Student Alcohol Consumption Analysis

Problem Description :

Problem Description: The Student Alcohol Consumption Analysis project aims to analyze a dataset containing information about students' alcohol consumption habits, demographics, and academic performance. By exploring and understanding the factors that influence alcohol consumption among students, we can gain valuable insights that can be used by educational institutions, parents, and policymakers to implement effective interventions and support systems.

Dataset Information: The dataset used in this project is a collection of survey data from students attending secondary school in Portugal. The data was collected as part of a study to understand various aspects of student life, including academic performance, family background, and habits like alcohol consumption. The dataset contains both numerical and categorical variables, providing a comprehensive view of the factors that may influence student alcohol consumption.

Background Information: Alcohol consumption among students is a significant concern as it can have adverse effects on their physical and mental health, academic performance, and overall well-being. Understanding the factors associated with alcohol consumption can help identify at-risk students and implement targeted interventions to address this issue.

The dataset includes various features such as:

- Demographics: Age, gender, family size, and parents' cohabitation status.
- Education: Mother's and father's education level.
- Travel Time: Time spent commuting to school.
- Failures: Number of past class failures.

- Alcohol Consumption: Variables indicating the frequency and quantity of alcohol consumption.

Objective: The primary objective of this project is to analyze the dataset and answer key questions related to student alcohol consumption. Some of the questions that can be explored are:

- What is the distribution of alcohol consumption among students?
- Are there any gender-based differences in alcohol consumption?
- Is there any correlation between alcohol consumption and academic performance?
- Which factors (e.g., age, family background) are associated with higher alcohol consumption?

By analyzing these questions, we aim to identify potential patterns and trends related to student alcohol consumption and provide actionable insights to address the issue effectively.

Note: The dataset used in this project should be obtained ethically and with proper consent from the students and educational institutions involved in the study. Privacy and data protection measures must be strictly followed throughout the analysis process.

Possible Framework :

1. **Importing Libraries and Data Loading**

- Import the necessary libraries, such as pandas, seaborn, and matplotlib, for data manipulation and visualization.
- Load the dataset using pd.read_csv() function and store it in a pandas DataFrame.

2. **Data Exploration and Understanding**

- Use df.columns to check the column names and ensure the dataset's completeness.
- Print the first few rows of the dataset using df.head() to understand its structure and format.
- Identify the numerical and categorical columns in the dataset for further analysis.

3. **Exploratory Data Analysis (EDA)**

- Conduct a thorough exploration of the data to gain insights and understand its distribution and relationships.
- Perform univariate analysis on categorical columns using count plots to visualize the frequency distribution of each category.
- Calculate and visualize the correlation matrix of numerical columns using a heatmap to identify potential correlations between variables.
- Group the data by gender and calculate the average number of failures using groupby() and mean() to compare the performance of male and female students.
- Use count plots to visualize the distribution of students across different schools, sexes, family sizes, and parents' cohabitation status.

4. **Descriptive Statistics**

- Compute descriptive statistics, such as mean, standard deviation, minimum, maximum, etc., for numerical columns using describe() to gain a better understanding of the dataset's central tendencies.
- Visualize the distribution of age using a histogram to observe the age group distribution among students.

5. **Data Visualization**

- Create visualizations to better understand the relationship between various features in the dataset.
- Use bar plots to compare the average mother's education level by family size and visualize the relationship between gender and age using violin plots.

6. **Data Preprocessing**

- Prepare the data for modeling by converting categorical variables into dummy variables using pd.get_dummies().
- Transform the target variable 'failures' into binary categories (0 or 1) to perform classification.

7. **Logistic Regression Model**

- Prepare the features (X) and target (y) variables for model training.
- Add a constant term to the features matrix using sm.add_constant() to account for the bias term in the logistic regression model.
- Train a logistic regression model using sm.Logit() and obtain the model summary using fit().

