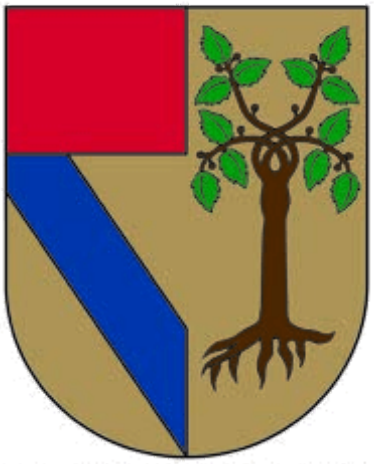


UNIVERSIDAD PANAMERICANA®

SG1_Team#2

Course:

Simulation & Data Visualization

Due date:

Thursday, May 29, 2025

Professor:

Gabriel Castillo Cortés

Team members:

Renata Calderón Mercado

Regina Ochoa Gispert

Miguel Angel Velez Olide

Repository:

https://github.com/Migue0244655/SG1_Team2.git

Introduction

Currently, students' academic performance is a constant concern for both educators and families. Multiple factors can influence academic success or failure, but among the most relevant are study habits and regular class attendance. Understanding how these variables impact academic performance can help design effective strategies to improve education and support at-risk students.

In this project, we use Machine Learning to explore a dataset about students' habits and performance. The choice of this topic stems from the conviction that technology and data analysis can reveal hidden patterns that might go unnoticed at first glance. Furthermore, Machine Learning allows us to build predictive models capable of anticipating a student's academic performance based on their habits, thus facilitating early identification of those who might need additional support.

Guided by this motivation, we set out to discover how variables such as time dedicated to studying and school absences affect the probability of passing or failing. To achieve this, we applied data visualization techniques that allowed us to clearly observe the relationship between these variables and academic performance. Through a classifier based on the K-Nearest Neighbors algorithm, we were able to not only identify patterns but also make concrete predictions about the academic success of new students.

This analysis not only provides valuable knowledge for educators and students, but also demonstrates the power of artificial intelligence to transform data into practical, evidence-based decisions. Understanding these patterns can make a difference in the academic lives of many young people, and that is why we consider this project relevant and worthy of attention.

Development

To better understand the dataset and prepare the foundation for analysis, we implemented a specific method to generate a complete data dictionary. This dictionary contains metadata and descriptive statistics for each variable, which allows us to know in detail the structure and characteristics of the data before applying any Machine Learning techniques.

The create_data_dictionary method performs an automatic analysis of each variable in the dataset, classifying it according to its type (numerical or categorical) and calculating the most relevant statistics for each case.

For numerical variables, it calculates:

- Mean
- Median
- Standard deviation
- Minimum and maximum
- Range (difference between maximum and minimum)

For categorical variables, it determines:

- Unique values
- Most frequent value and its frequency
- Frequency distribution for all values

Additionally, it records the total number of observations, valid values, and missing values for each variable, which allows us to identify possible data quality issues.

The dictionary was built from a prior definition of the variables, where each one has its type assigned, a brief description, and the domain to which it belongs. This facilitates the interpretation and organization of subsequent analysis.

With this methodology, we were able to generate a detailed summary of the dataset, which served as a starting point to detect patterns and prepare the data for predictive modeling.

Continuing with the exploratory analysis, we implemented the print_summary method to obtain a **concise and clear summary** of the dataset. This method prints key information to the console that allows us to quickly evaluate the main characteristics of the dataset.

The summary includes:

- The total number of students in the dataset, which gives us an idea of the size and representativeness of the sample.
- The total count of variables, indicating how many are numerical and how many are categorical, to understand the diversity of data types we are working with.

- Relevant statistics of selected variables that we consider crucial for the analysis, such as the final exam score (exam_score), daily study hours (study_hours_per_day), and class attendance percentage (attendance_percentage). For these variables, the mean and range are shown, which help understand the central tendency and dispersion of the data.

This summary not only facilitates quick interpretation of the data, but also serves as a reference point to validate the quality and consistency of the dataset before proceeding with the cleaning, visualization, and modeling stages.

