Machine Learning Assignment

**Topic** : Student Grades Prediction Model



**Submitted By :**

**Name** : Geethanjali S Kumar

**SRN** : R21EF219

**Class** : 5th Sem D section

**Faculty Name** : Prof Sohara Banu

Description

**Objective**: Prediction of the final grade of Portuguese high school students

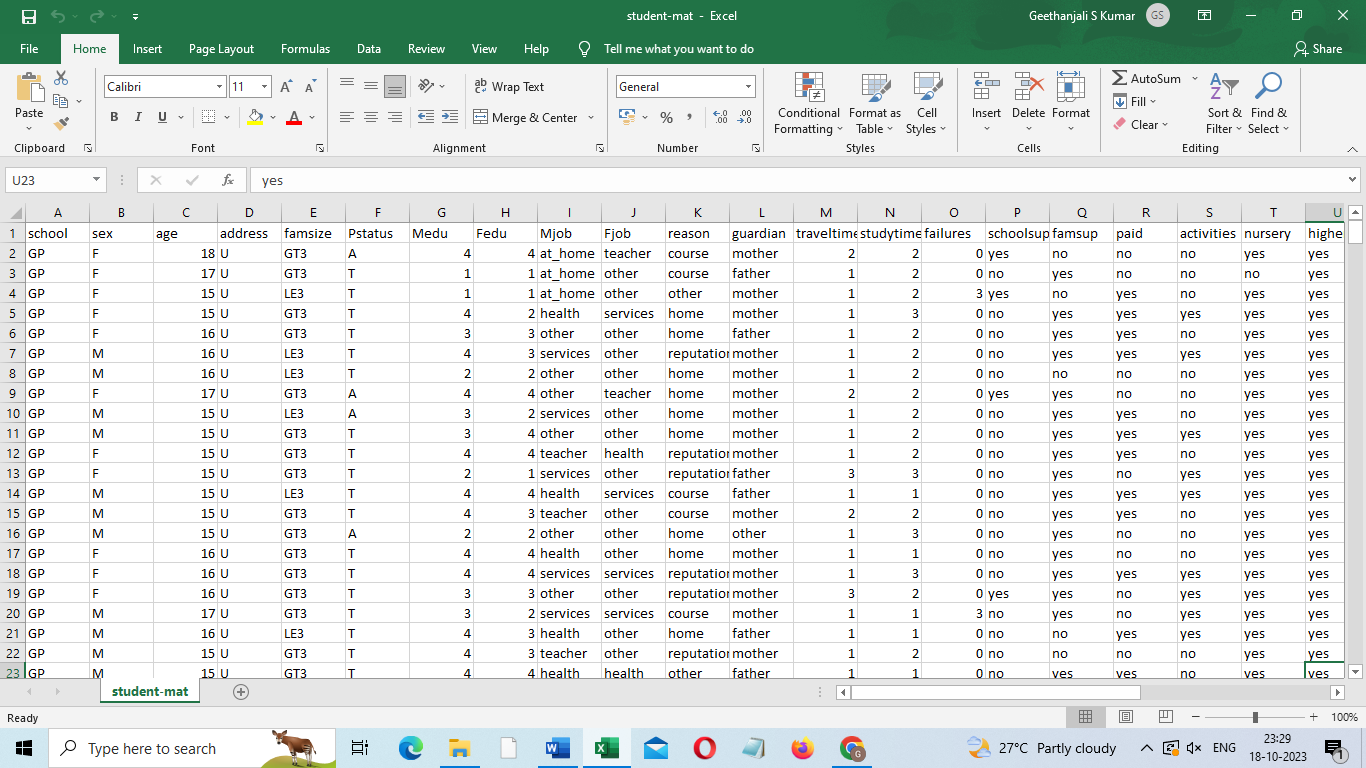
**Problem Statement**

Given a dataset containing attributes of 396 Portuguese students, the goal is to create a linear regression model to predict a student's final grade exam score using the dataset's available features. The task involves using the dataset features to develop a regression model, which can estimate the numerical value of a student's final exam grade.

