International Journal of Research in Engineering and Applied Sciences(IJREAS)
Available online at http://euroasiapub.org
Vol. 8 Issue 12, December-2018,
ISSN (O): 2249-3905, ISSN(P): 2349-6525 | Impact Factor: 7.196 |

Predictive Modeling for Student Performance: Harnessing Machine Learning to Forecast Academic Marks

Chaitanya Krishna Suryadevara
Department of Information Systems
Wilmington University
chaitanyakrishnawork123@gmail.com

Abstract :- This research investigates the impact of machine learning on higher education teaching and learning, as well as strategies for enhancing the learning environment. There has been a notable increase in student interest in online and digital courses, and platforms such as Course Era and Udemy have become increasingly popular. This study utilizes machine learning applications in teaching and learning, taking into account students' backgrounds, prior academic performance, and other relevant factors. However, due to the large class sizes, it may be challenging to provide personalized support to each student in open learning courses, which could lead to a higher dropout rate. To address this issue, the study employs linear regression, a machine learning algorithm, to predict outcomes.

Keywords :- Machine Learning, Dropout Rate, Linear Regression, Algorithm, Prediction, Outcomes, Coursera, Udemy.

INTRODUCTION

In the realm of education, the pursuit of academic excellence has always been a primary objective. For both educators and students alike, understanding and forecasting student performance is of paramount importance. It not only aids in identifying areas of improvement but also enables personalized interventions to enhance learning outcomes. With the advent of machine learning and data analytics, the education sector has witnessed a transformative shift in its ability to predict and support student success. This research focuses on the development and implementation of a predictive model that leverages machine learning algorithms to forecast student marks accurately. The model's objective is to harness the power of data-driven insights to provide educators, students, and educational institutions with a proactive tool for understanding and improving academic performance. Historically, educators have relied on conventional grading systems and summative assessments as primary indicators of student achievement. While these methods are valuable, they often offer limited opportunities for early intervention and individualized support. Moreover, educational systems are becoming increasingly complex, with diverse student populations and unique learning trajectories, making it challenging to cater to each student's needs effectively.

Machine learning offers a promising solution to these challenges by enabling the analysis of a multitude of variables that may influence student performance. By mining data related to students' prior academic records, attendance, study habits, and even socio-economic factors,

predictive models can identify patterns and correlations that are beyond human intuition. These insights can then be used to make data-informed decisions, such as identifying students at risk of underperforming and offering tailored interventions.

In this paper, we will explore the development and implementation of a student marks prediction model, examining the various factors that contribute to academic success. We will delve into the machine learning algorithms employed, the data sources used, and the evaluation metrics applied to gauge the model's effectiveness. Additionally, we will discuss the ethical considerations surrounding the use of predictive analytics in education and the potential benefits and challenges associated with this innovative approach. The integration of machine learning in education holds immense promise for optimizing the learning experience, enhancing teaching methodologies, and ultimately improving student outcomes. As we embark on this journey to explore student marks prediction using machine learning, we aim to contribute to the growing body of research that seeks to harness the power of data for the betterment of education. The quality of education is crucial for the development of a country, and the goal of educational institutions is to provide high-quality education to their students. One way to achieve this is by predicting students' academic performance and taking early action to improve both students' performance and teaching quality. Machine learning has been a popular topic of interest in recent years. This project aims to predict students' marks using linear regression, a machine learning algorithm. The goal is to determine how many hours of study are needed to achieve a certain percentage of marks and to predict the marks a student can expect to receive based on the number of hours they study per day. By using this information, schools and colleges can assess the performance of their students and take measures to improve it. The linear regression algorithm is used to train the model and make predictions.

METHODOLOGY

The first step in the implementation process is to gather the data set needed for the study. This is done by collecting data on students. To simplify the analysis, we can identify unique attributes in the data set and remove them if they are not relevant to the analysis. The data is then formatted for use. This process is called preprocessing of data, and it is an important step in extracting the necessary information from raw data. The higher the accuracy rate, the better the results. After preprocessing the data, the next step is to identify and remove any incomplete or irrelevant data from the dataset in order to produce accurate findings. This process is called data cleaning. There are several techniques that can be used for improved classification, such as linear regression, support or code vector machine, Naive Bayes Standard Classification, and mostly decision tree algorithms. In this research, we use the linear regression algorithm to implement the solution. We also need to select a training set from the dataset, identify the result attributes that determine the output, and begin classification. The problem statement for this study is that student marks are currently being analyzed and predicted based on guesswork and do not consider students' personal marks for academic evaluation.

