Application of AI in Education

SIDDHI JALAN
SAKSHI PRIYA
ANSHITA VERMA

STUDENT PERFORMANCE ANALYSIS AND PREDICTION (20% OF FOCUS) :

Goal: Analyze student performance data to identify patterns and predict future performance.

Analysis Techniques:

- Perform correlation analysis to identify relationships between performance in different subjects.
- Explore regression models to predict student performance based on background factors (if available in datasets).

Repository: https://github.com/anshita-21/SIT-ICOE-HACKATHON

---

1 Contents

2 Introduction & Motivation

Changing education is sometimes the most important thing, with predictive skills and understanding of student performance playing a vital role in creating effective teaching strategies that are tailored to each individual student. The advent of artificial intelligence (AI) technology has made it possible for us to explore deeper into students' data thus bringing about groundbreaking insights that can transform educational practices. This project aims at examining the complex ways through which AI may be employed in education, focusing on both student performance analysis and predictability on future academic trends so that it contributes to improving learning experiences as well as educational results in general.

3 Literature Review

The rising area of AI in education has seen a lot of academic works that have been useful in different aspects like individual study to predictive analytics. In particular, the studies by Baker and Siemens (2014) show how effective is predictive modeling while identifying students at risk for dropping out of school and what interventions can be put in place. Romero and Ventura (2013) also conducted a study on data mining highlighting how it could provide information to inform instructional design and curricula development, amongst others. Furthermore, Artificial Intelligence driven adaptive learning systems such as Khan Academy and Duolingo create an opportunity for personalizing education based on an individual's level of knowledge or learning style.

4 Methodology

a *Data Acquisition*

The initial step will be collecting student achievement data that is important from educational systems or from online learning platforms, which can include data such as academic records, demographic information and various subject assessment scores.

b *Model Development*

A range of AI methods are used to construct models that predict the future. Correlation analysis in schools examine associations between academic variables that identify potential influences. Following this are regression models like linear regression and decision trees which help in anticipating students' performances

using a combination of background factors such as socioeconomic status, previous academic achievements and learning styles.

C *Evaluation*

There is a need for rigorous evaluation processes to determine the efficiency of the developed models. This involves splitting data into training and testing sets and use of cross-validation approaches. Performance measures consist of accuracy, precision, recall and F1-score to form a complete picture about how well a model performs.

5 *Dataset Description*

The datasets used in this study consist of two separate files: student-mat.csv, representing the Math course, and student-por.csv, representing the Portuguese language course. Each dataset contains a set of attributes describing various aspects of the students' background and academic performance.

Data Source: Public datasets like UCI Machine Learning Repository's Student Performance dataset (https://archive.ics.uci.edu/ml/datasets/Student+Performance)

Table 1: The preprocessed student related variables

Attribute	Description (Domain)
school	student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
sex	student's sex (binary: "F" - female or "M" - male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: "U" - urban or "R" - rural)
famsize	family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
Pstatus	parent's cohabitation status (binary: "T" - living together or "A" - apart)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
Fjob	father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
reason	reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
guardian	student's guardian (nominal: "mother", "father" or "other")
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

Attribute	Description (Domain)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)

These grades are related with the course subject, Math or Portuguese:

G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20, output target)

6 *Data Cleaning: Handling Missing Values*

Before proceeding with the analysis and modeling, it's essential to check for missing values in the dataset and handle them appropriately to ensure the accuracy and reliability of the results.

- Imputation: Replace missing values with a suitable statistic such as mean, median, or mode.
- Deletion: Remove rows or columns with missing values if they are insignificant.
- Prediction: Utilize predictive models to estimate missing values based on other available data.
- Domain-specific Handling: Apply domain-specific knowledge to handle missing values appropriately.

count of missing values for each column in the dataset is null.

7 *Regression Analysis*

Regression analysis is a mathematically measured correlation of variables used as a predictive modelling method. You use regression modelling to predict numerical values depending on various inputs. For example, you can understand the relationship between an independent and dependen variable, allowing you to predict how the dependent variable changes along with its independent counterpart. In this case, the dependent variable is what you're measuring and the independent variable is the factor that causes change.

In business, regression analysis can help forecast trends, predict strengths and areas of weakness or establish cause-and-effect relationships to make informed business decisions and strategic plans. You often calculate regression analysis through machine learning or artificial intelligence, though there are also mathematical equations you can use. There are different analysis types that you can use based on the nature of the variables you're predicting and what information you'd like to gather from your analysis.

7.1 *Linear Regression Model for Predicting Student Grades*

Linear regression is a fundamental statistical technique used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). In the context of predicting student grades, linear regression can be employed to analyze how various factors such as parental education, study time, and previous grades influence the final grade achieved by students.

