Don Bosco Institute of Technology, Mumbai 400070
Department of Information Technology

Date: 06/04/21

Name: Shruti Chaube
Roll No.: 66

Experiment No.: 3

Title: Classification Algorithm

Problem Definition: Using an open source tool, Weka, to implement classifier algorithms like C4.5 (in Weka J48) on weather data.

Pre-requisite: Weka software, appropriate Dataset

Theory:

The idea of Classification Algorithms is pretty simple. You predict the target class by analyzing the training dataset. This is one of the most, if not the most essential concept you study when you learn Data Science.

What is Classification?

We use the training dataset to get better boundary conditions which could be used to determine each target class. Once the boundary conditions are determined, the next task is to predict the target class. The whole process is known as classification.

Target class examples:

- Analysis of the customer data to predict whether he will buy computer accessories **(Target class: Yes or No)**
- Classifying fruits from features like color, taste, size, weight **(Target classes: Apple, Orange, Cherry, Banana)**
- Gender classification from hair length **(Target classes: Male or Female)**

Applications of Classification Algorithms

- Email spam classification
- Bank customers loan pay willingness prediction.
- Cancer tumor cells identification.
- Sentiment analysis
- Drugs classification
- Facial key points detection
- Pedestrians detection in an automotive car driving.

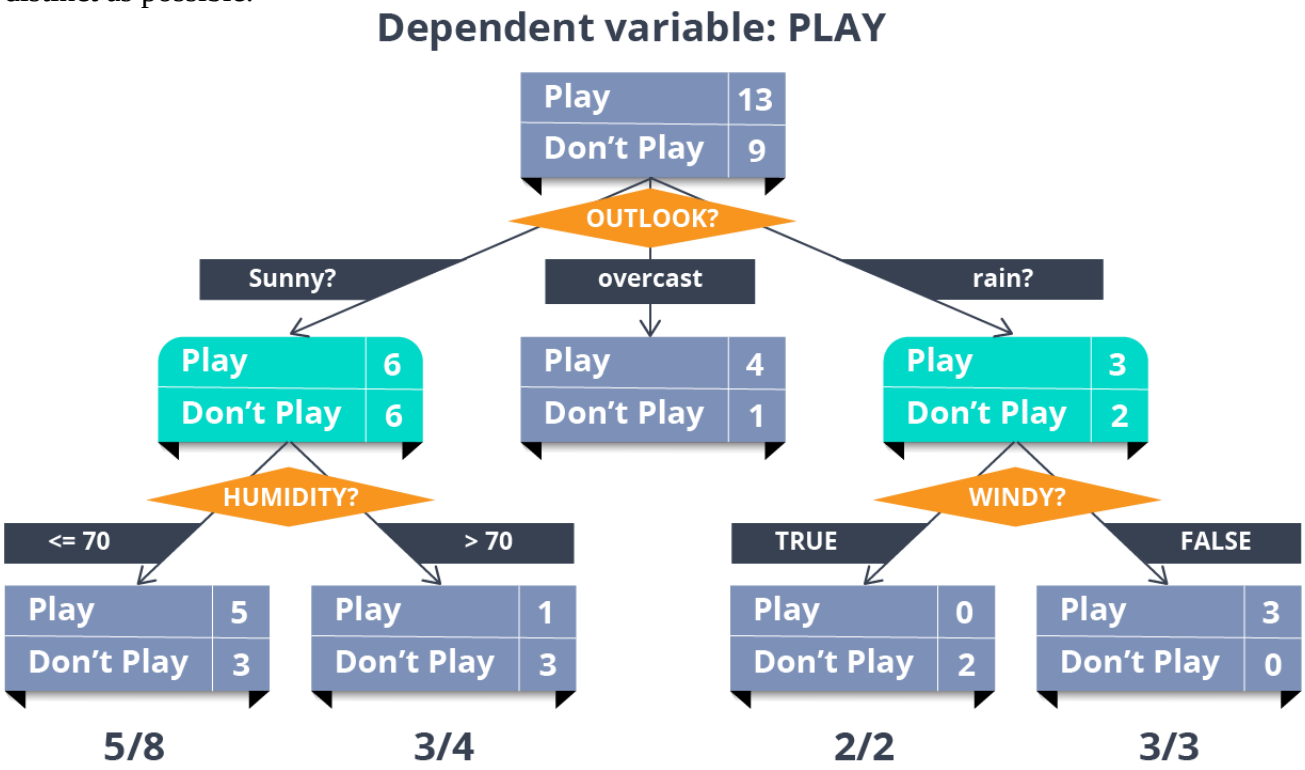
Types of Classification Algorithms

Classification Algorithms could be broadly classified as the following:

- **Linear Classifiers**
  - Logistic regression
  - Naive Bayes classifier
  - Fisher's linear discriminant
- **Support vector machines**
  - Least squares support vector machines
- **Quadratic classifiers**
- **Kernel estimation**
  - k-nearest neighbor
- **Decision trees**
  - Random forests
- **Neural networks**
- **Learning vector quantization**

Decision Trees

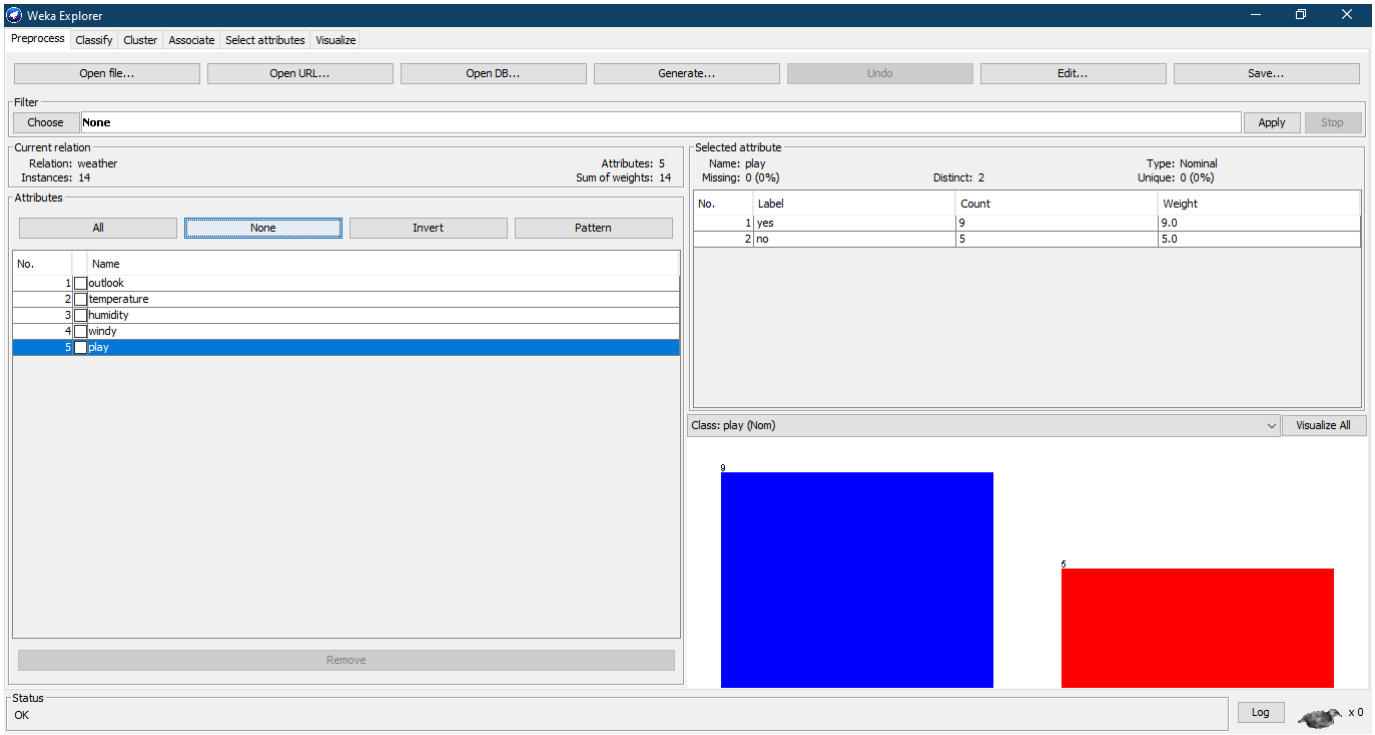
Now, the **decision tree** is by far, one of my favorite algorithms. With versatile features helping actualize both categorical and continuous dependent variables, it is a type of supervised learning algorithm mostly used for classification problems. What this algorithm does is, it splits the population into two or more homogeneous sets based on the most significant attributes making the groups as distinct as possible.



