit dependent variable.

The logistic regression model is a linear model that can be used to predict the probability of a binary outcome. The model is based on the logit function, which is the natural logarithm of the odds of success. The logit function is used to model the relationship between the independent variables and the binary outcome.

Logistic Regression is a significant machine learning algorithm because it has the capability to provide probabilities and classify contemporary data using continuous and discrete datasets. The logistic regression model can be used to estimate the probability of success based on the values of the independent variables. The model can be used to predict the probability of success for new data. The model can also be used to identify the factors that are associated with a higher probability of success.

2.2 Support Vector Machine

A support vector machine (SVM) is a supervised learning algorithm that is used mostly for classification and regression tasks. The algorithm is a discriminative classifier that not only finds the decision boundary between classes but also tries to maximize the margin between these classes. This results in a more robust model that is less likely to overfit the training data.

SVMs are more effective in high-dimensional spaces and are therefore well-suited for problems where there are many features, such as text classification and image classification. They are also effective in cases where the number of training examples is small.

The main idea behind an SVM is to find a hyperplane that maximally separates the data points of one class from the data points of the other class. The main advantage of using an SVM is that it is very effective in high-dimensional spaces. This is because the SVM finds a hyperplane by looking at the data points closest to it, which are called support vectors.

In other words, we are looking for the hyperplane that has the largest margin between the two classes. Once we have found this hyperplane, we can use it to make predictions on new data points. If we are dealing with a classification task, then we can predict that a new data point will belong to the class that is closest to the hyperplane. If we are dealing with a regression task, then we can predict the value of the target variable for a new data point by taking the value of the target variable that is closest to the hyperplane.

2.3 Naïve Bayes Algorithm

The Naive Bayes algorithm is a simple probabilistic classifier that is based on applying Bayes' theorem with strong (naive) independence assumptions between the features. The Naive Bayes algorithm is a simple, yet powerful machine learning technique for predictive modeling. It is a supervised learning algorithm that is based on the Bayes theorem of probability. The algorithm is designed to be "naive" in the sense that it makes strong assumptions about the independence of the input features.

Assuming that we have a dataset with two features and two classes, we can represent the joint probability distribution of the two features and the class labels are as follows:

$$P(X|Y) = P(Y|X) * P(X) / P(Y)$$

Where the probability that we are interested in calculating P(A|B) is called the posterior probability and the marginal probability of the event P(A) is called the prior.

The Naive Bayes algorithm makes the following two assumptions:

1. The features are not dependent on each other.
2. The class labels are mutually exclusive.

The Naive Bayes algorithm is commonly used in text classification tasks, such as spam detection or sentiment analysis. It is also often used in machine learning competitions, such as the Kaggle competitions. The Naive Bayes algorithm is trained on a training data set. The training data set is used to estimate the probabilities of the various class labels and the probabilities of the various feature values. These probabilities are then used to make predictions on new data.

The key advantage of the Naive Bayes algorithm is its simplicity.It is easy to understand and implement. It is also computationally efficient and scalable.

2.4 KNeighborsClassifier

It is a type of supervised machine-learning algorithm. KNN is a non-parametric and lazy learning algorithm where Non-parametric means there is no assumption for underlying data distribution. In other words, you don't need to know the data distribution to model the prediction. A lazy algorithm means that the algorithm does not need any training data points for model generation. This means modeling can be extremely fast.It stores all convenient cases and classifies the new data points based on a similarity measure (e.g. distance functions). The K-Nearest Neighbors (k-NN) algorithm is employed to solve both binary and multi-class problems. It is a non-parametric model that relies on proximity to define the relationship between instances in the training dataset. The number of neighbors is represented by the parameter k.

The basic idea of the KNeighborsClassifier is to calculate the distance between a query and all the examples in the trainingdata. The distance can, for example, be the Euclidean distance between two vectors. Once the distance is calculated, you select the top K examples that are closest to the query.

The majority of practical machine-learning applications can be solved quite effectively by a few simple algorithms like linear regression or support vector machines. However, there are situations where more complex algorithms could be useful. In particular, the KNeighborsClassifier is a good choice if your data isn't well-described by linear models and you need a non-linear model. KNN is a non-parametric method and simple algorithm that can be used for both classification and regression. KNN can be used for the classification of data with more than two classes. KNN can be used for regression by taking the average of the K nearest neighbors as the estimateof the dependent variable for a given independent variable.

This approach has several advantages:

1. There's no need to build a model, so you can avoid

overfitting.

2. The algorithm is easy to understand and implement.
3. All the data is utilized to build the nearest neighbor models.

3. METHODOLOGY

The methodology is the core component of any research-related work. The methods used to gain the results are shown in the methodology. Here, the whole research implementation is done using python. There are different steps involved to get the entire research work done which are as follows:

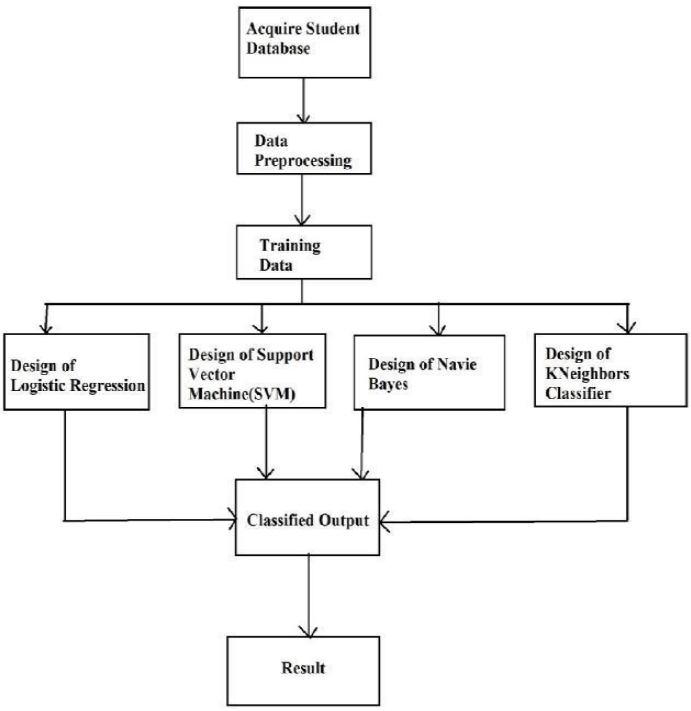

Fig -1: Methodology of Student Performance Predictor

3.1 Acquire Student Database

The UCI machine learning repository is a collection of databases and data generators that are used by the machine learning community for analysis purposes. The student performance dataset can be acquired from the UCI machine learning repository which is one of the websites that are available for downloading the datasets for free. The student file consists of two subjects' CSV files namely student-por.csv and student-mat.csv. Also, the dataset has multivariate characteristics. Since the data provided is not consistent and uniform, data-pre-processing is done for

checking and correcting the inconsistent behavior of trends in the dataset.

3.2 Data Preprocessing

After, Data acquisition the next step is to make the data uniform and process the data. The Dataset fetched has object-type features that need to be converted into numerical type. Thus, using the python dictionary and mapping functions the transformation of the data in the dataset is done. The final value is a five-level categorization consisting of 0 i.e. excellent or 'A' to 4 i.e. fail or 'F'.

The pre-processed dataset is further divided into two datasets namely the training dataset and the testing dataset. This is achieved by passing values like feature value, target value, and test size to the train-test split method of the sci-kit-learn package which makes the process easier.

After dividing this data into training and testing, the training data is sent to the following neural network designs

i.e. Logistic regression, Naive Bayes, SVM, and KNeighborsClassifier for training the neural network, and then test data that is used to predict the accuracy of the trained network model for better and more efficient results.

3.3 Design of Logistic regression, Support Vector Machine (SVM), Naive Bayes and KNeighborsClassifier

The design of logistic regression and support vector machinein the python environment is achieved through the NeuPy package which requires the standard deviation value as the most important parameter.

Along with it, the network comprises 30 inputs neuron, a pattern layer, a summation layer, and a decision layer for five-level classification whereas the design of Naive Bayes and KNeighborsClassifier neural network in a python environmentis achieved through the NeuPy package which requires the number of input features, the number of classes i.e. the classification result output neuron, learning rate. The network comprises 30 input features i.e. input neurons, a hidden layer,and the output layer for five-level classification.

Once the design for logistic regression, support vector machine, Naive Bayes, and KNeighborsClassifier is ready it is trained with the training data for accurate classification, andthen testing data is used for the trained neural network.

3.4 Testing and Classified Output

After the training of the designed neural network, the testingof logistic regression, support vector machine, Naive Bayes, and KNeighborsClassifier is performed using testing data. Based on testing data, the accuracy of the classifier is determined and the algorithm with the maximum accuracy isused for the best possible outcome or prediction.

In this project, the highest accuracy is found in logistic regression and hence the calculations and computations are done using logistic regression. The computation may take some time but the results displayed are 93.16% accurate.

4. CONCLUSION

[2] At present, the Logistic Regression algorithm gives us thebest accuracy among all the other algorithms. But in the future, if any new techniqueor approach is introduced.

This might be better than Logistic Regression or the one which would provide better accuracy, implementation of that particular algorithm would be easy because the program via which the application is made is flexible and changes can be made quite smoothly.

[3] A student performance predictor is a tool that can be used by educators to help identify students who may be at risk foracademic difficulties. This tool can be used to help target interventions and supports for those students who may needthem the most. In conclusion, the meta-analysis on predicting students' performance has motivated us to carry out further research which can be applied in our educational institutes. Hence, this model will be helpful for the educational system to review the student's performance in a systematic manner.

5. FUTURE WORK

There are many possible directions for future work on studentperformance prediction. One direction could be to use more sophisticated machine learning methods, such as deep learning, to improve predictive accuracy. Another direction could be to incorporate additional data sources, such as data on student demographics, into the predictive models. Anotherway to improve the predictor would be to use a larger and more representative dataset. Finally, the predictor could be made more user-friendly, for example by providing a graphicalinterface.

REFERENCES

[1] Ditika Bhanushali, Seher Khan, Mohammad Madhia, Shoumik Majumdar " Student Performance Prediction And Analysis" IJARCCE (2018)

[2] Kalpesh P. Chaudhari "Student Performance Prediction System using Data Mining Approach" IJARCCE (2017).

[3] Isha D Shetty "Student Performance Prediction" International Journal of Computer Applications Technology andResearch (2019)

[4] Dr. Chandrashekhar Raut, "STUDENT PERFORMANCE PREDICTION USING DATA MINING TECHNIQUES" International Research Journal of Engineering and Technology(IRJET) 2020.