ASSUMPTIONS:

- The variables should be measured at a continuous level. Examples of continuous variables are time, sales, weight and test scores.
- Use a scatterplot to find out quickly if there is a linear relationship between those two variables.
- The observations should be independent of each other (that is, there should be no dependency).
- Your data should have no significant outliers.
- Check for homoscedasticity — a statistical concept in which the variances along the best-fit linear-regression line remain similar all through that line.

a *Data Preprocessing*

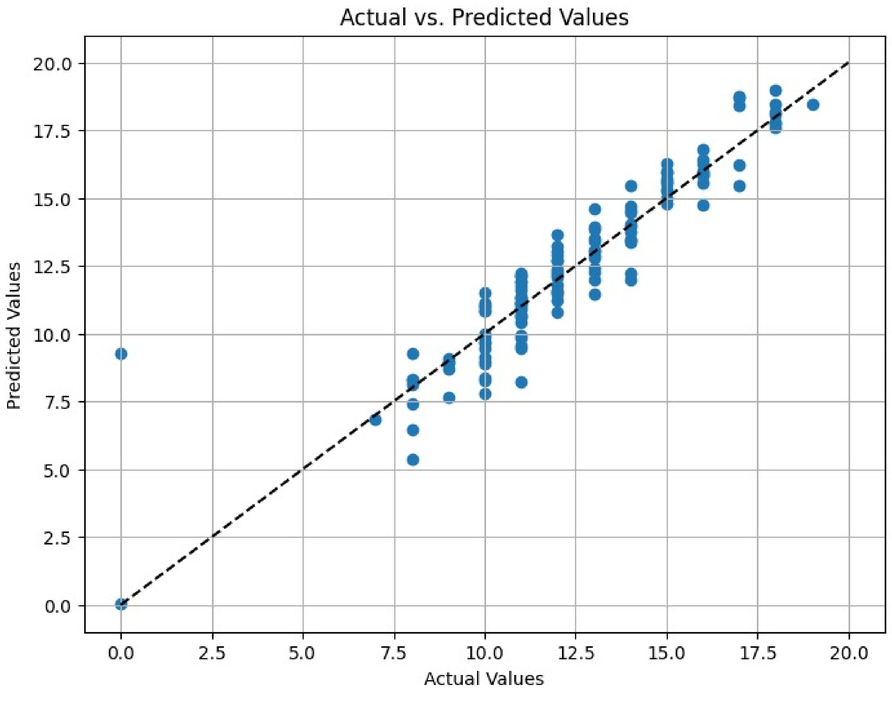
- Feature Selection: The dataset is loaded and split into features (X) and the target variable (y), where the target variable represents the final grade (G3) obtained by students.
- Identification of Variable Types: Categorical variables are identified as ordinal or nominal, based on their nature and potential impact on the target variable.
- Preprocessing Steps: Ordinal variables are standardized, while nominal variables are encoded using one-hot encoding to convert them into numerical format.

b *Model Building*

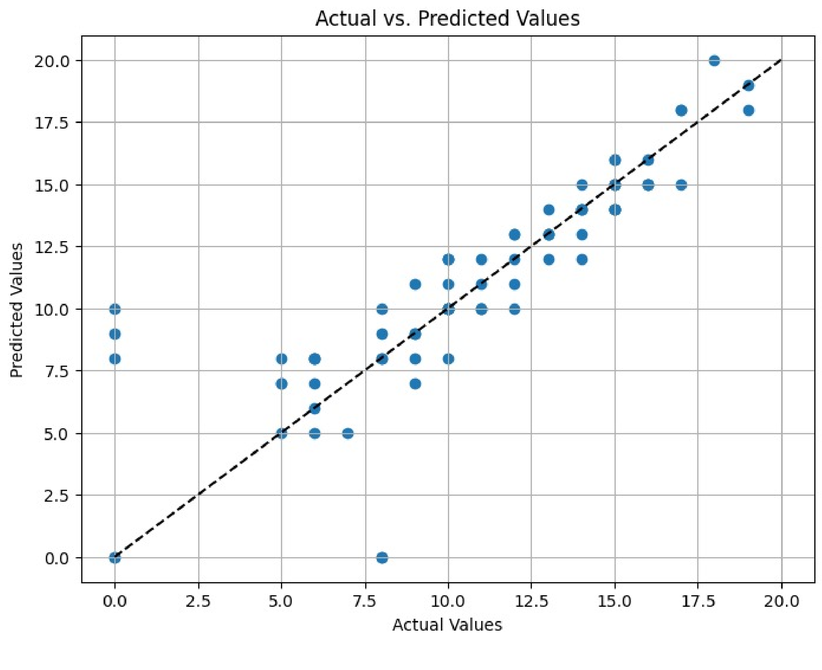
- Pipeline Construction: A pipeline is constructed to sequentially apply preprocessing steps and train the linear regression model.
- Preprocessor: A ColumnTransformer is used to apply one-hot encoding to nominal variables and standardize ordinal variables.
- Model: Linear regression model is instantiated and included in the pipeline.

c *Training and Evaluation*

- Data Splitting: The dataset is split into training, validation, and testing sets using the train_test_split function from sklearn.model_selection.
- Model Fitting: The pipeline, comprising preprocessing and linear regression model, is fitted to the training data.
- Model Evaluation: The pipeline's performance is evaluated on the testing data using the coefficient of determination (R^2 score) to assess how well the model predicts the final grades.



Mathematics



Portugese

7.2 *Decision Tree Regression for Predicting Student Grades*

In this analysis, we aim to predict students' final grade (`G3`) in mathematics by utilizing various features from the "student-mat.csv" dataset. We employ a machine learning pipeline that includes preprocessing steps and a decision tree regression model.

a *Data Preprocessing*

The dataset contains several features, including both categorical and numerical variables. Categorical variables are further classified into ordinal and nominal types. Ordinal variables, such as `Medu` (mother's education) and `Fedu` (father's education), have a natural order, while nominal variables, such as `school` and `sex`, do not.