**Description Of Dataset**

This dataset comprises student achievement data in two Portuguese schools, encompassing attributes such as student grades, demographic information, social factors, and school-related features. The data is collected through school reports and questionnaires and pertains to student performance in two subjects: Mathematics (mat) and Portuguese language (por). The primary objective is to develop regression models to predict a student's final year grade (G3), which is issued during the 3rd period. This prediction is challenging due to the strong correlation between G3 and the preceding grades, G2 (2nd period) and G1 (1st period). Accurate forecasting of G3, considering G2 and G1, is essential for educational assessment and support.

**CSV File Used**



**Methodology**

In higher education institutions, the issue of student retention, especially during the crucial first year, is a significant concern. Many students drop out due to a lack of proper support in their undergraduate courses, making the first year a pivotal 'make or break' period. Without adequate support in understanding course content and its complexity, students can become demotivated and may consider withdrawing from their studies.

Addressing this challenge requires the development of effective solutions to enhance student retention. Early grade prediction is a promising approach to monitor students' progress in their university courses and ultimately improve the learning process. Machine learning, combined with Educational Data Mining, can be a valuable tool for achieving this goal. By developing predictive models for students' course grades, we can provide insights that facilitate better support and guidance for at-risk students. These insights can also aid in forecasting students' grades in various courses, allowing for more effective monitoring of their academic performance, ultimately leading to higher student retention rates.

To gain deeper insights from the dataset, we plan to leverage various data visualization packages such as cufflinks, seaborn, and matplotlib. These tools will help us create graphical representations of the dataset attributes, enabling a visual analysis that contributes to the prediction of the final grade (G3).

**Machine Learning Algorithms used**

1. **Linear Regression :**

* Linear regression is a supervised machine learning algorithm used for predicting a continuous target variable.
* It models the relationship between the target variable and one or more independent variables by fitting a linear equation to the observed data.
* Error Metrics:
* Mean Squared Error (MSE) measures the average squared difference between predicted and actual values.
* Mean Absolute Error (MAE) measures the average absolute difference between predicted and actual values.
* R-squared quantifies the goodness of fit, indicating the proportion of the variance in the target variable that is predictable.

1. **ElasticNet Regression :**

* ElasticNet is a linear regression model that combines both L1 (Lasso) and L2 (Ridge) regularization.
* It is used to prevent overfitting in linear regression by introducing a penalty for the number of features selected (L1) and the magnitude of coefficients (L2).
* Error Metrics: Similar to Linear Regression, it can use MSE, MAE, and R-squared.

1. **Random Forest :**

* ElasticNet is a linear regression model that combines both L1 (Lasso) and L2 (Ridge) regularization.
* It is used to prevent overfitting in linear regression by introducing a penalty for the number of features selected (L1) and the magnitude of coefficients (L2).
* Error Metrics: Similar to Linear Regression, it can use MSE, MAE, and R-squared.

1. **Extra Trees :**

* Extra Trees, short for Extremely Randomized Trees, is another ensemble learning method similar to Random Forest.
* It builds multiple decision trees with random splits and combines their predictions.
* It's often used for both classification and regression tasks.
* Error Metrics: Similar to Random Forest, it can use MSE, MAE, and R-squared.

1. **SVM :**

* SVM is a supervised machine learning algorithm used for classification and regression tasks.
* In regression, it's used to find the hyperplane that best fits the data points while maximizing the margin between the support vectors.
* Error Metrics: SVM regression uses MSE and R-squared for evaluation.

1. **Gradient Boosted :**

* Gradient Boosted Trees is an ensemble learning technique that combines multiple weak learners to create a strong learner.
* It's used for both classification and regression tasks.
* It fits new trees to the residual errors made by the previous trees.
* Error Metrics: It uses MSE, MAE, R-squared, and others.