8. **Decision Tree Classifier Model**

- Split the data into training and testing sets using train_test_split().
- Train a Decision Tree Classifier model using DecisionTreeClassifier() and fit it with the training data.
- Use the trained model to predict the target variable for the test set and calculate the accuracy using accuracy_score().

9. **Results and Conclusion**

- Interpret the results obtained from the logistic regression model's summary to identify significant predictors of student failures.
- Evaluate the performance of the Decision Tree Classifier model based on the accuracy score.
- Summarize the findings and insights from the analysis and provide recommendations to address alcohol consumption issues among students.

10. **Final Thoughts and Further Analysis**

- Discuss potential limitations and assumptions made during the analysis.
- Suggest additional analyses that can be performed to further explore the dataset and gain deeper insights.
- Conclude the project by emphasizing the importance of data-driven decision-making in addressing student alcohol consumption issues.

Code Explanation :

*If this section is empty, the explanation is provided in the .ipynb file itself.

1. **Data Loading and Exploration**

- In the beginning, the code starts by importing the necessary libraries, such as pandas, seaborn, and matplotlib. These libraries will help us manipulate and visualize the data effectively.
- Next, the code reads the dataset from a CSV file using pd.read_csv('data.csv') and stores it in a pandas DataFrame called data.
- To get a quick overview of the dataset, the code prints the column names using df.columns and displays the first few rows of the dataset using df.head(). This helps us understand the structure and format of the data.

2. **Frequency Distribution of Categorical Columns**

- The code iterates through the categorical columns (['school', 'sex', 'famsize', 'Pstatus']) and calculates the frequency distribution of each category using the value_counts() function.
- It then prints the frequency distribution for each column, showing the number of occurrences of each category. This gives us an idea of how the data is distributed across different categories.

3. **Correlation Matrix and Heatmap**

- Next, the code calculates the correlation matrix for numerical columns (['age', 'Medu', 'Fedu', 'traveltime', 'failures']) using the corr() function.
- It visualizes the correlation matrix using a heatmap generated by sns.heatmap(). The heatmap displays the correlation coefficients between pairs of variables, with cool colors indicating negative correlations and warm colors indicating positive correlations. This helps us identify any relationships between numerical variables.

4. **Average Grades by Gender**

- The code groups the data by gender using groupby('sex') and calculates the average number of failures for male and female students using mean().
- It then prints the average grades for each gender. This allows us to compare the performance of male and female students in terms of failures.

5. **Count Plots for Categorical Columns**

- The code uses sns.countplot() to create count plots for the categorical columns, including 'school', 'sex', 'famsize', and 'Pstatus'.

- Each count plot displays the number of occurrences for each category, providing insights into the distribution of students across different categories.

6. **Descriptive Statistics for Numerical Columns**

- The code calculates descriptive statistics, such as mean, standard deviation, minimum, maximum, etc., for numerical columns using describe().
- The summary statistics are printed to provide an overview of the central tendencies and distributions of numerical features.

7. **Correlation Heatmap for Numerical Columns**

- Similar to the earlier heatmap, the code generates a correlation heatmap for numerical columns (['age', 'Medu', 'Fedu', 'traveltime', 'failures']) using sns.heatmap().
- This heatmap provides a visual representation of the correlations between numerical features, helping us identify any strong relationships between them.

8. **Histogram of Age Distribution**

- The code uses sns.histplot() to create a histogram of the age distribution. The histogram displays the number of students in different age groups, providing insights into the age distribution among the students.
- Bar Plot - Average Mother's Education Level by Family Size
- The code uses sns.barplot() to create a bar plot that shows the average mother's education level by family size.
- This visualization helps us understand how the mother's education level varies with different family sizes.

9. **Additional Insights and Data Analysis**

- The code calculates the count and percentage of students based on their gender using value_counts() and normalize=True.
- It also creates a cross-tabulation table (pd.crosstab()) to compare gender and parental cohabitation status, providing insights into the relationship between these variables.
- Furthermore, the code generates box plots and violin plots to visualize the distribution of mother's education level and age by gender, respectively.

