6/10/2024

Assignment 2

FA24-RCS-013

Submitted by: Umar Farooq (FA24-RCS-013)

Submitted to: Sir Dr. Muhammad Sharjeel

Subject: Machine Learning

Submission date: 6 October, 2024

Hand crafted features:

I have completed the task of manually examining the text to extract features (names of people) and created the necessary arff files.

First, I made three separate arff files, each focusing on different features of the names.

First arff file:

In this file, I took two specific attributes:

- **Names that start with a single vowel** (a, e, I, o, u) not two or more vowel at start.
- **The second letter of the name is "h"**.

These two features were used to categorize the names in this file.

Second arff file:

In the second file, I focused on two other features:

- **Names that start with any vowel**.
- **The length of the name** (less the 5 character the +)

I used these two attributes to create this version of the arff file.

Third arff file:

For the last arff file, I combined some attributes from the previous files:

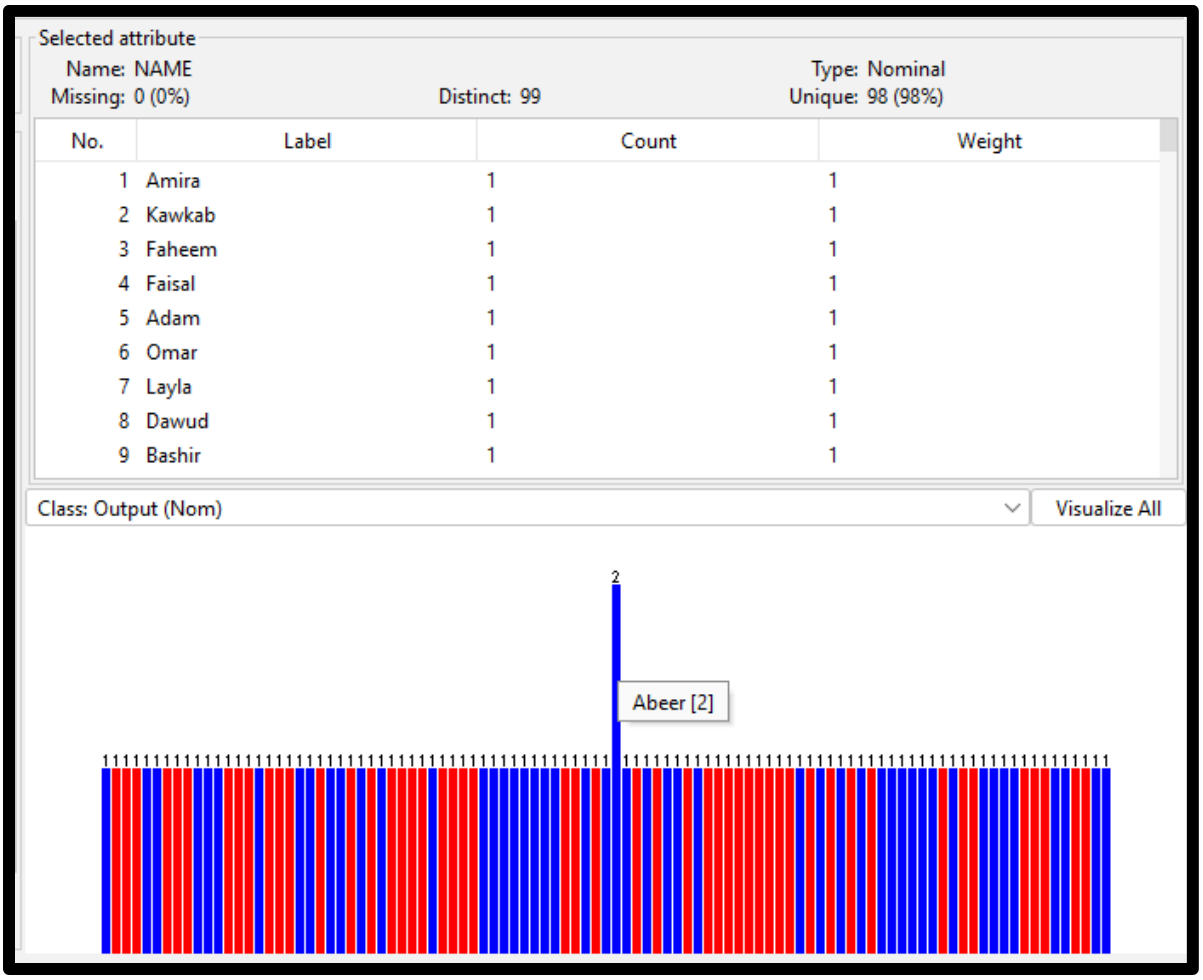
- **Names that start with a vowel**.
- **The second letter of the name is "h"** (same as the first file).

