COMSATS UNIVERSITY ISLAMABAD LAHORE CAMPUS

COMPUTER SCIENCE DEPARTMENT

ASSIGNMENT: #2

MACHINE LEARNING

SUBMITTED BY:

MAHAM NASIR (SP24-RCS-020)

SUBMITTED TO:

DR. MUHAMMAD SHARJEEL

DATE OF SUBMISSION:

6TH OCTOBER 2024

3. Write a paragraph about your experience of working with the standard ML pipeline in your own words.

Working with a standard machine learning (ML) pipeline is a systematic and structured process. It starts with data collection and preprocessing, where data is cleaned, transformed, and split into training and testing sets. Feature engineering often follows, involving the selection or creation of relevant features to improve model performance. Afterward, various models are trained and tuned using techniques like cross-validation to find the best fit for the data. Evaluation metrics, such as accuracy, are then used to assess model performance. Finally, the model is deployed, and its performance is monitored and updated over time to ensure continued accuracy. Each stage is critical, and working through the entire pipeline helps in building strong and scalable solutions.

1. Data Collection and Preprocessing

The dataset used comprised 80 instances with 7 features, collected from [Google Drive shared folder]. Preprocessing steps included handling missing values, removing duplicate entries, and normalizing numerical features. The dataset was split into training (50%) and testing (50%) sets to facilitate model evaluation.

2. Feature Engineering

Key features were selected based on a combination of domain knowledge and correlation analysis. After scaling all numerical features using standardization, 3 features were retained for the model training process. This step helped to enhance model performance by reducing irrelevant information.

3. Model Training and Tuning

Three machine learning algorithms were chosen for comparison:

- **J48**
- **Random Forest**
- **REPTree**

4. Evaluation Metrics

The models were evaluated on the test set using several metrics, including accuracy, Correctly Classified Instances and Incorrectly Classified Instances. Below is a comparison of the models:

Model	Accuracy	Correctly Classified Instances	Incorrectly Classified Instances
J48	85%	34	6
Random Forest	**90%**	36	4
REPTree	87.5 %	35	5

The Random Forest model outperformed the others across most evaluation metrics, achieving the highest accuracy. The Random Forest model was the best-performing model in terms of accuracy (90%), making it the ideal candidate for deployment.