Method create_distributions_plot

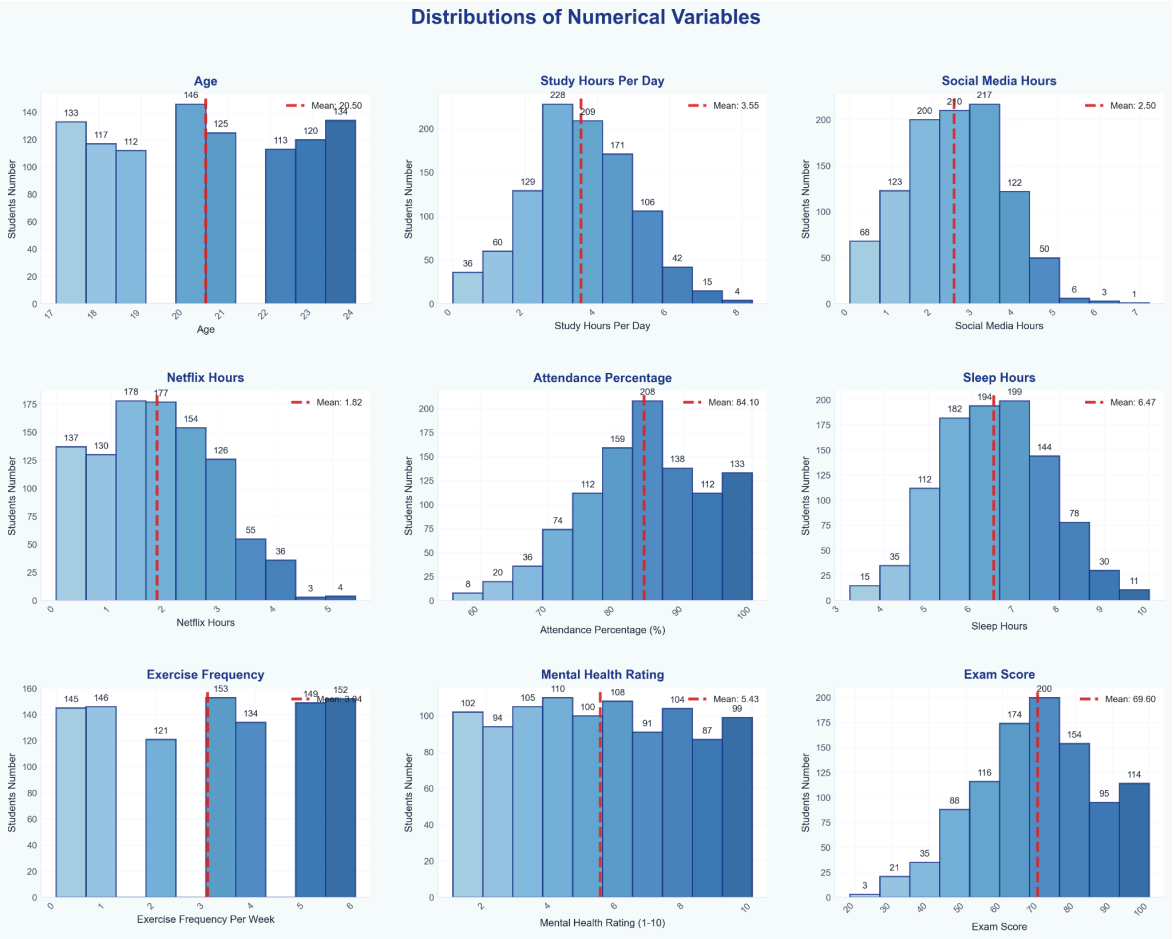
To deepen the exploratory analysis of numerical data, we developed the create_distributions_plot method, whose purpose is to visualize the distribution of up to nine key numerical variables from the dataset. This allows us to better understand the dispersion, concentration, and trends of variables that impact student performance and habits.

Why did we create this method?

Visual exploration is a fundamental tool for detecting patterns, anomalies, and relevant characteristics in data that are not evident with numerical statistics alone. By representing the distributions of variables such as daily study hours, attendance percentage, hours dedicated to social media, or final exam scores, we can:

- Identify if the data is skewed or has a uniform distribution.
- Detect the presence of outliers or inconsistencies.
- Observe variability among students and how certain habits cluster.
- Guide the cleaning process and variable selection for subsequent modeling.

This visual approach allowed us to detect, for example, that while most students dedicated between 2 and 4 hours to daily study, a significant percentage had low exam scores, which suggested investigating more about other factors such as attendance and mental health. Thus, these graphs were a key foundation for exploration and subsequent predictive analysis.



