Question Code: 141

Data Fellowship iykra

Fathurrahman Hernanda Khasan

1. Exploratory Data Analysis (EDA) :
   a. Load the dataset using pandas.
   b. Display the first few rows of the dataset to understand its structure.
   c. Generate summary statistics to understand the distribution of numerical variables and identify any immediate discrepancies.
   d. Create visualizations such as histograms, box plots, and scatter plots to explore the relationships between different features and the target variable (self reported health status).

2. Data Processing and Cleaning :
   a. For handling missing values, i used the mean strategy. Because the most are normally distributed. Because of that me replace missing values with the mean of each column, and after eda there is no missing data, id do not handling missing value.
   b. i identified outliers using the Z-score method. Rows with any numerical value exceeding 3 standard deviations from the mean are considered outliers and removed.
   c. i used one-hot encoding to convert the categorical variable 'gender' and ‘self_reported_health_status’ into a numerical form suitable for machine learning models. This is because gender has no ordinal relationship between categories, so one-hot encoding is appropriate and self_reported_health_status is a label.

3. Feature Engineering :
   a. I created a new categorical feature bmi_category based on BMI range to capture different weight statuses.
   b. I created a combined feature activity_to_fruit_veggie_ratio to capture the relationship between physical activity and dietary habits.
   c. I label-encoded the bmi_category feature for modeling.
   d. Fo feature selection, i used ANOVA F-value to select the top k features. In this example, we selected the top 5 features.
   e. For dimensionality reduction, i applied PCA to reduce the dimensionality of the feature space to 5 principal components.

Link Answer : https://github.com/fahernkhan/141---Technical-Test-Essay---Data-Fellowship-Batch-12