To handle these variables appropriately, we use a ColumnTransformer` with two transformers:

- OneHotEncoder` for nominal variables to convert them into a format suitable for the model.
- StandardScaler` for ordinal variables to standardize their values.

b *Model Building*

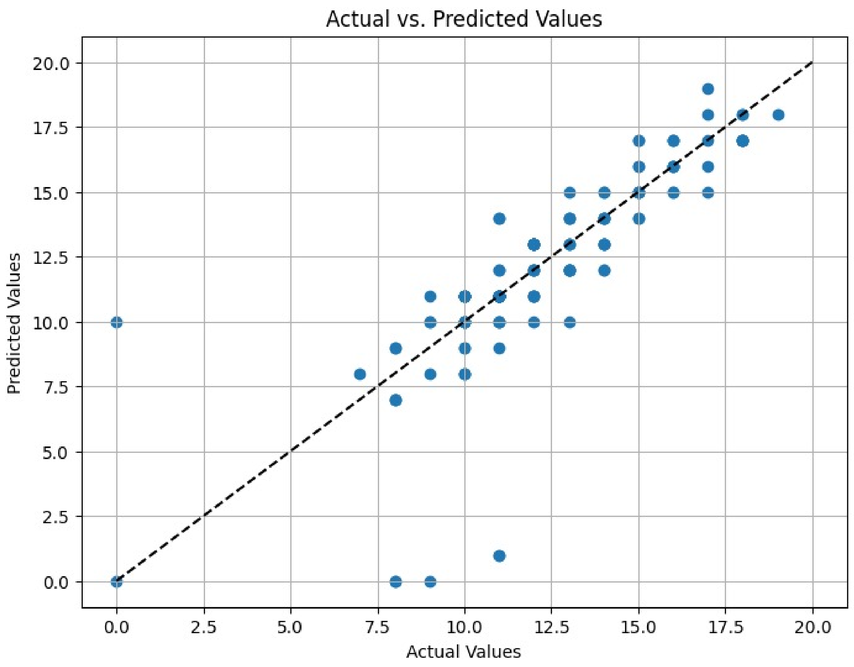
A `DecisionTreeRegressor` is chosen for its ability to capture non-linear relationships and interactions between features. The model is incorporated into a `Pipeline` along with the preprocessing steps to streamline the workflow

c *Training and Testing*

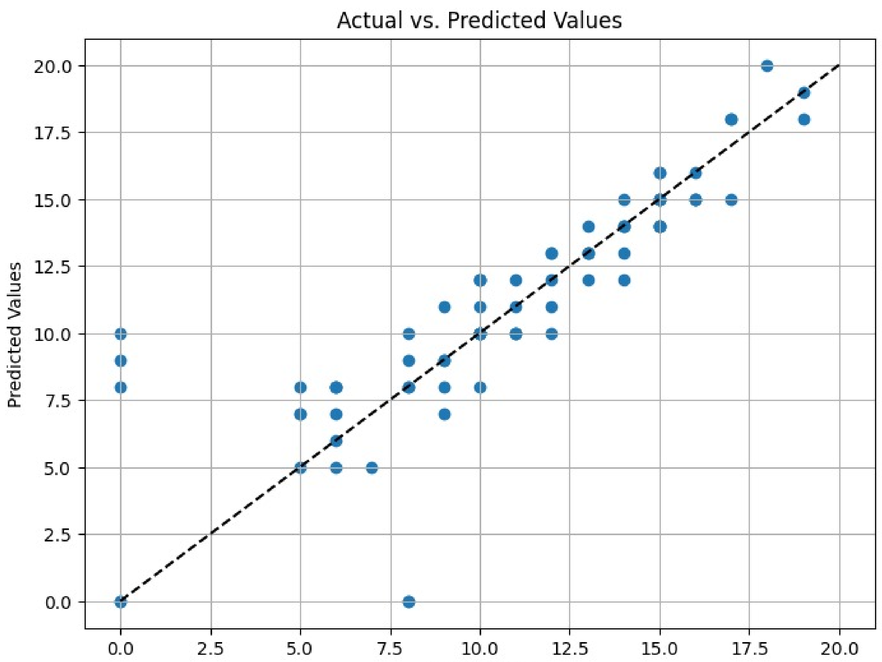
The dataset is split into training, validation, and testing sets using `train_test_split`. The model is trained on the training set, and its performance is evaluated on the testing set.

d *Model Evaluation*

The model's predictions are compared against the actual `G3` values. A scatter plot visualizes the relationship between the predicted and actual values, with a diagonal line representing the ideal prediction scenario. The mean squared error (MSE) is calculated to quantify the model's accuracy.



Mathematics



Portugese

7. *RandomForest Regression for Predicting Student Grades*

Aims to predict students' final grades (`G3`) in a math course using a dataset named "student-mat.csv". We employ a RandomForestRegressor within a machine learning pipeline

a *Data Preprocessing*

The dataset comprises both categorical and numerical variables. Categorical variables are divided into ordinal and nominal categories. Ordinal variables, such as `Medu` (mother's education) and `Fedu` (father's education), have a natural order, while nominal variables, such as `school` and `sex`, do not.