Method create_categorical_plots

This method was designed to visualize the distribution of categorical variables from the dataset, clearly showing the frequency and proportion of each category within a set of up to six relevant categorical variables.

Why did we implement this method?

Categorical variables contain qualitative information that cannot be directly analyzed with conventional numerical statistics. Therefore, we use bar charts to understand how categories are distributed among students. This helps us to:

- Identify the most common and least frequent categories.
- Detect possible imbalances or biases in the categories.
- Obtain a clear perspective of the composition of the studied group in qualitative variables.

- Support the interpretation of results and planning of subsequent analyses based on these variables.

These graphs allowed us to observe, for example, how categories of extracurricular activity attendance are distributed or the proportion of students according to school type. This provides a clear view of the group's heterogeneity and helps focus statistical and predictive analyses based on these variables.

Method create_correlation_matrix

This method aims to calculate and visualize correlations between all numerical variables in the dataset through a correlation matrix represented with a heatmap.

Why did we implement this method?

Correlations between numerical variables are fundamental to:

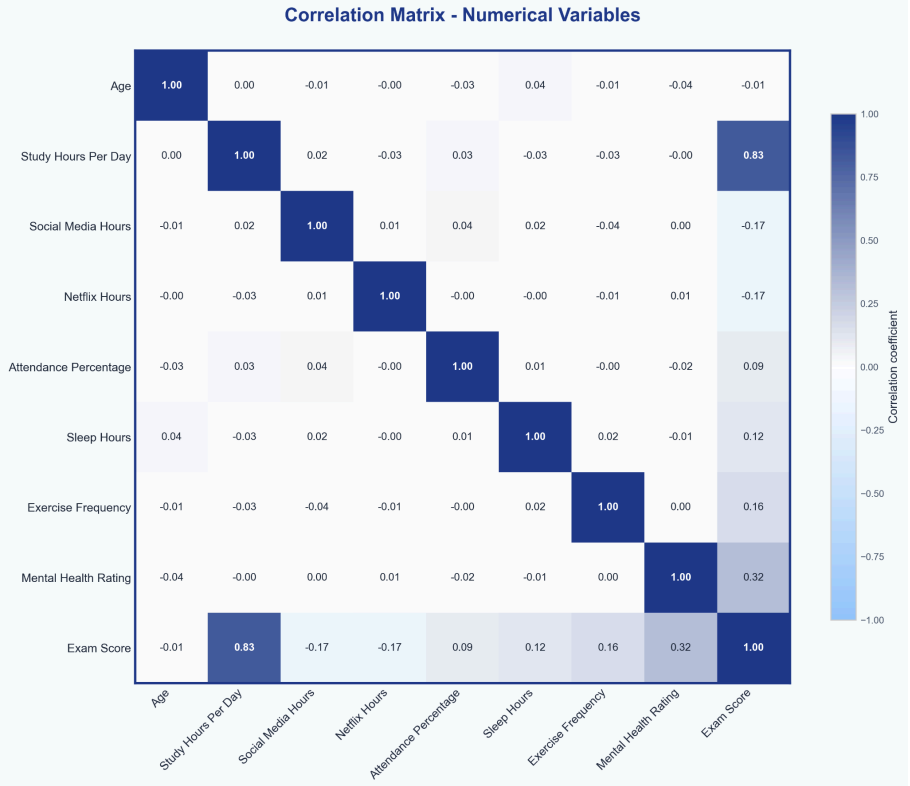
- Identify linear relationships between variables, for example, if an increase in one variable is associated with an increase or decrease in another.
- Detects variables that are highly related, which can be useful for simplifying predictive models by eliminating redundant variables.
- Explore possible dependencies or hidden patterns within the data, helping to better understand the behavior of the study group.

- Facilitate decision-making about which variables may be most relevant for subsequent analyses or for building statistical or machine learning models.

Description of the generated graph:

- Correlation matrix: Each cell shows the Pearson correlation coefficient between two numerical variables, with values between -1 (perfect negative correlation) and 1 (perfect positive correlation).
- Heatmap: The color of each cell reflects the strength and direction of the correlation:
    - Intense blue tones indicate strong negative correlations.
    - White represents null or near-zero correlation.
    - Reddish tones indicate strong positive correlations.