10. **Logistic Regression Model**

- The code preprocesses the data by converting the 'failures' column into binary categories (0 or 1) using apply() and lambda.
- It creates dummy variables for categorical variables ('school', 'sex', 'famsize', 'Pstatus') using pd.get_dummies().

- The features and target variables are defined, and a logistic regression model is trained using sm.Logit() and fit().
- The result summary of the logistic regression model is printed, which provides insights into the significance of the predictors.

11. Decision Tree Classifier Model

- The code prepares the features and target variables for the Decision Tree Classifier model.
- The dataset is split into training and testing sets using train_test_split().
- The Decision Tree Classifier model is trained on the training data using DecisionTreeClassifier() and fit().
- The model's accuracy is calculated using accuracy_score() on the test set to evaluate its performance.

Future Work :

Introduction: The future work for the Student Alcohol Consumption Analysis project aims to build upon the existing analysis and enhance its capabilities. The goal is to gain deeper insights into student behaviors and factors affecting alcohol consumption. This can help educational institutions and policymakers develop effective interventions to promote student well-being and address alcohol-related issues.

Step-by-Step Guide:

1. Collect More Data:

- To improve the analysis, gather more comprehensive and diverse data related to student behaviors, socio-economic factors, family background, and academic performance. Consider surveys, questionnaires, or data from multiple schools or regions for a broader perspective.

2. Feature Engineering:

- Explore additional features that may influence alcohol consumption, such as peer influence, mental health indicators, extracurricular activities, or parental involvement. Perform feature engineering to create new variables that capture meaningful patterns.

3. Data Preprocessing and Cleaning:

- As the dataset grows, ensure proper preprocessing and data cleaning to handle missing values, outliers, and inconsistencies. This step is crucial for maintaining data quality and accurate analysis.

4. Advanced Data Visualization:

- Utilize advanced visualization techniques like interactive plots, 3D visualizations, or animated graphs to present complex relationships between variables more effectively. Interactive visualizations can help users explore the data interactively.

5. Statistical Analysis:

- Conduct in-depth statistical analysis, including regression analysis, hypothesis testing, and ANOVA, to quantify the relationships between variables and identify significant factors affecting alcohol consumption.

6. Clustering and Segmentation:

- Apply clustering algorithms like K-means or hierarchical clustering to group students based on similar characteristics. This can reveal distinct student segments with different alcohol consumption patterns and provide targeted interventions.

7. Sentiment Analysis:

- Implement sentiment analysis on textual data, such as students' responses to surveys or social media posts, to understand their emotional well-being and sentiments related to alcohol consumption.

8. Machine Learning Models:

- Utilize advanced machine learning models like random forests, support vector machines, or neural networks to predict alcohol consumption patterns more accurately. Evaluate the models' performance using metrics like accuracy, precision, recall, and F1 score.

9. Time-Series Analysis:

- If data is collected over time, perform time-series analysis to identify trends, seasonality, and cyclic patterns in alcohol consumption. This can provide insights into temporal variations and help predict future trends.

10. Data Ethics and Privacy:

- Ensure data ethics and privacy are maintained throughout the analysis. Anonymize sensitive information and follow ethical guidelines to protect students' confidentiality and privacy.

Conclusion: By following this step-by-step guide and implementing the suggested future work, the Student Alcohol Consumption Analysis project can achieve more comprehensive insights and make valuable contributions to student health and well-being. The results can be used to implement evidence-based interventions and policies that support students in leading healthy and successful academic lives.

Concept Explanation :

Alright, let me tell you about an algorithm called Decision Trees, but let's give it a fun twist and imagine it as a wizard's adventure!

Once upon a time in a magical forest, there was a wise and quirky wizard named Decisiono the Tree Wizard. Decisiono had a peculiar talent – he could help anyone make decisions by asking them a series of magical yes-or-no questions.

Imagine you're a brave adventurer wandering into Decisiono's forest, seeking his guidance on whether you should take a treacherous path or the safer one to find the hidden treasure. Exciting, right?