- ColumnTransformer is used to apply appropriate preprocessing to these variables:
- OneHotEncoder transforms nominal variables into a format suitable for modeling.
- StandardScaler standardizes the ordinal variables to have a mean of zero and a standard deviation of one.

b *Model Building*

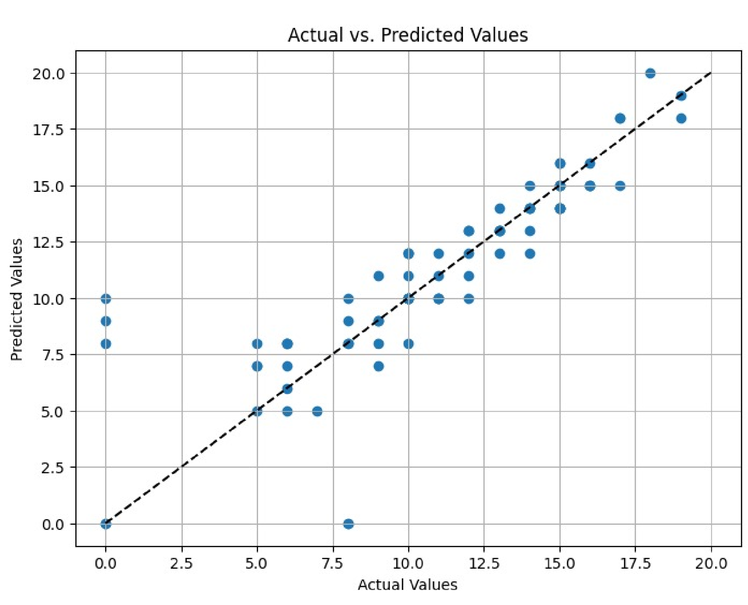
The RandomForestRegressor is chosen for its robustness and ability to handle complex interactions between features. It is integrated into a `Pipeline` with the preprocessing steps, ensuring a seamless workflow from data transformation to model training.

c *Training and Testing*

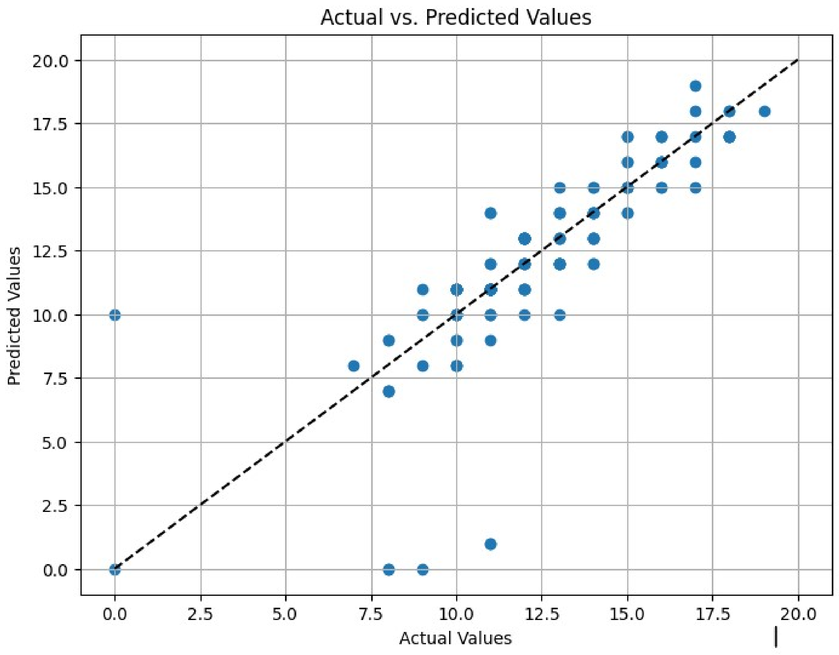
The data is split into training, validation, and testing sets. The model is trained on the training set and evaluated on the testing set to assess its performance.

d *Model Evaluation*

The RandomForest pipeline provides a comprehensive approach to predicting student grades, leveraging the strengths of ensemble learning. The preprocessing steps ensure that the model receives well-formatted and standardized input, leading to more reliable predictions.



Mathematics



Portugese

7.4 *Support Vector Regression (SVR) for Predicting Student Grades*

In this analysis, we aim to predict students' final grades (G3) in a mathematics course using the "student-mat.csv" dataset. We employ the Support Vector Regression (SVR) model within a machine learning pipeline that includes preprocessing steps for both categorical and numerical features.

a *Data Preprocessing*

- We use a ColumnTransformer to apply different preprocessing techniques to the ordinal and nominal variables.
- For nominal variables, we apply one-hot encoding using OneHotEncoder.
- For ordinal variables, we standardize their values using StandardScaler.
- The remainder="passthrough" ensures that any remaining features are passed through without transformation.