This analysis allows us, for example, to discover if variables such as study hours and exam scores are strongly related or if school attendance has any association with performance. This knowledge guides interpretation and subsequent steps in data analysis.



Method create_performance_analysis

With the objective of analyzing more deeply the relationship between different factors and students' academic performance, we developed the create_performance_analysis method. This method generates a set of graphs that allows visual observation of how key variables such as study hours, school attendance, gender, and having a part-time job influence final exam scores.

Why was this method implemented?

A comparative visualization between explanatory variables and academic performance is crucial for:

- Detecting linear or non-linear relationships between individual factors and obtained scores.
- Identifying significant differences between groups (for example, by gender or employment conditions).
- Evaluating the relative influence of each variable on academic performance.
- Generating visual evidence that complements statistical analysis and supports the predictive modeling process.

Description of generated graphs

The method generates a figure composed of four subplots arranged in a 2x2 grid, each addressing a different dimension of the analysis:

Study hours vs. exam score A scatter plot is used to represent the relationship between daily study hours and the score obtained on the final exam. Additionally, a linear regression line is superimposed to visualize the general trend. The Pearson correlation coefficient (r) is calculated and displayed on the graph, providing a quantitative measure of the relationship between both variables.

School attendance vs. exam score Similar to the previous graph, this scatter plot shows how the percentage of class attendance relates to exam results. A regression line is also included to facilitate visual interpretation of the correlation.

Academic performance by gender A box plot is presented that compares the distribution of exam scores between different gender categories. This analysis allows identification of possible performance differences and evaluation of the dispersion and median of each group.

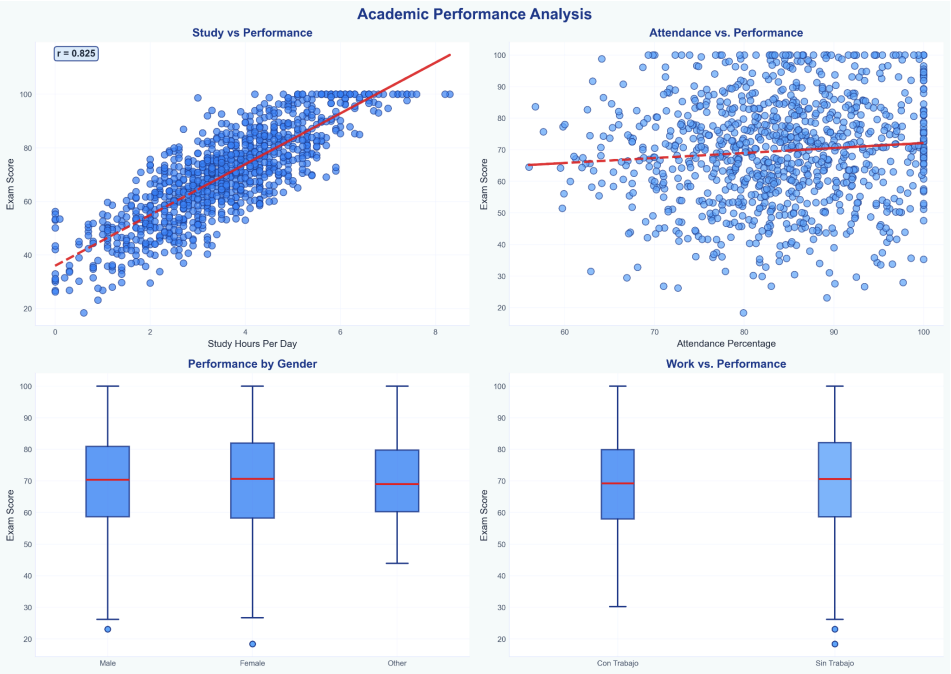
Academic performance and part-time employment Another boxplot compares the scores of students who have a part-time job with those who don't. This graph helps explore whether work load significantly affects academic performance.

Expected results

Through these graphs, visual analysis facilitates the identification of important patterns, such as:

- The positive correlation between study time and academic performance.
- The impact of school attendance on exam results.
- Possible gender gaps in performance.
- The influence of external employment on students' academic capacity.

This graphical representation was consolidated as a key tool within the project's exploratory analysis, allowing us to support hypotheses with clear visual evidence and serve as a guide for future modeling phases and decision-making.