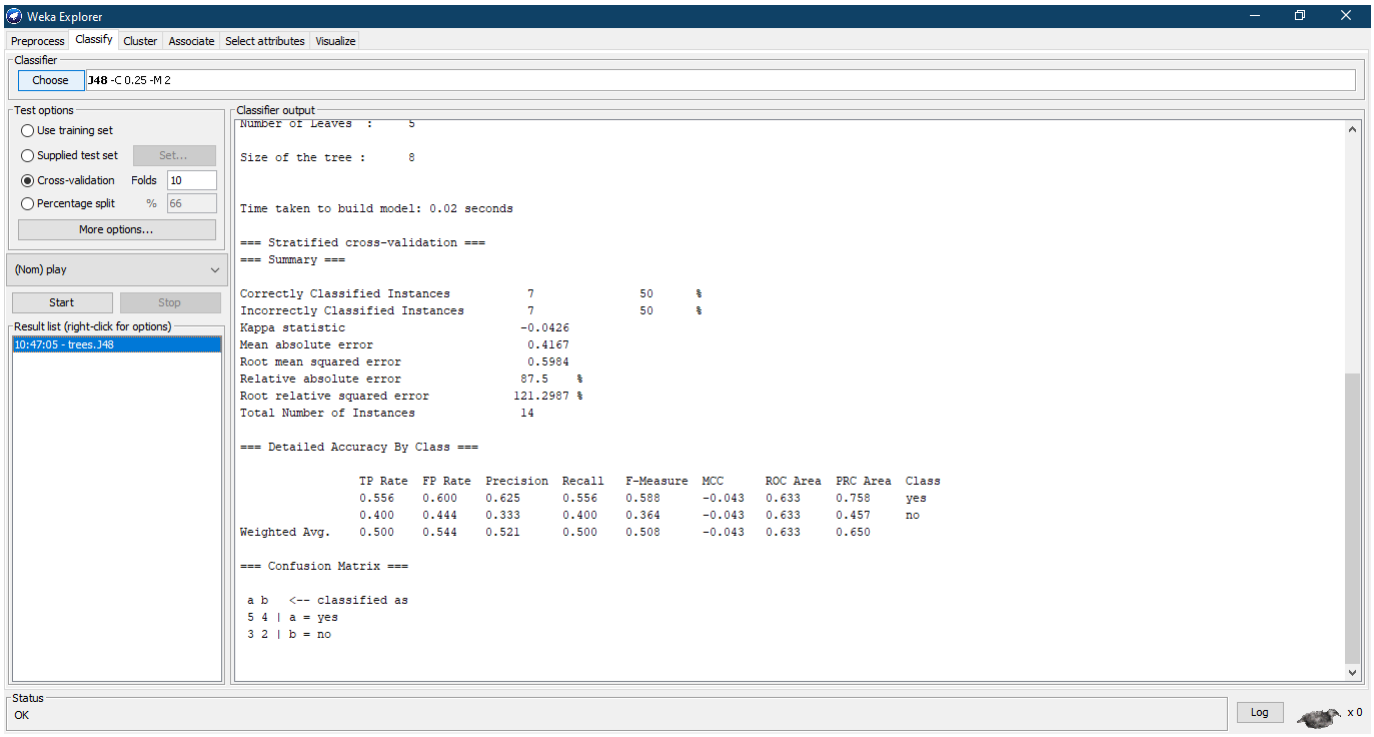
In the image above, you can see that population is classified into four different groups based on multiple attributes to identify 'if they will play or not'.