b *Model Building*

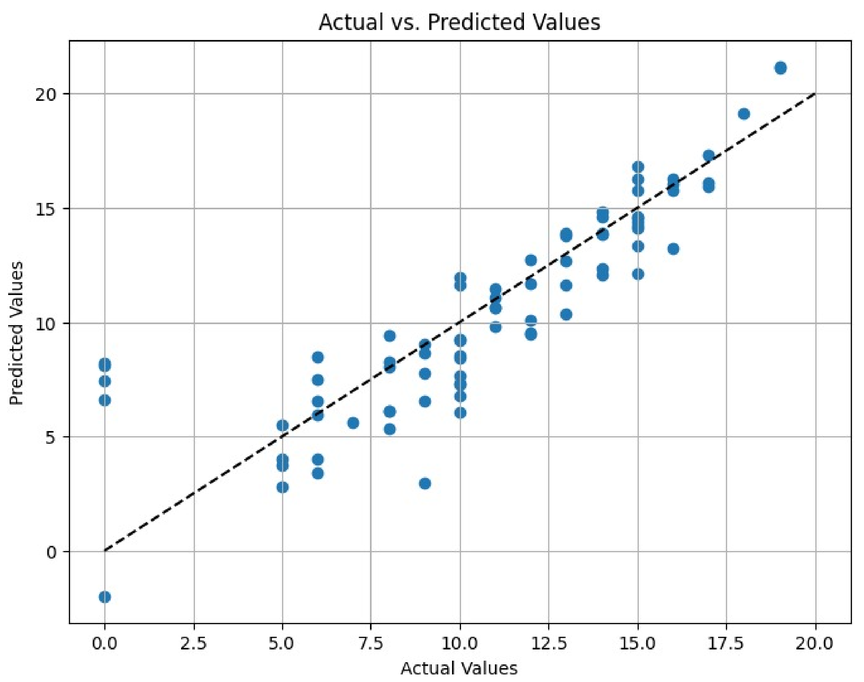
- We choose the Support Vector Regression (SVR) model for its ability to handle non-linear relationships and outliers.
- The SVR model aims to find a hyperplane that best fits the data while minimizing the error.
- The model is integrated into a Pipeline along with the preprocessing steps, ensuring a streamlined workflow.

c *Training and Testing*

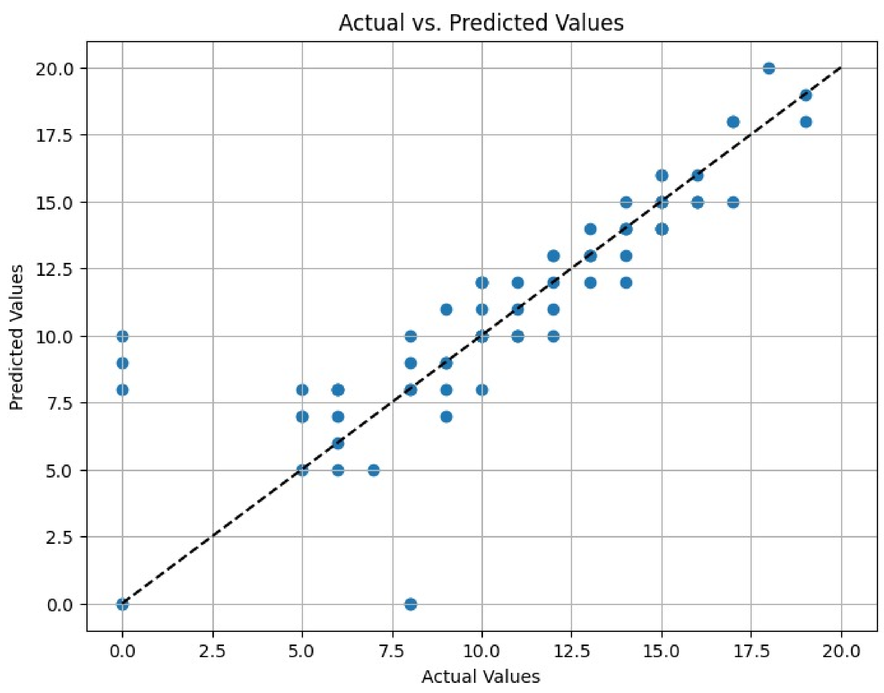
- We split the data into training, validation, and testing sets using train_test_split.
- The model is trained on the training set (X_train, y_train) and evaluated on the testing set (X_test, y_test)

d *Model Evaluation*

- We calculate the R-squared value (coefficient of determination) to assess how well the model explains the variance in the target variable.
- A scatter plot visualizes the relationship between the predicted and actual G3 values, with a diagonal line representing the ideal prediction scenario.



Mathematics



Portugese

Regression is a data science task of predicting the value of target (numerical variable) by building a model based on one or more predictors (numerical and categorical variables).

Thus, through the above regression models we have succesfully predicted G3 student performance.

8 Correlation Analysis

A correlation is a statistical measure of the relationship between two variables. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the data can be visually represented in a scatterplot. Using a scatterplot, we can generally assess the relationship between the variables and determine whether they are correlated or not.

Correlation analysis, also known as bivariate, is primarily concerned with finding out whether a relationship exists between variables and then determining the magnitude and action of that relationship.