Method create_lifestyle_analysis

The create_lifestyle_analysis method was designed to analyze how various lifestyle habits impact students' academic performance. Through a visual approach, this analysis seeks to identify correlations between lifestyle-related factors and scores obtained on the final exam.

Why was this method implemented?

In the context of machine learning and predicting student performance, it is fundamental to understand which non-academic variables can influence results. This method allows for:

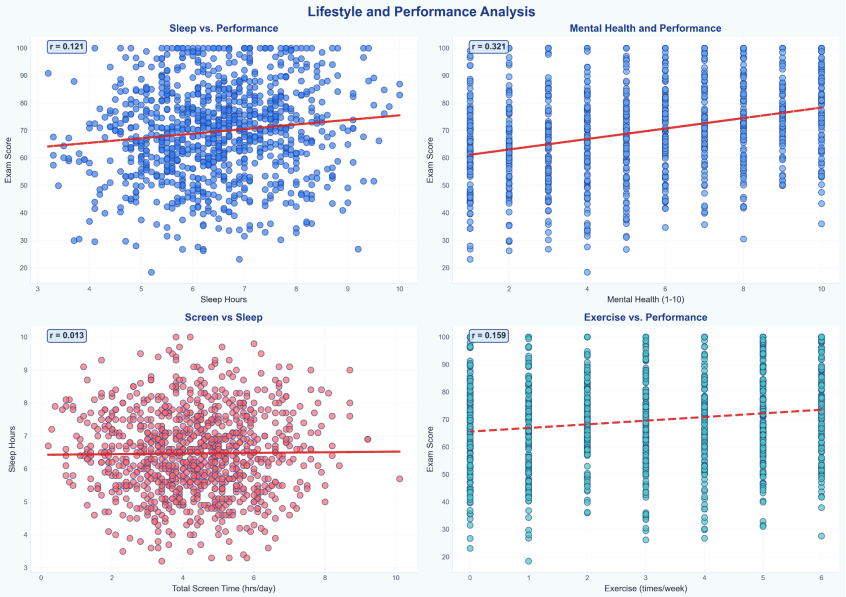
- Evaluating the impact of sleep, mental health, screen time, and exercise on academic performance.
- Determining correlations that could be integrated as predictive features in machine learning models.
- Offering a more comprehensive perspective of the student, considering both cognitive and wellness factors.

Generated visualizations

The method produces a figure composed of four graphs distributed in a 2x2 grid, each exploring a key relationship:

1. Sleep hours vs. exam score Through a scatter plot, this graph shows how exam scores vary based on time dedicated to sleep. A linear regression line is included to represent the general trend, as well as the Pearson correlation coefficient (r) to quantify the relationship.
2. Mental health vs. exam score This graph examines the relationship between self-assessment of mental health status (on a scale of 1 to 10) and academic performance. The inclusion of the regression line and correlation coefficient allows observation of how emotional well-being could be linked to academic success.
3. Screen time vs. sleep hours Here the impact of total screen time (sum of hours on social media and content consumption on platforms like Netflix) on sleep hours is analyzed. The goal is to demonstrate whether greater screen exposure time compromises rest, which could indirectly influence academic performance.
4. Exercise frequency vs. exam score This graph evaluates how regular physical activity (measured in times per week) relates to students' scores. It is expected that an active

lifestyle may be positively associated with better academic results, given its relationship with physical and mental health.



Method create_dashboard

This method generates an interactive and summarized visual dashboard that allows clear and accessible observation of the most relevant information from the study. Its main function is to summarize key data from the analyzed student group, combining informative graphs with descriptive statistics, all within a single visual figure.

What does it specifically do?

