simge, sembol, yazı tipi, logo, grafik içeren bir resim

Açıklama otomatik olarak oluşturuldu

**CSE4288-Introduction to Machine Learning**

**Term Project**

**Data Preprocessing Report**

**Group17**

**150120024-Enes Haksoy Öztürk**

**150123842-Enes Akturan**

**150121025-Ahmet Abdullah Gültekin**

**150120005-Ayşe Gülsüm Eren**

**150119762-Muhammet Ali Özkul**

**Gender-Based Wage Disparity Analysis Across Sectors**

**Dataset Structure and Content Summary**

The dataset appears to capture demographic, educational, and professional attributes alongside salary information. This information is likely intended for analyzing salary trends based on different features.

The dataset includes the following attributes:

* Age (Numeric)
* Gender (Categorical)
* Education Level (Categorical)
* Job Title (Categorical)
* Years of Experience (Numeric)
* Salary (Numeric)

**Data Cleaning**

We performed a series of data cleaning operations on the dataset.

* We identified columns with missing values and counted how many were missing in each column. Rows containing any missing values were removed, ensuring a clean dataset for further analysis.
* We removed rows with identical values across all columns, leaving only unique rows in the dataset.
* Numerical columns were converted to integer type to improve compatibility with certain analysis methods and ensure data uniformity.
* We cleaned the categorical data by removing unnecessary spaces that could have caused inconsistencies (e.g., "Male " vs. "Male"). This ensured the categories were clean and uniform, reducing the risk of mismatched or duplicate entries.
* Categorical columns were converted lowercase to improve compatibility with certain analysis methods and ensure data uniformity.
* Categorical columns (Gender, Education Level, and Job Title) were encoded using LabelEncoder to convert categorical values into numeric formats suitable for correlation analysis.

**External Data Analysis with Visualizations**

**Correlation Matrix:** A heatmap was plotted to examine correlations between numerical variables.

* Strong positive or negative correlations indicate potential relationships worth further exploration.
* Example: Years of Experience and Salary showed a moderate positive correlation.

**Education Level Distribution:** The distribution of education levels was visualized using a pie chart:

* + The most common education levels were highlighted.
  + This distribution provides insights into the qualifications of the workforce.

**Gender and Education Level:** A bar chart was created to compare the number of workers based on gender and education level:

* + Provides insights into gender representation across different education levels.

**Outliers Analysis:** Histograms and box plots were used to visualize the distributions of Salary, Age, and Years of Experience.

* + These visualizations helped identify potential outliers, such as extreme values in Salary.

**Salary Distribution vs Gender:** A box plot was created to compare salary distributions across genders:

* + Reveals potential gender-based disparities in salaries.

**Overall Dataset Distribution:** A histogram was plotted for all numerical columns to analyze their distributions:

* + Provides a high-level overview of the variability in the dataset.

The cleaned and preprocessed data was saved. This dataset is now ready for advanced analysis or machine learning applications.

**Rationale for Feature Selection or Engineering**

1. **Categorical to Numerical Conversion:**
   * Categorical features (Gender, Education Level, Job Title) are converted to numerical values using LabelEncoder to facilitate correlation analysis and machine learning model training.
2. **Handling Missing Values:**
   * Rows with missing values are dropped to ensure the dataset is clean and complete for analysis.
3. **Correlation Analysis:**
   * The correlation matrix helps in understanding the relationships between numerical features, which can inform feature selection for modeling.
4. **Visualization for Insights:**
   * Various visualizations (heatmap, pie chart, bar chart, histograms) are used to gain insights into the data distribution and relationships between features.

These steps ensure the dataset is clean, well-prepared, and provides valuable insights through exploratory data analysis (EDA).