8.1 *Spearman Correlation Analysis of Student Grades in Mathematics and Portuguese*

a *Introduction*

Spearman correlation analysis is a statistical method used to evaluate the strength and direction of the relationship between two variables. In the educational context, it can help determine if there is a significant correlation between student grades in Mathematics and Portuguese courses. This analysis provides insights into the extent to which performance in one subject may influence performance in another.

b *Defining Columns*

We define two sets of columns:

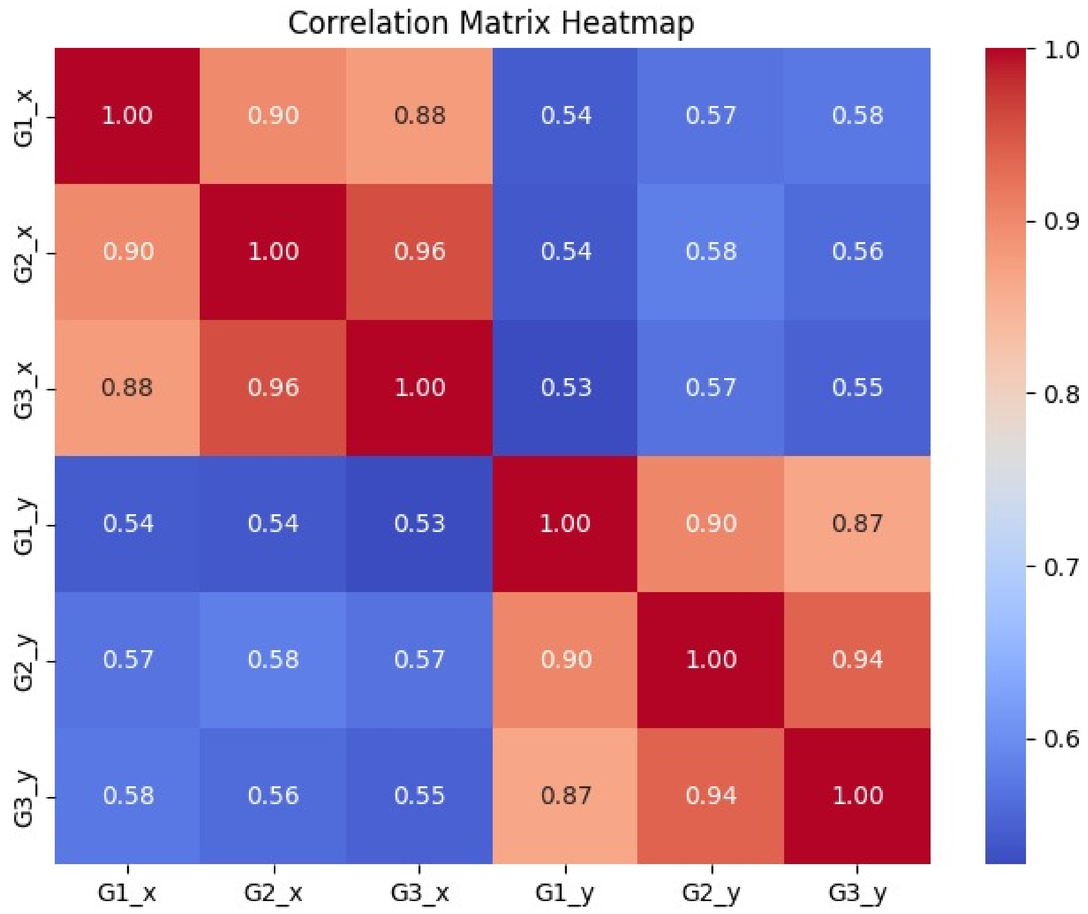
- math_columns : Represents grades in mathematics (`G1_x`, `G2_x`, `G3_x`).
- portuguese_columns : Represents grades in Portuguese (`G1_y`, `G2_y`, `G3_y`).

c *Correlation Matrix Calculation*

- Spearman correlation coefficient is computed between the specified columns using the corr method with method='spearman'.
- The coefficient ranges from -1 to 1, indicating the strength and direction of the correlation:
  - ρ = 1: Perfect positive correlation
  - ρ = -1: Perfect negative correlation
  - ρ = 0: No correlation

d *Visualization*

To better interpret the results, we visualize the correlation matrix using a heatmap from the `seaborn` library. The heatmap provides a color-coded representation of the correlation coefficients, making it easier to identify strong or weak relationships.



b *Insights*

The correlation analysis reveals insights into how grades in mathematics and Portuguese are related. This information can be valuable for educators to understand the interdependencies between different subjects and how they impact student performance.

Strong Positive Correlation Within Subjects:

- The diagonal elements (the correlation of each subject with itself) are all 1. This is expected since a variable perfectly correlates with itself.
- Within each subject (G1, G2, G3), there is a strong positive correlation between grades. For example, G1_x has a correlation of around 0.88 to 0.90 with G2_x and G3_x.

Moderate Positive Correlation Between Subjects:

- There is a moderate positive correlation between corresponding grades in mathematics and Portuguese. For example:
- G1_x has a correlation of around 0.54 to 0.57 with G1_y, indicating a moderate positive relationship between the grades in the first grading period for each subject.

Correlations Across Grading Periods:

- The correlation coefficients between subjects tend to be fairly consistent across grading periods. For instance, the correlation between G1_x and G1_y is similar to the correlation between G2_x and G2_y, and so on. This suggests a consistent relationship between the grades in different subjects across different grading periods.