The method performs four main tasks, distributed across different sections of the dashboard:

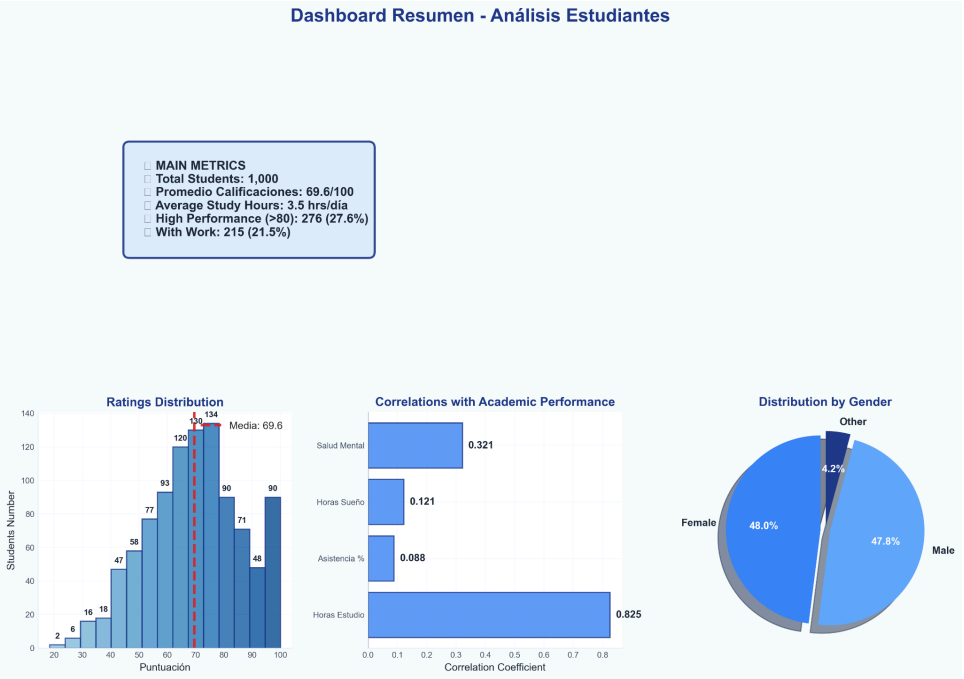
1. Main group metrics An "executive summary" type section is created with the following statistics:
    - Total students in the dataset.
    - General grade average.
    - Average study hours per day.
    - Percentage of students with high performance (grade higher than 80).
    - Percentage of students who work.
2. This information provides an initial overview of the evaluated group.

3. Grade distribution (Histogram) A histogram is shown that allows viewing how student grades are distributed. The mean is highlighted with a dotted line, and each bar is labeled with the number of students in that range. This allows detection of biases or anomalous concentrations.
4. Correlations with academic performance In a horizontal bar chart, correlations between different variables (such as study hours, attendance, sleep, and mental health) and academic performance are visualized.
    - The bars indicate the magnitude of the correlation.
    - The color shows whether the relationship is positive or negative.
5. This helps identify which factors might be most related to academic success.
6. Distribution by gender (Pie chart) The proportion of students by gender in the dataset is visualized. This segmentation allows observation of whether there is balance or predominance of any group, which could influence other analyses.

Why was it done?

The purpose of this dashboard is to facilitate global interpretation of the dataset, allowing researchers, teachers, or administrators to:

- Identify general trends quickly.
- Detect behavioral or performance patterns.
- Visualize possible factors associated with academic success.



Method clean_and_prepare_data

This function's purpose is to clean, transform, and prepare the dataset data so it can be effectively used in analysis and Machine Learning models. It represents a key stage in the process known as data preprocessing.

What does it specifically do?

1. Dataset Loading The function begins by loading the raw dataset from its source.
2. Data Cleaning (Quick Cleaning)
    - Missing value imputation:
        - For numerical variables (numerical_cols), missing values are filled with the median.
        - For categorical variables (categorical_cols), missing values are filled with the mode (most frequent value), or 'Unknown' if no mode exists.
    - Outlier adjustment:
        - Uses the Interquartile Range (IQR) method to detect extreme values.
        - Values below or above the acceptable range are adjusted to the limit using np.clip().

3. Feature Engineering (Derived Variable Creation) Three new variables are generated to enrich the analysis:
    - total_entertainment: sum of time spent on social media and Netflix.
    - study_ratio: relationship between study hours and entertainment hours.
    - wellness_index: a composite index based on sleep, exercise, and mental health.
4. Categorical Variable Encoding Categorical variables are transformed into numerical variables through Label Encoding, creating new columns with the suffix _enc.
5. Processed DataFrame Return Finally, it returns the new DataFrame ready to be used in analysis or for training Machine Learning models.

Why was it implemented?

