EDA quiz

8 out of 8 correct

1. Which feature(s) in the Wine Quality dataset exhibit non-normality?

- ○ Fixed acidity
- ◉ Residual sugar
- ○ pH
- ○ Alcohol

Explanation: The residual sugar feature in the Wine Quality dataset exhibits non-normality, as confirmed by the EDA analysis. The histogram and Q-Q plot of this feature show a non-normal distribution with a positive skew. Other features, such as fixed acidity, pH, and alcohol, exhibit normal or nearly normal distributions.

2. Which transformation(s) can be applied to the non-normal features in the Wine Quality dataset to improve normality?

- ○ Logarithmic transformation
- ○ Square root transformation
- ○ Exponential transformation
- ◉ All of the above

Explanation: To improve normality in non-normal features, various transformation methods can be applied, including logarithmic transformation, square root transformation, and exponential transformation. All of these transformations can help to normalize skewed data by reducing the influence of outliers and improving the distribution's symmetry. Therefore, option D is the correct answer.

3. What is the most important feature for predicting wine quality according to the feature importance ranking?

- ○ pH
- ◉ Alcohol
- ○ Density
- ○ Sulphates

Explanation: According to the feature importance ranking analysis, the alcohol feature is the most important feature for predicting wine quality in the Wine Quality dataset. This indicates that wines with higher alcohol content tend to have higher quality ratings. Other features, such as volatile acidity, sulphates, and citric acid, also play a role in determining wine quality but are less important than alcohol. Therefore, option B is the correct answer.

4. What feature(s) in the Wine Quality dataset have the highest correlation with wine quality?

- ○ Chlorides
- ○ Total sulfur dioxide
- ○ Citric acid
- ◉ All of the above

Explanation: According to the correlation analysis of the Wine Quality dataset, all three features - Chlorides, Total sulfur dioxide, and Citric acid - have the highest correlation with wine quality. This indicates that these features have a strong impact on the overall quality rating of wines. Other features, such as pH, density, and residual sugar, also have a moderate correlation with wine quality. Therefore, option D is the correct answer.

5. What is the purpose of feature engineering in machine learning?

- ○ To increase the size of the dataset
- ◉ To extract useful information from the raw data
- ○ To make the dataset more complex
- ○ None of the above

Explanation: The purpose of feature engineering in machine learning is to extract useful information from the raw data and transform it into a set of features that can be used to train a predictive model. Feature engineering involves selecting and extracting relevant features from the raw data, transforming and scaling the features to a common range, and creating new features by combining existing ones. The goal is to maximize the information content of the features while minimizing the number of irrelevant or redundant features. Therefore, option B is the correct answer.

6. Which feature selection technique uses tree-based models to rank the importance of each feature in predicting the target variable?

- ◯ Correlation analysis
- ◯ Principal component analysis
- ◯ Model-based selection
- ◉ Feature importance ranking

Explanation: Feature importance ranking is a technique that uses tree-based models to rank the importance of each feature in predicting the target variable. This technique involves fitting a decision tree or a random forest to the data and using the resulting model to measure the impact of each feature on the prediction accuracy. Features that have a high impact on the prediction accuracy are considered important and retained, while features that have a low impact are considered irrelevant and removed. Therefore, option D is the correct answer.

7. What is the main goal of analyzing students' performance in exams?

- ◯ To evaluate the quality of teaching
- ◯ To predict future performance
- ◯ To identify areas for improvement
- ◉ All of the above

Explanation: Analyzing students' performance in exams has several goals, including evaluating the quality of teaching, predicting future performance, and identifying areas for improvement. By analyzing exam results, educators can evaluate the effectiveness of their teaching methods and identify areas where students may be struggling. They can also use the data to predict future performance and identify students who may need additional support or

intervention. Additionally, analyzing exam results can help to identify patterns and trends in student performance, which can inform decisions about curriculum development and instructional design. Therefore, option D is the correct answer.

8. Which feature in the Students Performance dataset has the strongest correlation with math scores?

- ◯ Parental education
- ◯ Test preparation course
- ◉ Lunch
- ◯ None of the above

Explanation: In the Students Performance dataset, the "lunch" feature has the strongest correlation with math scores. This can be confirmed by calculating the correlation coefficient between the "lunch" feature and the "math score" feature using a correlation matrix. The "lunch" feature is a proxy for socio-economic status, and students who receive free or reduced-price lunches are typically from lower-income families. Therefore, the strong correlation between the "lunch" feature and math scores suggests that socio-economic status is an important factor in predicting math performance. Therefore, option C is the correct answer.

Submit