Methodology

1. Data Collection and Preprocessing

- **Data Sources**: Gather historical student performance data from educational institutions, including academic records, assessment scores, attendance records, demographic information, and potentially additional factors such as extracurricular activities.
- **Data Cleaning**: Preprocess the data by handling missing values, outliers, and data inconsistencies. Ensure data quality and consistency for accurate modeling.
- **Feature Engineering**: Create relevant features, such as average exam scores, study hours, and participation in study groups, that may contribute to predicting student marks. Convert categorical variables into numerical representations if necessary.

2. Data Splitting

- **Training and Testing Sets**: Divide the dataset into training and testing subsets. Typically, an 80-20 or 70-30 split is used, with the larger portion allocated for model training.

3. Model Selection

- **Algorithm Selection**: Choose appropriate machine learning algorithms for the prediction task. Common choices include linear regression, decision trees, random forests, and gradient boosting algorithms.
- **Hyperparameter Tuning**: Fine-tune algorithm hyperparameters to optimize model performance. This may involve techniques like grid search or random search.

4. Model Training

- **Training the Model**: Train the selected machine learning model on the training dataset. The model will learn the relationships between input features (e.g., student attributes) and the target variable (e.g., final exam scores).

5. Model Evaluation

- **Performance Metrics**: Assess the model's performance using appropriate evaluation metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
- **Cross-Validation**: Implement k-fold cross-validation to ensure model robustness and minimize overfitting. Adjust the model as needed based on cross-validation results.

6. Predictive Modeling and Testing

- **Testing the Model**: Apply the trained model to the testing dataset to predict student marks. Evaluate how well the model generalizes to unseen data.

7. Model Interpretation and Analysis

- **Feature Importance**: Analyze feature importance to understand which student factors have the most significant impact on predicting marks. This can provide valuable insights for educators.

8. Ethical Considerations

- **Privacy and Consent**: Ensure compliance with data privacy regulations and obtain informed consent if the data includes personally identifiable information.
- **Bias and Fairness**: Investigate potential bias in the predictive model and address any fairness concerns to prevent unintended discrimination.

9. Model Deployment and Integration

- **Implementation**: If the model performs satisfactorily, consider integrating it into the educational system to provide real-time predictions or early warning systems for students at risk of underperforming.

10. Continuous Improvement

- **Feedback Loop**: Establish a feedback mechanism to continually update and refine the model as new data becomes available. This iterative process can lead to more accurate predictions over time.

This methodology outlines the systematic approach for developing and implementing a student marks prediction model using machine learning. It encompasses data collection, preprocessing, model selection, training, evaluation, ethical considerations, deployment, and ongoing improvement, with the ultimate goal of enhancing educational outcomes and supporting student success.

Advantages:

Predictions can be made before final marks are evaluated.

Automation of marks prediction can be achieved through the use of machine learning.

Disadvantages:

Most of these methods rely on data mining techniques, which are based on completed data.

Early stage evaluation is not possible with these methods.

ALGORITHM USED

Linear Regression

Linear regression is a Machine Learning algorithm that is used to model the linear is the relationship between a depending variable and one or even more independent variables. This is done by fitting a straight line, called the regression line, to the data. The regression line shows the relationship between the dependent variable, which is typically plotted on the y-axis, and the independent variables, which are plotted on the x-axis. If the dependent variable increases as the independent variable increases, the relationship is positive. If the dependent variable decreases as the independent variable increases, the relationship is negative. Linear regression is often used to make predictions for continuous or numeric variables, such as sales, salary, age, or product price.

Linear regression is a statistical model that aims to predict the value of a dependent variable based on the value of one or more independent variables. It is called linear regression because it models the relationship between the dependent and independent variables using a linear equation. This means that the model assumes that there is a linear relationship between the dependent and independent variables, such that the change in the dependent variable is directly proportional to the change in the independent variable. The linear regression model can be used to make predictions about continuous variables, such as sales, salary, age, and product price.