This function was implemented to:

- Correct common problems in raw data: missing values, outliers, non-numerical variables.
- Facilitate Machine Learning algorithm training, which generally requires clean and numerical data.
- Extract additional information through new variables that help improve the model's predictive capacity.

What is its purpose within the project?

In this project, data cleaning and preparation enables:

- Having a reliable and complete database for academic performance analysis.
- Improving prediction model quality by reducing noise and enhancing student representation.
- Exploring more complex relationships between variables, such as the balance between study and leisure or overall wellness.

This preprocessing stage is fundamental for ensuring the subsequent analysis and modeling phases produce accurate and meaningful results.

Method create_ml_visualization

This function is responsible for training a supervised Machine Learning model with the objective of predicting students' exam scores (exam_score) based on variables related to habits, wellness, and socio-educational conditions.

What does it do step by step?

1. Predictor Variable Selection
    - Defines a list of independent variables (feature_cols) considered relevant for predicting academic performance. These variables include:
        - Personal data: age, gender_enc
        - Study habits: study_hours_per_day, study_ratio
        - Health and wellness: sleep_hours, exercise_frequency, mental_health_rating, wellness_index
        - Lifestyle: total_entertainment, part_time_job_enc, diet_quality_enc
        - Family context: parental_education_level_enc
    - Available variables are filtered to avoid errors if columns are missing.
2. Separation of Predictor Variables (X) and Target Variable (y)
    - X: contains the values of selected variables.
    - y: corresponds to the variable to be predicted (exam_score).
3. Variable Scaling
    - Standard scaling (StandardScaler) is applied to normalize predictor variables, improving model stability.
4. Training and Test Set Division
    - The dataset is divided into:
        - Training: 80%
        - Testing: 20%
    - This separation allows evaluating model performance with previously unseen data.
5. Model Training
    - A Random Forest Regressor model is trained, which is an ensemble of decision trees.
    - It is robust against correlated variables and allows interpretation of each variable's importance.
    - Defined parameters:

        - n_estimators=100: number of trees.
        - max_depth=10: maximum depth to prevent overfitting.
        - random_state=42: ensures reproducibility.
6. Model Evaluation
    - Predictions are made on the test set.
    - Three evaluation metrics are calculated:
        - R^2 (coefficient of determination): measures what proportion of grade variability can be explained by selected variables.
        - RMSE (root mean squared error): indicates average error on the same scale as the target variable.
        - MAE (mean absolute error): measures the average difference between predicted and actual values.
7. Feature Importance
    - The relative importance of each variable for the model is calculated.
    - The Top 5 most influential variables in academic performance prediction are printed.
8. Results Return
    - Returns a dictionary containing:
        - The trained model
        - Evaluation metrics
        - Variable importance table
        - Variables actually used

Why was it implemented?

This step is fundamental in the project because it allows:

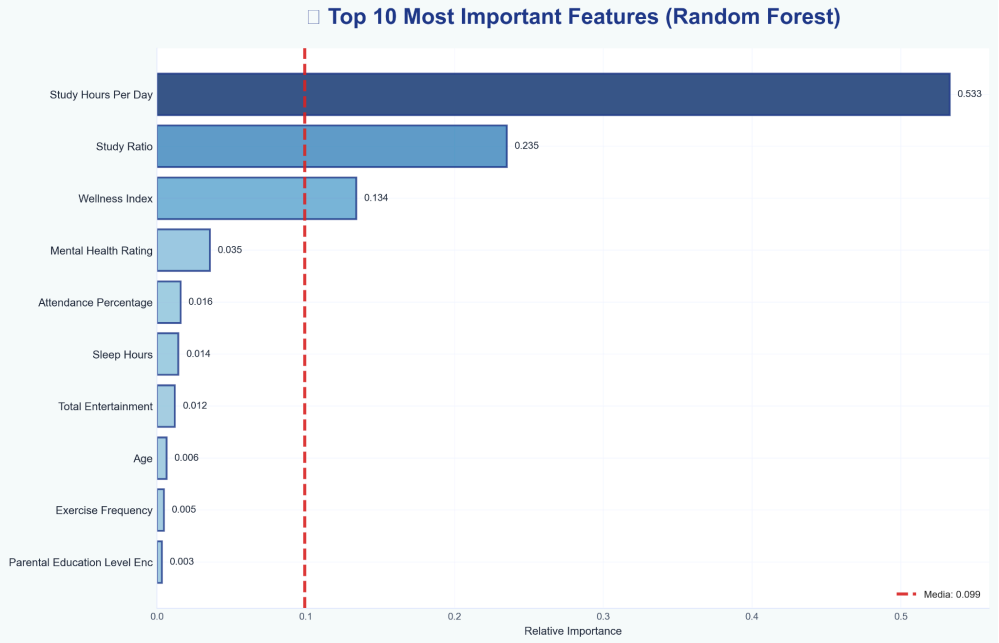
- Identifying which factors carry the most weight in academic success.
- Validating whether performance prediction is possible with the collected data.
- Evaluating model quality through quantitative metrics.
- Providing a trained model for future predictions or additional analyses.