Procedure:

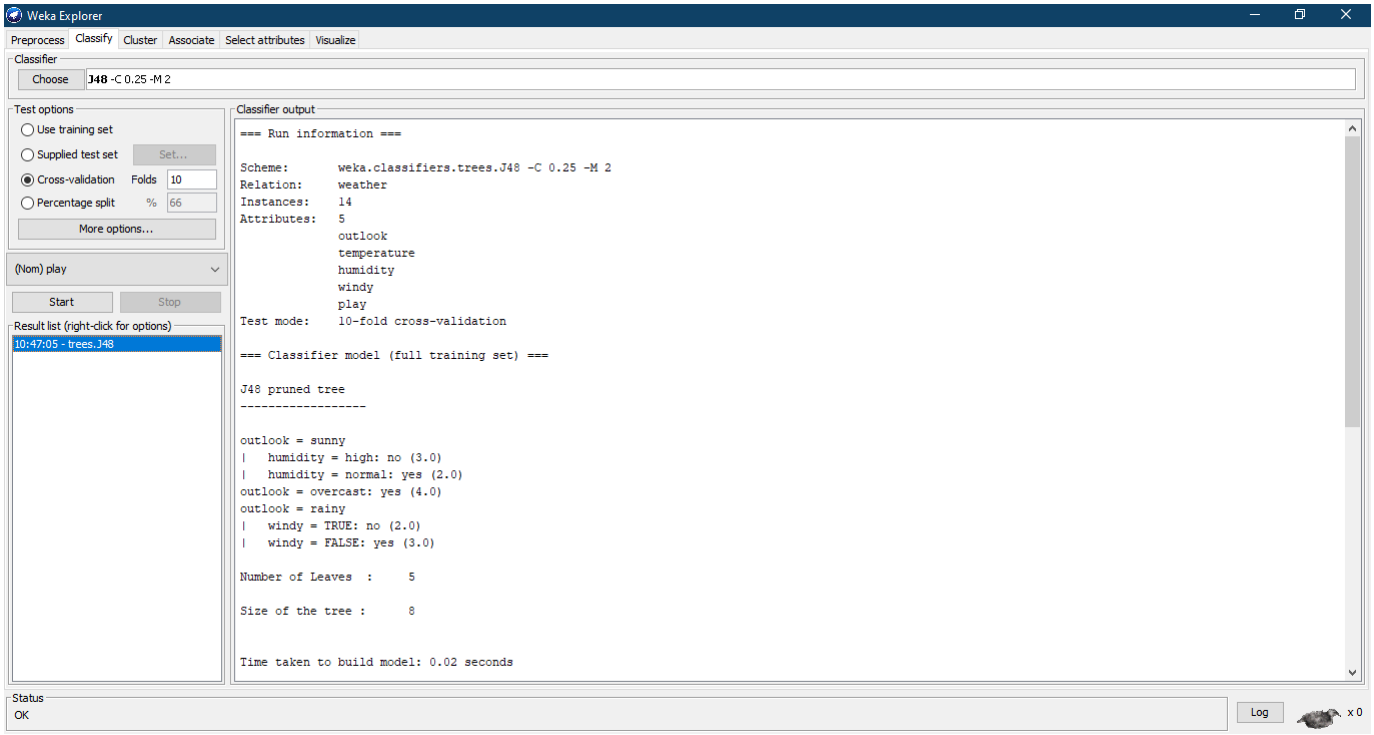
In order to classify our data, first we need to load the dataset. This will be done in wekaexplorer window. Here, we have loaded wheather dataset having 14 instances and 5 attributes.