You start your quest by approaching Decisiono, who stands tall like an ancient tree. "Greetings, young adventurer! I am Decisiono the Tree Wizard. Are you ready for a magical journey of decisions?" he asks with a gleam in his eye.

You nod eagerly, and he conjures a magical tree right in front of you – it's the Decision Tree! Each branch of the tree represents one of Decisiono's questions. The tree has many levels, and each level has a question leading to new branches.

The First Question: "Do you have a map?" A yes or no answer determines your next move. If you have a map, Decisiono points you to the safe path. But if you don't, he asks the second question: "Have you ever traveled this way before?"

Now, let's say you don't have a map, and you answer "No" to the second question. Decisiono smiles and guides you to the third question: "Do you have a compass?" Depending on your answer, you move further down the tree, reaching more questions like "Is it daylight?" or "Are you good with stars?"

As you continue answering the questions, Decisiono keeps leading you through the branches of the Decision Tree, and eventually, you reach the final level – the "Decision Leaves". These leaves hold the treasure's location! Each leaf represents a different decision.

Finally, you reach a leaf with the decision: "Take the treacherous path." This means, after considering all your answers, Decisiono believes you should go for the adventurous route. You gasp, wondering if it's the right choice, but you trust the wisdom of the Decision Tree.

And so, your journey begins! Following the guidance of the Decision Tree, you embark on the treacherous path, overcoming obstacles, solving riddles, and eventually finding the hidden treasure.

In the real world, Decision Trees work similarly! They use a series of questions and logical splits to help us make decisions based on data. Each question (or "split") divides the data into smaller groups, leading to a specific decision or outcome. The algorithm recursively creates this tree until it reaches leaves, which represent the final decisions.

So, next time you need some magical decision-making assistance, just remember the quirky wizard Decisiono and his amazing Decision Tree adventure! ⍰♂⍰⍰

Exercise Questions :

1. **What are the categorical and numerical columns in the dataset?**

Answer: The categorical columns are ['school', 'sex', 'famsize', 'Pstatus'], and the numerical columns are ['age', 'Medu', 'Fedu', 'traveltime', 'failures'].

2. **How can you visualize the correlation between numerical columns using a heatmap?**

Answer: You can use the seaborn library's heatmap function to create a correlation matrix and then visualize it as a heatmap.

3. **What does the correlation between 'Medu' and 'Fedu' tell us?**

Answer: The correlation between 'Medu' (Mother's Education) and 'Fedu' (Father's Education) indicates the strength and direction of their relationship. A positive correlation suggests that as one variable increases, the other tends to increase as well, and vice versa.

4. **What is the average number of failures for each gender?**

Answer: You can use the groupby function to group the data by 'sex' and calculate the mean of 'failures' for each gender.

5. **How can you visualize the count of schools in the dataset?**

Answer: You can use the seaborn library's countplot function to create a bar chart showing the count of each school in the dataset.

6. **What is the distribution of ages among the students?**

Answer: You can use the seaborn library's histplot function to create a histogram showing the distribution of ages among the students.

7. **Is there any relationship between mother's education level ('Medu') and family size ('famsize')?**

Answer: You can use the seaborn library's barplot function to create a bar chart showing the average mother's education level for each family size.

8. **What percentage of students are male and female in the dataset?**

Answer: You can calculate the count and percentage of male and female students using the value_counts and normalize functions.

9. **Is there any association between the variables 'sex' and 'Pstatus'?**

Answer: You can create a cross-tabulation between 'sex' and 'Pstatus' using the crosstab function to see if there is any association between the two variables.

10. **How can you build a logistic regression model to predict student failures based on 'age', 'Medu', 'Fedu', and 'traveltime'?**

Answer: You can use the statsmodels library to build a logistic regression model. First, create the design matrix X with the selected features and add a constant term using add_constant. Then, define the target variable y as 'failures'. Finally, use the Logit function to create the logistic regression model and fit it to the data. You can then use the model to make predictions.