Lower Correlation Between Different Grading Periods of the Same Subject:

- The correlation coefficients between different grading periods of the same subject are slightly lower compared to the correlations between corresponding grading periods of different subjects. This could indicate some variability or differences in performance between grading periods within the same subject.

8.2 *Pearson Correlation Analysis of Student Grades in Mathematics and Portuguese*

a *Introduction*

Pearson correlation analysis is a statistical method used to measure the strength and direction of the linear relationship between two continuous variables. In the context of student performance analysis, Pearson correlation can help determine if there is a significant linear correlation between grades obtained in Mathematics and Portuguese courses.

b *Defining Columns*

We define two sets of columns:

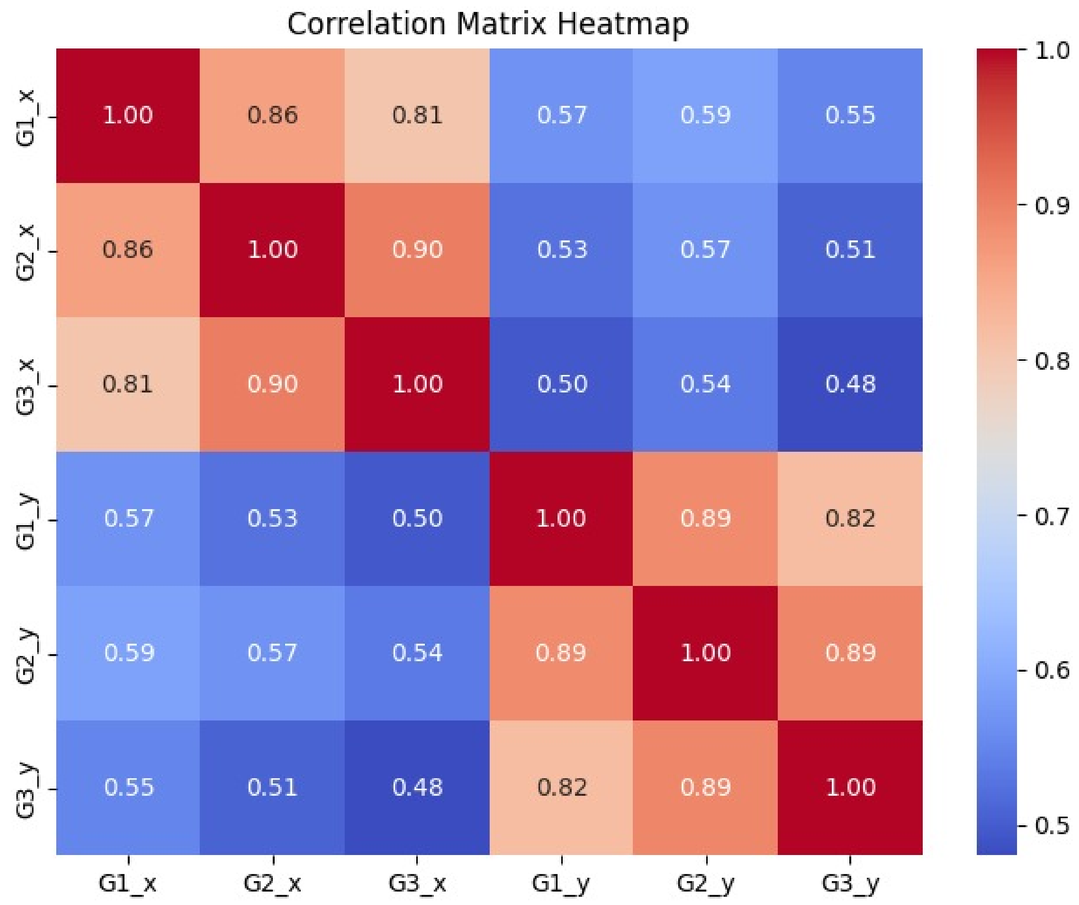
- math_columns : Represents grades in mathematics (`G1_x`, `G2_x`, `G3_x`).
- portuguese_columns : Represents grades in Portuguese (`G1_y`, `G2_y`, `G3_y`).

c *Correlation Matrix Calculation*

- Pearson correlation coefficient is computed between the specified columns using the corr method with method='pearson'.
- The coefficient ranges from -1 to 1, where:
- r = 1 indicates a perfect positive linear correlation.
- r = -1 indicates a perfect negative linear correlation.
- r = 0 indicates no linear correlation.

d *Visualization*

- The correlation matrix is visualized using a heatmap generated with Seaborn.
- The heatmap provides a color-coded representation of correlation coefficients, with warmer colors indicating stronger positive correlations and cooler colors indicating stronger negative correlations.



b *Insights*

The correlation analysis using Pearson's method provides valuable insights into the linear associations between student grades in mathematics and Portuguese. Such information can be instrumental for educational strategies and interventions.

Strong Positive Correlation Within Subjects:

- As with the Spearman correlation, the diagonal elements (the correlation of each subject with itself) are all 1, indicating a perfect correlation.
- Within each subject (G1, G2, G3), there is a strong positive correlation between grades. For example, G1_x has a correlation of around 0.80 to 0.86 with G2_x and G3_x.