Linear regression is a statistical method that is used to model the relationship between a depending variable and one or even more independent variables. It is a popular Machine Learning algorithm that is used for predictive analysis. Linear regression models make predictions for continuous or numeric variables, such as sales, salary, age, product price, etc. The linear regression algorithm shows a linear relationship between a dependent variable (y) and one or more independent variables (x), which means that it shows how the value of the dependent variable changes in relation to the value of the independent variable. The linear regression model is represented by a sloped straight line that shows the relationship between the variables. When

one value increases, the other also increases or decreases, depending on the type of linear relationship. A linear line that shows the relationship between the depending and independing variables is called the regression line. Regression lines can show two types of relationships: a positive linear relationship, where the dependent variable increases on the Y-axis as the independent variable increases on the X-axis; or a negative linear relationship, where the dependent variable decreases on the Y-axis as the independent variable increases on the X- axis.

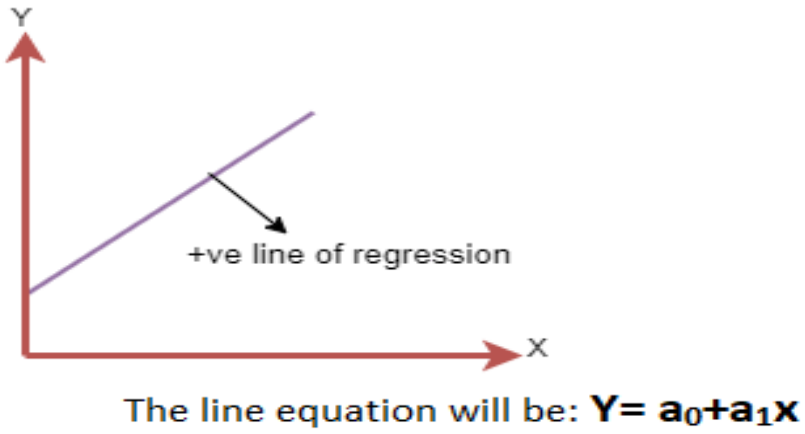


Figure 1 Linear Regression

A. Positive linear regression

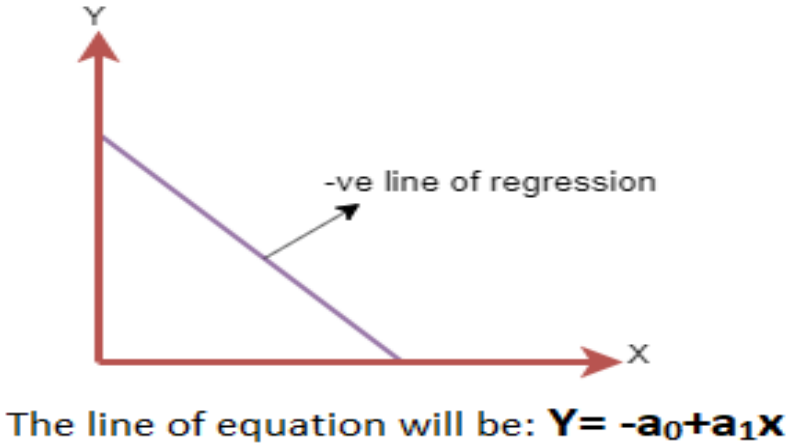


Figure 2 negative regression

B. Negative linear regression

Consider the below image:-

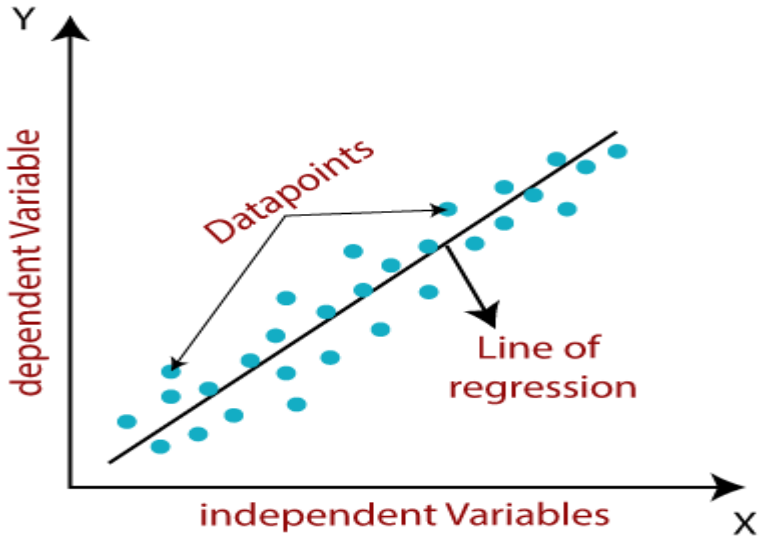


Figure 3 Positive regression

C. Linear regression (independent vatiables vs dependent variable)

Mathematically, also we can represent a linear regression as below :-

$$y= a_0+a_1x+ \varepsilon$$

Here,

Y=Dependent Variable (Target Variable)
X=Independent Variable (predictor Variable)
a0=intercept of the line (Gives an additional degree of freedom)
a1=Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and the y variables are training datasets for the Linear Regression model representations.