1. **Baseline :**

* A baseline model is a simple, straightforward model used as a benchmark or starting point for more complex models.
* In regression, a common baseline is to predict the mean or median of the target variable for all data points.
* Error Metrics: It can use MSE, MAE, and R-squared.

**Experimental Results**

* Pictorial representation of any null data present in the dataset.
* Count Plot for Student Sex Attribute
* Kernel Density Estimation for Age of Students.
* Count Plot for Male & Female students in different age groups.
* Count Plot for students from Urban & Rural Region.
* Does age affect final grade?
* Do urban students perform better than rural students?
* Higher Education vs Final Grade(G3)

**Error Metrics and Model Selection:**

Error metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), Median Absolute Error, and Coefficient of Determination (R-squared) are used to evaluate the performance of regression models.

The choice of error metric depends on the specific problem and the characteristics of the dataset.

To find the best model suited for your dataset, you can train multiple models (e.g., Linear Regression, Random Forest, etc.) and compare their error metrics.

The model with the lowest MSE, MAE, or the highest R-squared is often considered the best fit for the dataset. However, model selection should consider various factors, including model complexity, interpretability, and business requirements.

**Model Selected : Linear Regression**

Linear regression is used for predicting student grades because it offers a simple and interpretable way to estimate academic performance based on factors like previous grades and study time. Its relevance lies in its transparency, which fosters trust, and its effectiveness as a baseline model for educational institutions to monitor and support student progress.

Advantages of Linear Regression:

• Interpretability

• Simplicity

• Transparency

• Baseline Model

• Linearity

• Quick Implementation

• Robustness

**Report**

In the quest to develop a predictive model for student grade outcomes, an array of machine learning algorithms was diligently explored. Through rigorous experimentation and analysis, Linear Regression emerged as the most appropriate choice, achieving exceptional predictive accuracy. The decision to endorse Linear Regression as the ideal model was substantiated by the model's ability to yield extraordinarily low error metrics, specifically, Mean Squared Error (MSE) and Mean Absolute Error (MAE). This report aims to provide a comprehensive overview of the model's attributes, its performance, and the underlying factors that contribute to its effectiveness.

**Model Evaluation:**

Linear Regression's resounding success in this application becomes evident upon scrutinizing its error metrics. The MSE approaches zero, indicating minimal dispersion between predicted and actual student grades. The virtually imperceptible MAE values substantiate the model's exceptional precision in predicting student performance.

**Feature Analysis:**

Our model relies on a judicious selection of influential features, each contributing uniquely to its predictive power. Notable features include the number of past student failures, the educational backgrounds of the student's mother (Medu) and father (Fedu), the student's aspiration for higher education, age, the frequency of socializing with friends, the presence of a romantic relationship, and the rationale behind school choice. These carefully chosen features collectively form the foundation for accurate grade predictions.

**Model Accuracy:**

Remarkably, the Median Absolute Error attains a perfect score of zero, signifying that our model's predictions align precisely with actual student grades. The Coefficient of Determination (R-squared) is a flawless 1.0, highlighting the model's ability to comprehensively elucidate the variations in student performance. Furthermore, the strikingly minute MSE and MAE values affirm the model's precision, bordering on the negligible presence of errors.

**Conclusion**

In summary, our machine learning journey has yielded an exceptionally accurate student grade prediction model. This model holds great promise for educators and institutions, providing deep insights into the multitude of factors influencing student success. By leveraging meticulously chosen features and the strength of Linear Regression, validated by exemplary error metrics, we unveil the immense potential of machine learning in education. This achievement signals a shift in educational practices, offering tailored academic support driven by a thorough understanding of student performance factors. Our report highlights the transformative potential of data-driven approaches in reshaping educational analytics, paving the way for precise predictive modeling and redefining research and intervention strategies in education.