**Programming For AI**

**(AI2001)**

**Fall-2024**



**Final Project Report**

**Submitted to : Miss Mehak Mazhar**

**Team Members:**

**Raghib Rizwan (23K-0012) and Hafsa Rashid (23K-0064)**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**NATIOAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES (NUCES), KARACHI**

**Data Analysis and Visualization of the WineQT Dataset with a User Interface**

**Objective:** The aim of this project was to analyze the WineQT dataset using data visualization techniques and machine learning (K-Means clustering) to extract meaningful insights. Additionally, a user-friendly interface was developed using HTML and CSS to showcase the analysis results interactively.

**Dataset Description:**

* Dataset Name: WineQT
* Number of Rows: 1,143
* Number of Columns: 13

The dataset comprises physicochemical attributes such as pH, alcohol content, and quality scores, which are critical for understanding the characteristics of wine.

**Tools and Techniques Used:**

1. **Python Libraries for Analysis:**
   * **Seaborn**: For correlation heatmap, distribution plots, and box plots.
   * **Pandas**: For numerical computations during analysis.
   * **HTML and CSS**: For designing and implementing the UI app to showcase visualizations.

**Analysis Workflow:**

**1. Correlation Heatmap:**

* **Purpose:**  
  To identify relationships and dependencies between features.
* **Insights Gained:**
  + Strong correlations were observed between certain features, such as alcohol and quality, suggesting that higher alcohol content might improve wine quality.
  + Weak correlations between other features provided insights into which variables had minimal impact on quality.
* **Applications:**  
  Correlation analysis is essential for feature selection in predictive modeling.

**2. Distribution Plot:**

* **Purpose:**  
  To visualize the distribution of numerical features to understand data spread and detect skewness or outliers.
* **Insights Gained:**
  + Features like alcohol and pH displayed normal distributions, while others showed skewed patterns.
  + Highlighted overlapping ranges between features, aiding in understanding their variability.

**3. Box Plot and Bar Graph:**

* **Purpose:**
  + Box plots: To detect outliers and compare medians across features.
  + Bar graphs: To provide an aggregated view of feature values.
* **Insights Gained:**
  + Identified extreme values (outliers) for key features such as residual sugar and citric acid.
  + Highlighted the variability and central tendencies of features, enabling better comparisons.

**4. Scatter Plot:**

* **Purpose:**  
  To explore relationships between specific features (e.g., pH vs. alcohol) and detect linear or non-linear patterns.
* **Insights Gained:**
  + Scatter plots revealed clustering patterns, indicative of wine categories.
  + Highlighted potential feature pairings for predictive models.

**5. K-Means Clustering (with PCA):**

* **Purpose:**  
  To group data into clusters based on similarities, with PCA reducing dimensionality for visualization.
* **Insights Gained:**
  + Clustering revealed natural groupings within the dataset.
  + Provided a simplified view of high-dimensional data, aiding segmentation analysis.

**Purpose of Data Visualizations:**

The visualizations were designed to simplify complex datasets, uncover trends, and provide actionable insights. Each graph served a specific role in enhancing understanding:

* **Correlation Heatmap:** Clarified feature dependencies and selection priorities.
* **Distribution Plot:** Revealed data distributions and potential preprocessing requirements.
* **Box Plot and Bar Graph:** Highlighted variability and outliers.
* **Scatter Plot:** Illustrated relationships and clustering tendencies.
* **K-Means Clustering:** Grouped similar data points to identify patterns.

**How Visualization Enhances Understanding:**

1. **Simplification of Complexity:**Visualizations break down complex numerical datasets into comprehensible insights, avoiding the need for manual analysis of raw numbers.
2. **Trend Identification:**Graphs like scatter plots and correlation matrices highlight trends such as the influence of alcohol on quality.
3. **Outlier Detection:**Box plots effectively expose extreme values that could skew analyses or models**.**
4. **Feature Relationships:**Scatter plots and heatmaps provide clarity on how features interact, aiding in feature engineering.
5. **Segmentation and Grouping:**K-Means clustering simplifies high-dimensional data, revealing hidden patterns and segmentations.

**User Interface Implementation:**

A web-based application was developed to present the analysis results interactively.

* **Features of the UI:**
  + Graphs and visualizations embedded within the app for intuitive exploration.
  + A clean, responsive design implemented using HTML and CSS to ensure accessibility across devices.
* **Purpose of the UI:**  
  To make data insights accessible to non-technical users, bridging the gap between analysis and decision-making.

### ****Conclusion:****

This project combined data analysis, visualization, and front-end development to analyze the WineQT dataset effectively. By leveraging Python libraries for insights and web technologies for presentation, the project demonstrated the power of data visualization in understanding and communicating complex datasets.

**Acknowledgments:**

Special thanks to the course instructor for their guidance and support throughout this project.