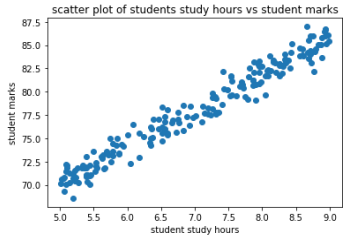
RESULT

By using this model you can make out that how many hours you need to study in order to get good marks and get a good percentage and percentage required to get admission in your wanted colleges and schools. It will make sure that no students fails and could fetch minimum required marks atleast to pass the exam. The Accuracy of the project is 95%

Model	Result
Linear Regression	0.9514124242154466

OUTPUTS

```
[ ]  plt.scatter(x =df.study_hours, y = df.student_marks)
     plt.xlabel("student study hours")
     plt.ylabel("student marks")
     plt.title("scatter plot of students study hours vs student marks")
     plt.show()
```



▾ Prepare data for Machine Learning Algorithm

Figure 4 Machine learning result

CONCLUSION

So this is how you can solve the problem of student marks prediction with machine learning. It is a good regression problem for data science beginners as it is easy to solve and understand. I hope you liked this article on Student marks prediction with machine learning using Python. Feel free to ask valuable questions in the comments section below.

FUTURE SCOPE

In this study, Classification techniques are used to predict student performance. Classification generally refers to the mapping of data items into predefined groups and classes. A classification algorithm analyses the training data during the learning phase, and the test data are used to evaluate the accuracy of the classification algorithms during the classification phase. Classification techniques such as Support vector machine (SVM), Random Forest (RF), Decision tree, and K-nearest neighbour (KNN) algorithms are used to predict the performance of the students. The success of the student in competition is significantly influenced by their performance in the semester exams. this approach aids students who are at risk of having a weak performance by offering an early prediction that improves their performance in the upcoming semesters. The early prediction of marks helps the students to reach their goals and can perform better by studying hard. In this analysis, the Random Forest has given more accuracy of 74% compared to other algorithms. A user interface was developed so that after logging in, individuals can use their preview semester marks to predict their results. In future work, we can include more factors, and using more datasets the performance of the model can be improved and students' performance can be improved.

REFERENCES

[1] Student Marks Predictor using Machine Learning - GOEDUHUB
[2] Hadzic and D. R. Morgan (2009). "On packet selection criteria for clock recovery," Proceedings of the National Academy of Sciences, vol. IEEE Int. Symp. Precision Clock Synchronization Meas. Contr. Commun.
[3] C. S. Turner (2008). "Slope filtering: an FIR approach to linear regression", IEEE Signal Process. Mag., vol. 25, pp. 159-163.
[4] GeeksforGeeks | A computer science portal for geeks
[5] D. Veitch, J. Ridoux and S. B. (2009). "Robust synchronization of absolute and difference clocks over networks," by Korada in IEEE/ACM Trans. Networking, vol. 17, pp. 417-430.
[6] D. R. Morgan and I. Hadžić: Non-uniform linear regression with block wise sample-minimum preprocessing", IEEE Trans. Signal Process
[7] Ankitha A Nichat, Dr. Anjali B Raut (2017). "Predicting and Analysis of student Performance Using Decision Tree Technique", International Journal of Innovative Research in Computer and Communication Engineering, Vol.5, Issue 4.
[8] S.A. Oloruntoba, J.L. Akinode (2017). "Student Academic Performance Prediction Using Support Vector Machine".
[9] Dhanashree Mane, Pranali Namdas, Pooja Gargade, Dnyaneshwari Jagtap, S.S. Rathi (2018). Predicting student Performance Using Machine Learning Approach". VJER Vishwakarma Journal of Engineering Research, Volume 2 Issue 4.
[10] Students Marks Prediction Using Linear Regression – 1000 Projects