Going through the ML pipeline highlighted the importance of creating good features. Extracting features from the text improved my understanding of the data and demonstrated how small details can impact model performance. Using WEKA made it simple to experiment with algorithms and visualize results. Observing the J48 algorithm's behavior reinforced the critical role of feature selection in machine learning.

**Results:**

**Model Type**: Pruned J48 Decision Tree

**Structure**:

* Based on conditions involving starting characters, name length, and number of vowels.
* Tree has **5 leaves** and a size of **9**.

**Performance Summary**

**Correctly Classified Instances**: 85%

**Incorrectly Classified Instances**: 15%

**Kappa Statistic**: 0.7

**Mean Absolute Error**: 0.1937

**Root Mean Squared Error**: 0.3504

**Total Instances**: 100

**Summary of Run Information**

**Classifier Used**: J48 (C = 0.25, M = 2)

**Dataset**: namesWithFeatures

**Instances**: 100

**Attributes**: 10 features including Name, LengthOfName, and NumberOfVowelsInName.

**Test Mode**: 10-fold cross-validation

**Classifier Model**

**Model Type**: Pruned J48 Decision Tree

**Structure**:

1. Based on conditions involving starting characters, name length, and number of vowels.
2. Tree has **5 leaves** and a size of **9**.

**Performance Summary**

1. **Correctly Classified Instances**: 85%
2. **Incorrectly Classified Instances**: 15%
3. **Kappa Statistic**: 0.7
4. **Mean Absolute Error**: 0.1937
5. **Root Mean Squared Error**: 0.3504
6. **Total Instances**: 100

**Detailed Accuracy by Class**

**Class TRUE**:

1. TP Rate: 88%
2. FP Rate: 18%
3. Precision: 83%
4. F-Measure: 85%

**Class FALSE**:

1. TP Rate: 82%
2. FP Rate: 12%
3. Precision: 87%
4. F-Measure: 84%

**Confusion Matrix**

1. **True Positives (TRUE)**: 44
2. **False Positives (TRUE)**: 6
3. **False Negatives (FALSE)**: 9
4. **True Negatives (FALSE)**: 41

The J48 decision tree classifier was tested on a dataset called "namesWithFeatures," which has 100 examples and 10 characteristics related to names. Using a method called 10-fold cross-validation, the model achieved an accuracy of 85%, meaning it correctly classified 85 out of 100 examples but got 15 wrong.

The model was particularly good at identifying the TRUE class, achieving:

88% true positive rate (it correctly identified 88% of the TRUE cases)

83% precision rate (when it predicted TRUE, it was correct 83% of the time)

**Screenshots for each attribute:**





















