**Name:** Arya Yemul

**Roll No:** 281004

**Batch:** A1

**Assignment 2**

**Problem Statement:**

Perform the following operations using R/Python on the data sets:

a) Compute and display summary statistics for each feature available in the dataset. (e.g. minimum value, maximum value, mean, range, standard deviation, variance and percentiles

b) Illustrate the feature distributions using histogram.

c) Data cleaning, Data integration, Data transformation, Data model building (e.g. Classification)

**Objectives:**

1. To explore and understand the dataset through summary statistics such as mean, standard deviation, range, and percentiles.
2. To visualize feature distributions using histograms for identifying patterns and potential anomalies.
3. To perform essential data preprocessing steps including cleaning, integration, and transformation.
4. To build and evaluate a classification model based on the transformed dataset for predictive analysis.

**Resources used:**

1) Software used: Jupiter Notebook

2) Libraries used: Pandas, Matplotlib, Seaborn, SKLearn

**Theory:**

1. **Mean (Average)**

* The sum of all values divided by the number of values.
* Indicates the central tendency of the data.

1. **Range**

* The difference between the maximum and minimum values.
* Shows the spread of the data**.**

1. **Standard Deviation (SD)**

* Measures the average distance of each data point from the mean.
* Indicates how spread out the data is.

1. **Variance**

* The square of the standard deviation.
* Shows the degree of variation in the dataset.

1. **Percentiles**

* Values that divide the data into 100 equal parts.

**Methodology:**

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

* Generate summary stats: min, max, mean, std, variance, percentiles.
* Understand data distribution, outliers, and feature scales.

**2. Data Visualization**

* Plot histograms for numerical features.
* Detect skewness, outliers, and distribution patterns.

**3. Data Preprocessing Pipeline**

* **Cleaning**: Handle missing values, remove duplicates, fix data types.
* **Integration**: Merge multiple datasets if needed.
* **Transformation**:
  + Encode categorical variables.
  + Scale numerical features (standardize/normalize).
  + Create or remove features based on relevance.

**4. Model Development**

* Split data into training and testing sets.
* Choose a classification algorithm (e.g., Logistic Regression, Random Forest, SVM).
* Train the model on the training data.

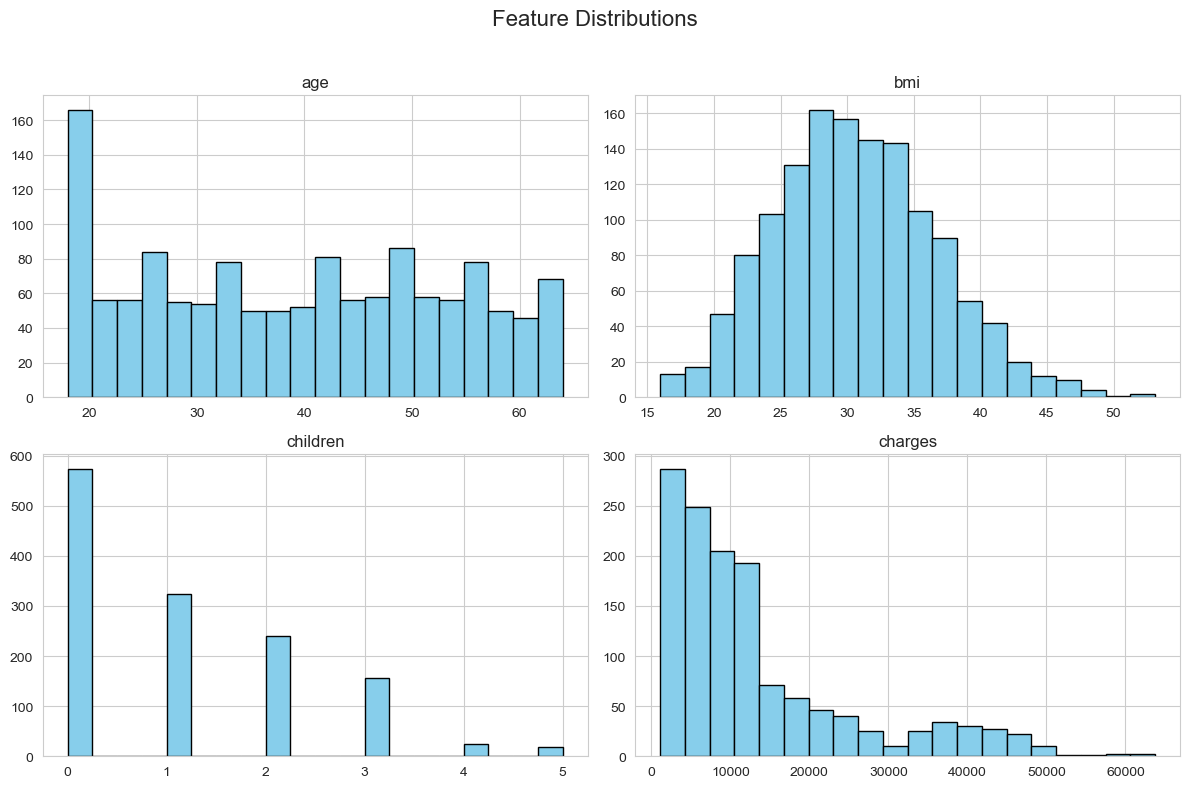
**5. Model Evaluation**

* Evaluate using Accuracy, Precision, Recall, F1-score, ROC-AUC.
* Apply cross-validation and hyperparameter tuning (e.g., Grid Search).
* Handle class imbalance if needed.

**Results:**

|  |  |  |  |
| --- | --- | --- | --- |
| age | bmi | children | charges |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **count** | 1338.000000 | 1338.000000 | 1338.000000 | 1338.000000 |
| **mean** | 39.207025 | 30.663397 | 1.094918 | 13270.422265 |
| **std** | 14.049960 | 6.098187 | 1.205493 | 12110.011237 |
| **min** | 18.000000 | 15.960000 | 0.000000 | 1121.873900 |
| **25%** | 27.000000 | 26.296250 | 0.000000 | 4740.287150 |
| **50%** | 39.000000 | 30.400000 | 1.000000 | 9382.033000 |
| **75%** | 51.000000 | 34.693750 | 2.000000 | 16639.912515 |
| **max** | 64.000000 | 53.130000 | 5.000000 | 63770.428010 |



Confusion Matrix:

[[205 11]

[ 1 49]]

Classification Report:

precision recall f1-score support

0 1.00 0.95 0.97 216

1 0.82 0.98 0.89 50

accuracy 0.95 266

macro avg 0.91 0.96 0.93 266

weighted avg 0.96 0.95 0.96 266

**Conclusion:**

We generated summary statistics, visualized feature distributions using histograms, and built a complete data pipeline—covering cleaning, integration, transformation, and classification modeling. The process highlighted the importance of exploratory analysis and data preprocessing in developing accurate and reliable predictive models.