COMSATS University Islamabad, Lahore Campus

Roll No: FA23-RCS-011

Name: Zain Ul Abdeen

Assignment 2 Report: J48 Classification Results

Course Title: Machine Learning

Course Code: CSC668 - PCS 716

Resource Person: Dr. Muhammad Sharjeel

Marks: 10

Due Date: October 6, 2024

Introduction:

In this assignment, I explored the complete machine learning (ML) pipeline by applying the J48 classification algorithm to a dataset containing names. The objective was to understand how to extract features, prepare data, and evaluate the performance of a classifier using WEKA.

Feature Extraction:

To begin, I manually observed the dataset and extracted several input features from the names of individuals. The features I identified included:

- **Names:** The actual names of individuals.
- **Gender:** A categorical feature indicating the gender associated with each name.
- **Even or odd:** A feature indicating whether the length of the name is even or odd.
- **Starting From Vowel:** A binary feature indicating if the name starts with a vowel.
- **Signs:** A feature representing the output class, which I converted to numeric values (1 for positive and 0 for negative) and named the column as "**Output in 0 and 1**".

I compiled these features into an ARFF file, which served as the input for the J48 algorithm.

J48 Classification Results

After loading the ARFF file into WEKA, I ran the J48 classification algorithm with the following parameters:

- **Confidence Factor (C):** 0.25
- **Minimum Number of Instances per Leaf (M):** 2
- **Test Mode:** 10-fold cross-validation

Summary of Results

The results of the classification were quite promising:

- **Total Instances:** 100
- **Correctly Classified Instances:** 100 (100%)
- **Incorrectly Classified Instances:** 0 (0%)
- **Kappa Statistic:** 1
- **Mean Absolute Error:** 0
- **Root Mean Squared Error:** 0
- **Relative Absolute Error:** 0%
- **Root Relative Squared Error:** 0%

Detailed Accuracy by Class:

The detailed accuracy metrics for each class were as follows:

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
0	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
1	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000

The confusion matrix indicated that all instances were classified correctly, with no false positives or negatives.

Model Complexity:

The J48 model produced a pruned tree with:

- **Number of Leaves:** 2
- **Size of the Tree:** 3

The time taken to build the model was negligible, indicating efficient processing.

Experience with the ML Pipeline:

Working through the standard ML pipeline was an enlightening experience. I learned the importance of feature selection and how it directly impacts the performance of the classifier. The J48 algorithm's ability to classify the dataset with perfect accuracy was surprising and highlighted the effectiveness of decision trees in handling categorical data. Additionally, the process of using WEKA for model training and evaluation was user-friendly and intuitive, making it easier to visualize the results and understand the underlying mechanics of the algorithm.

Overall, this assignment not only reinforced my understanding of machine learning concepts but also provided practical experience in applying these concepts to real-world data.