This file also uses two features but in a different combination compared to the other two.

Each of these files had different ways of organizing the data based on these selected features, and I saved them separately to test in weka later.

Characteristics of the data (weka's main window)

After I uploaded the arff files into weka's main window to check the data, I noticed something strange. I had added 100 names in the file, but weka was only showing 99 names. This confused me at first because I thought I had everything right. As I looked closer at the dataset in weka, I found the reason for the missing name: one of the names was listed twice in the arff file. Because of this duplication, weka counted only 99 unique names instead of 100. The repeated name (Abeer) had a taller line in weka's display as shown in image below, which made it easy to spot that this name was included twice. This feature helped me quickly see the duplication without having to search through the file manually.

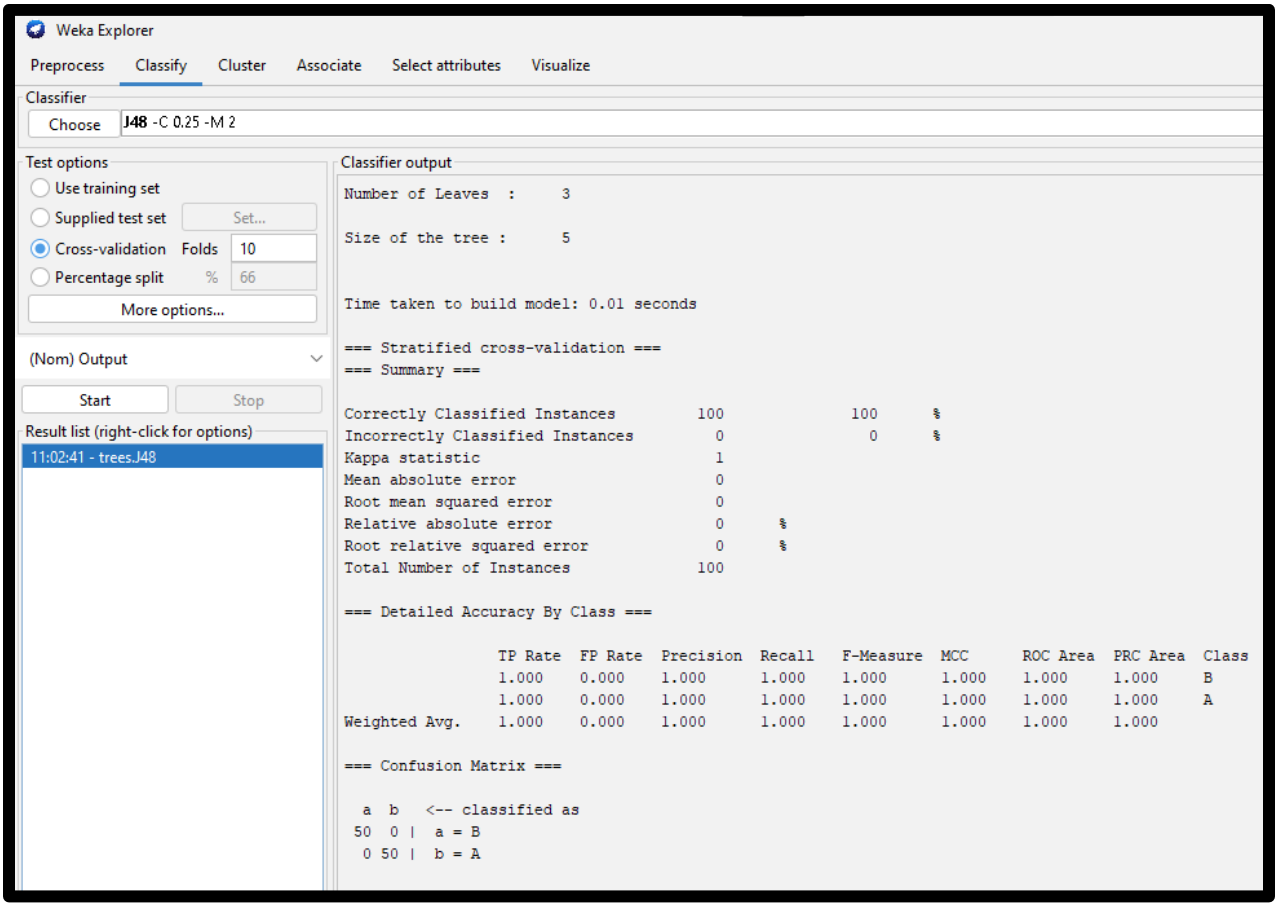


Run the j48 classification algorithm on 1st arrf file

After running the first arff file in weka, I was really happy to see that I got a perfect result: **100% accuracy**. This means that every name in my dataset was classified correctly based on the features I chose. The two attributes I used were names that start with a single vowel (like a, e, I, o, u) and the second letter being "h."

Thanks to these magic features, all the names got the right tags, and there were no mistakes 0% incorrect classifications. This means every name matched exactly what I was looking for, and that felt great!

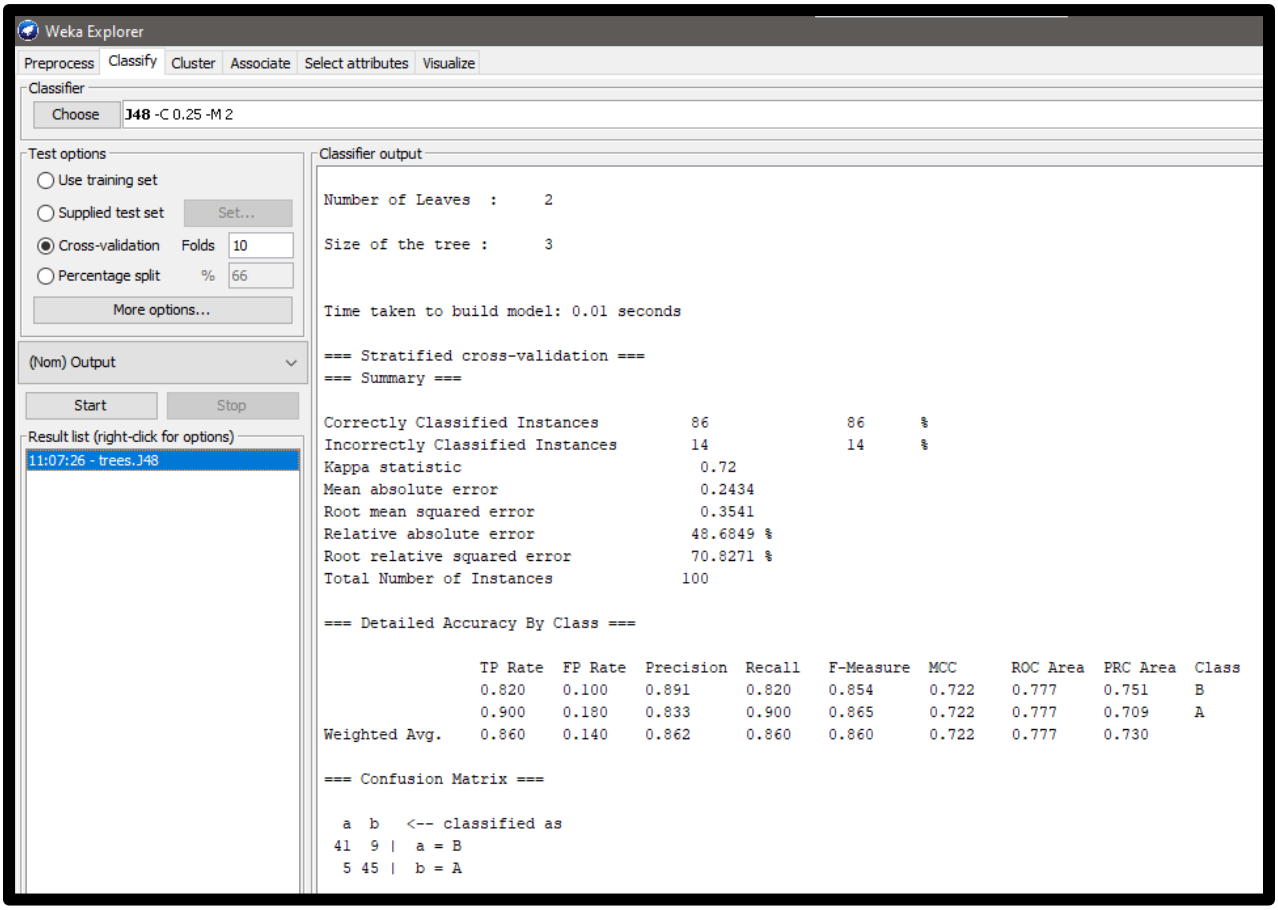
This success showed me how important it is to pick the right features when doing a classification task. By choosing these particular attributes, I was able to get clear and accurate results. It really helped me understand that when the features are relevant, the classifier performs much better.