What is its purpose within the project?

Within the comprehensive academic performance analysis, this function:

- Applies real Machine Learning techniques to an educational case.
- Provides data-based evidence about which factors are most related to good grades.
- Supports decision-making for students, institutions, or researchers by prioritizing key variables such as study habits, wellness, or family conditions.

This predictive modeling approach transforms descriptive analysis into actionable insights that can inform educational strategies and interventions.



Method generate_ml_insights

This function aims to interpret Machine Learning model results and extract actionable conclusions based on data, highlighting factors that distinguish successful students and generating practical recommendations.

What does it specifically do?

1. Analysis of Feature Importance
    - Examines which variables the model identified as most influential in predicting academic performance.
    - Ranks factors by their predictive power, providing a data-driven hierarchy of what matters most for student success.
2. Performance Pattern Identification

    - Analyzes the relationship between top predictor variables and exam scores.
    - Identifies thresholds or ranges where students typically perform better.
    - Discovers non-obvious patterns that might not be apparent through simple correlation analysis.
3. Student Success Profile Creation
    - Develops a comprehensive profile of high-performing students based on model insights.
    - Combines multiple variables to create a holistic view of successful academic behaviors and conditions.
4. Actionable Recommendation Generation
    - Translates statistical findings into practical, implementable advice.
    - Provides specific guidance for students, educators, and parents based on the most impactful factors.
    - Prioritizes recommendations based on their potential impact and feasibility.
5. Results Communication
    - Presents findings in clear, non-technical language accessible to various audiences.
    - Uses data-driven insights to support each recommendation with quantitative evidence.

Why was it implemented?

- Models alone are insufficient: it's necessary to translate technical results into comprehensible and useful messages.
- This function serves as a "bridge" between predictive analysis and practical decision-making, showing success patterns based on real data.
- Enables communication of results to non-technical audiences (such as educators, students, or parents), facilitating their understanding and impact.

What is its purpose within the project?

Within the context of academic performance analysis with Machine Learning, this function:

- Summarizes clearly and accessibly the most important conclusions from the model.
- Provides quantitative evidence to justify academic and personal recommendations.

- Supports the project's objective of generating actionable knowledge, not just predictions.
- Transforms complex statistical relationships into practical guidance that can improve student outcomes.
- Validates the practical utility of the Machine Learning approach by demonstrating real-world applicability.

This interpretation phase is crucial because it ensures that sophisticated analytical work translates into meaningful improvements in educational practices and student success strategies.

Conclusion

This project demonstrates how machine learning techniques can extract valuable and practical insights from educational data. Through rigorous data cleansing, visual analysis, and predictive modeling, clear patterns were identified between study habits, lifestyle, and academic performance.

The most significant findings include:

- A strong positive correlation between time spent studying and scores obtained on the final exam.

- School attendance is a critical factor in academic performance, showing that students with better attendance tend to achieve better grades.

- Well-being variables, such as sleep hours, mental health status, and exercise frequency, also showed significant associations with performance, demonstrating that academic success is not solely dependent on academic factors.

- The Random Forest classification model achieved an acceptable level of accuracy in predicting student performance, validating the predictive approach's usefulness in

identifying at-risk students.

This analysis not only provides quantitative results but also narratives about students' realities. These stories, supported by visualizations and data, provide a better understanding of the educational environment and suggest opportunities for early intervention and personalized support.

In conclusion, this project demonstrates that analyzing educational data with artificial intelligence is not only possible but highly useful. Machine learning tools can be key allies for educational institutions when making evidence-based decisions, with the goal of improving student performance and well-being.