# Data Operations and Machine Learning Practical 2

## 1) 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.  
b) Data Visualization - Create a histogram for each feature in the dataset.  
c) Data cleaning, Data integration, Data transformation, Data model building (e.g., Classification).

## 2) Libraries Used

Python:  
Pandas: For data manipulation and preprocessing.  
NumPy: For numerical operations and array manipulation.  
Matplotlib: For data visualization.  
Seaborn: For enhanced data visualization.  
Scikit-learn: For model building and evaluation.

## 3) Theory

The theoretical basis of the practical involves understanding statistical distributions, data quality improvement techniques, and the application of machine learning models to predict outcomes based on cleaned and transformed data.

A diagram of data processing

Description automatically generated

## 4) Methods

Methods include detailed steps for data preprocessing, visualization using histograms, data cleaning techniques, integration of multiple data sources, transformation of features, and the building and evaluation of a classification model.

## 5) Advantages

Advantages of the chosen methods include:

* • Enhanced understanding of data through visualization.
* • Improved data quality through rigorous cleaning processes.
* • Accurate predictions from well-prepared data feeding into sophisticated models.

## 6) Disadvantages

Disadvantages include:

* • Time-consuming data cleaning and preparation processes.
* • Complexity in integrating and transforming data from multiple sources.
* • Risk of overfitting in machine learning models if not properly tuned.

## 7) Working

The working section describes the practical application of theoretical knowledge and methods to real dataset problems, focusing on achieving actionable insights through machine learning.

## 8) Conclusion

The practical demonstrates the effectiveness of systematic data analysis and machine learning in making informed decisions based on data. The importance of comprehensive data preparation is underscored by the success of the applied models in achieving high accuracy.