Run the j48 classification algorithm on 2nd arrf file

When I ran the second arff file in weka, I got a good result of 86% accuracy. This means that 86% of the names in my dataset were classified correctly based on the features I selected. The attributes I used for this file were names that start with a vowel and names with a length of less than 5 characters. While I didn't achieve a perfect score this time, I was still happy with the 86% accuracy. This indicates that most of the names were correctly identified according to the criteria I set. However, there were some names that did not match the expected classification, leading to the remaining 14% incorrect classifications.

This experience taught me that even when using relevant features, there can still be challenges in classification. Some names that started with a vowel may have been longer than 4 characters, causing them to be misclassified. This result made me realize the importance of carefully selecting and testing features to improve accuracy.

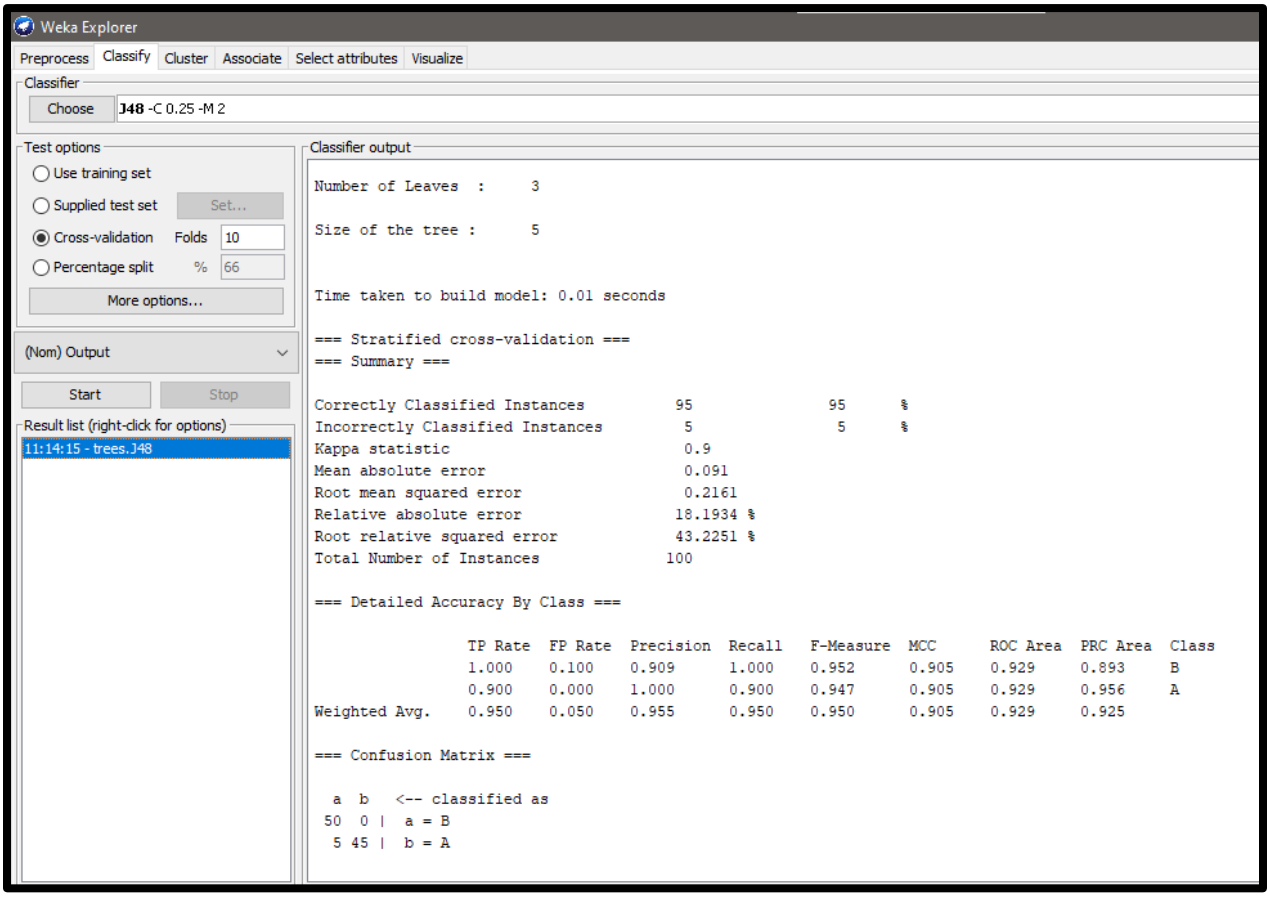


Run the j48 classification algorithm on 3rd arrf file

When I ran the third arff file in weka, I was pleased to see that I achieved 95% accuracy. This means that 95% of the names in my dataset were correctly classified based on the features I selected. For this file, I used the same attribute as the first file, which was names that start with a vowel, along with the attribute that the second letter of the name is "h."

This result shows that the combination of these two features worked very well for classification. Even though it wasn't a perfect score, the 95% accuracy indicates that only 5% of the names were misclassified. This is a strong performance, and it reflects the effectiveness of the features I chose.

Running this experiment helped me understand how slight changes in features can greatly impact the results. It also reinforced my belief that well-chosen attributes can lead to much better accuracy in machine learning tasks. Overall, I was very satisfied with the outcome of this file, and it motivated me to continue refining my feature selection in future experiments.



My Experience

Working with the standard machine learning pipeline was a very interesting and helpful experience for me. I started by looking closely at the text data to find important features, focusing on names and their specific characteristics. This first step of finding features was important because I learned that choosing the right features is essential for getting good results. After I created my arff files, I uploaded them into weka, where I could see the data and understand it better. I ran different classification algorithms like j48, which showed me how well my chosen features worked. I got different results**, including a perfect score of 100% accuracy**, an 86% accuracy, and a strong 95% accuracy with different combinations of features. Each result taught me something new about how important it is to pick the right features for accurate classification. Overall, this process helped me understand the ml pipeline better and the steps needed to prepare and analyze data for successful machine learning projects.