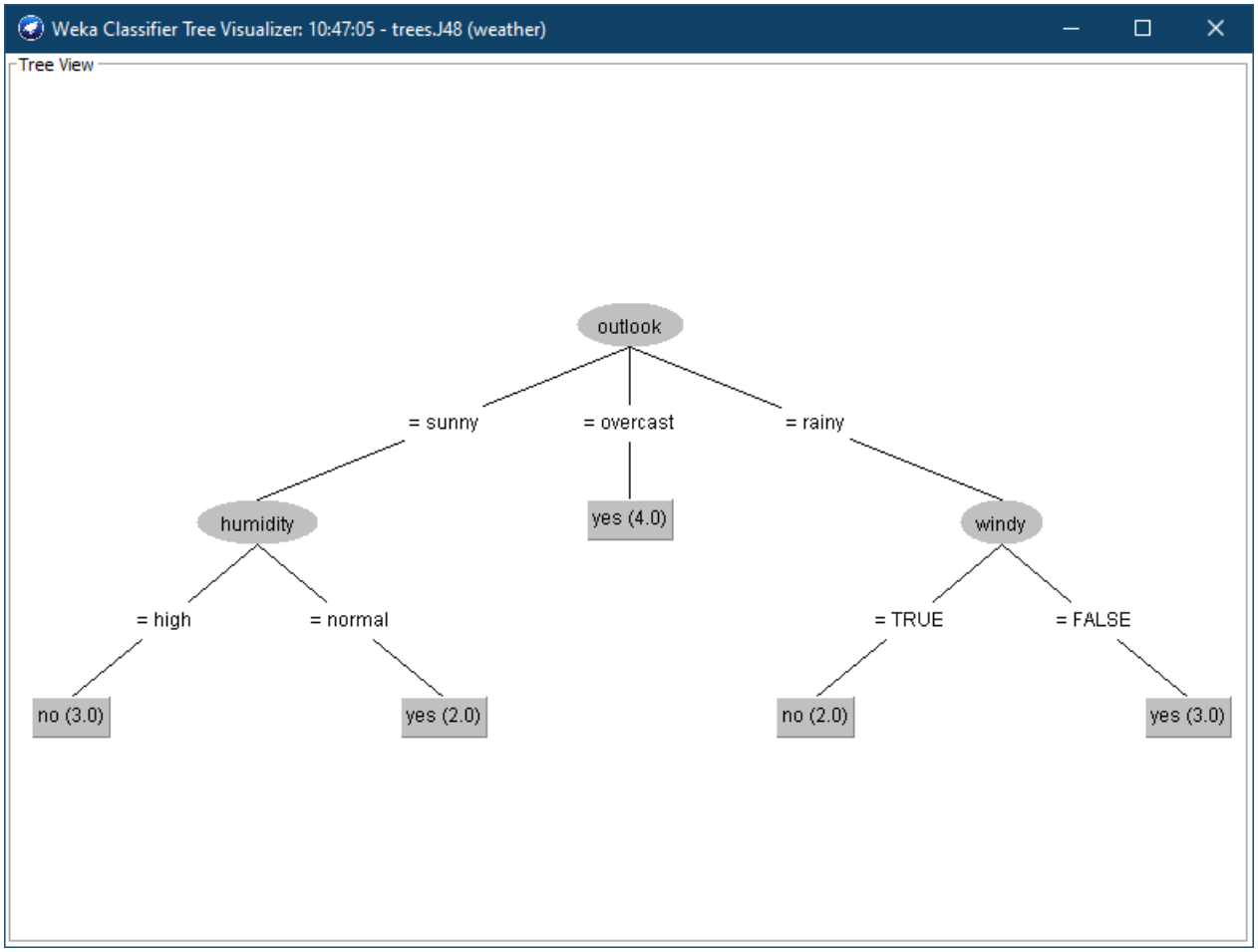
We need to click on classify tab. This window consists of various classifiers like bays, functions, lazy, meta and tree etc. available in weka. We first click on trees, then choose J48 (c4.5 is termed as J48 in weka software) which results in following figure.



After setting parameters we choose one of the test methods as four are given here. Finally click on the start button to build decision tree. The right hand side window shows classifier output, which shows some text and numeric values which is not easy to interpret.



So let us look at graphical representation of this tree by choosing visualize tree option. In this way weka grow decision tree applying c4.5 algorithm of data classification technique in data mining.



Conclusion:

Thus, we learn to implement classifier algorithms like C4.5 (in Weka J48) on weather data.

References:

https://www.edureka.co/blog/classification-algorithms/

http://www.ijsrp.org/research-paper-1013/ijsrp-p2278.pdf