Moderate Positive Correlation Between Subjects:

- There is a moderate positive correlation between corresponding grades in mathematics and Portuguese, similar to the Spearman correlation. For example:
- G1_x has a correlation of around 0.55 to 0.57 with G1_y, indicating a moderate positive relationship between the grades in the first grading period for each subject.

Correlations Across Grading Periods:

- The correlation coefficients between subjects tend to be fairly consistent across grading periods, similar to the Spearman correlation. This suggests a consistent relationship between the grades in different subjects across different grading periods.

Lower Correlation Between Different Grading Periods of the Same Subject:

- As with the Spearman correlation, the correlation coefficients between different grading periods of the same subject are slightly lower compared to the correlations between corresponding grading periods of different subjects. This could indicate some variability or differences in performance between grading periods within the same subject.

Overall, the Pearson correlation matrix provides similar insights to the Spearman correlation matrix, indicating a positive relationship between grades in mathematics and Portuguese, with slightly different correlation coefficients due to their emphasis on linear relationships.

9 Results and Analysis

Findings derived from the analysis of student performance data are disseminated through the results section. A range of visualizations illustrate links between academic variables, while regression models make future performance predictions. Also, it is important to note that evaluation metrics help in assessing how effective predictive models are for educators and policy makers who thereby can obtain actionable information to customize interventions and support systems for students.

Since we suspected that the G1 and G2 grades would have a high impact, three input configurations were tested for each model:

- **A-with all variables from Table 1 except G3 (the output);**
- **B-similar to A but without G2 (the second period grade); and**
- **C- similar to B but without G1 (the first period grade).**

9.1 Regression Model Performance

- Linear Regression: This model provides a baseline for predicting student performance based on linear relationships between input features and output grades. While simple and interpretable, it may not capture complex nonlinear patterns in the data.
- Random Forest Regression: By utilizing an ensemble of decision trees, the Random Forest model can capture nonlinear relationships and interactions between features, potentially leading to improved prediction accuracy compared to linear regression.
- Decision Tree Regression: Decision trees offer a straightforward way to understand the decision-making process of the model. However, they may suffer from overfitting and lack generalization ability, especially when dealing with complex datasets.
- Support Vector Regression (SVR): SVR aims to find the optimal hyperplane that best separates data points while maximizing the margin. It can handle nonlinear relationships through the use of kernel functions, making it suitable for capturing complex patterns in the data.

Among the regression models, the Random Forest Regression model demonstrates the highest prediction accuracy and robustness to nonlinear patterns in the data. Its ensemble nature allows it to capture complex relationships between input features and output grades, making it a suitable choice for predicting student performance.

8. Correlation Analysis Results:

- Pearson Correlation: Pearson correlation measures the linear relationship between two continuous variables. It is suitable for identifying linear associations between academic factors but may not capture nonlinear relationships.
- Spearman Correlation: Spearman correlation assesses the monotonic relationship between variables, making it more robust to outliers and non-normal distributions. It can capture both linear and nonlinear associations, providing a more comprehensive understanding of the data.

For correlation analysis, Spearman correlation may be preferred as it can capture both linear and nonlinear associations between academic variables, providing a more nuanced understanding of their relationships.

10 Future Implementation Plan

The project has plans of expanding as well as refining in the future. On one hand, this integration allows dynamic predictions and interventions due to real time data streams, which is conducive to adaptability in educational practices. Additionally, more developed machine learning algorithms incorporating neural networks and ensemble methods have improved

prediction accuracy dramatically. This deployment will also involve continuous monitoring of such models in education set up through strong feedback loops for improving them.

a *Roadmap for Further Development and Deployment*

1. Model Refinement and Optimization:
   - Conduct thorough analysis of model performance metrics and identify areas for improvement.
   - Explore advanced algorithms and techniques to enhance prediction accuracy and generalization.
2. Integration of Real-Time Data Streams:
   - Develop robust data pipelines to ingest, process, and integrate real-time data streams from educational platforms.
   - Design mechanisms for handling streaming data efficiently and updating the predictive model in near real-time.
3. Continuous Monitoring and Feedback Mechanisms:
   - Establish monitoring dashboards and performance metrics to track the deployed model's performance.
   - Implement feedback loops to gather input from educators and stakeholders for ongoing model improvement.

11 Conclusion

Consequently, AI and education merging creates a realm of possibilities that is transformative, transcending the conventional frameworks and ushering into an era of personalized learning and improved educational outcomes. With this, teachers can effectively predict students who are likely to fail in advance through data analysis hence intervening quickly to prevent them from dropping out. This project demonstrates how artificial intelligence can be utilized to improve the way of teaching and enable us to succeed in today's digital world.

12 References

- BreimanL.,2001.RandomForests.MachineLearning, 45,no.1,5–32.
- BreimanL.;FriedmanJ.;OhlsenR.;andStoneC.,1984. Classification and Regression Trees. Wadsworth, Monterey,CA
- Minaei-Bidgoli B.; KashyD.; Kortemeyer G.; and PunchW.,2003.Predictingstudentperformance: an application of dataminingmethodswith an educa tionalweb-basedsystem. InProc.of IEEEFrontiers inEducation.Colorado,USA,13–18.