# Self-Review Questions

1.Define the following terms (compare and contrast concepts) :

a) Supervised learning and Unsupervised learning

Supervised learning is a type of machine learning which uses algorithm to give out the result or predict the future using the data which are identified or validated.

Unsupervised learning is a machine learning process which uses to algorithm to predict the outcome or discover the relationship among the data which are not labelled or non-validated.

b) Classification and Prediction

Classification is the process of breaking down of data into similar data and categories. This process helps in learning the patterns and algorithms to predict the future or give out the new or unseen data after the training data.

Prediction is the process of using the training model to make estimation for a target variable based on a set of input variables and features.

c) Confusion matrix (Contingency table), Accuracy, and Error rate

A confusion matrix helps us to visualize and summarize the result of algorithm of the data which we input and it is seen as a table.

Accuracy is the measurement of the ratio of correct prediction made by the machine or prediction made after running the algorithm.

Error rate is the number of mistakes made when the process of prediction takes place in machine learning.

d) Cross-validation and Bootstrap

Cross-validation divides a dataset into numerous subgroups or folds. The model is trained on a subset of the data, with the remainder utilized for validation.

Bootstrap is a statistical resampling method for calculating the sampling distribution of a statistic. By randomly sampling the original dataset with replacement, numerous bootstrap samples are generated.

2.Outline major steps required by the classification process.

Major steps required by the classification process are

1. Data collection
2. Data preparation
3. Data splitting
4. Model selection
5. Model evaluation
6. Model training
7. Model turning
8. Deployment

3.With examples, explain the difference between accuracy and precision.

Accuracy is the overall correctness of the result of machine learning or the prediction of the process.

Precision is a performance metric that measures the proportion of true positive predictions among all positive predictions made by a model.

For example

Consider a model that has been trained to determine whether or not an email is spam. It is applied to 1000 emails, with the following outcomes:

True positives (spam emails correctly classified as spam) = 150

False positives (non-spam emails incorrectly classified as spam) = 50

True negatives (non-spam emails correctly classified as non-spam) = 750

False negatives (spam emails incorrectly classified as non-spam) = 50

Using these results, we can calculate the accuracy and precision of the model:

Accuracy = (150 + 750) / (150 + 750 + 50 + 50) = 90%

Precision = 150 / (150 + 50) = 75%

4.Why is a confusion-matrix-based evaluation method better than a simple accuracy based measure?

Confusion-matrix-based evaluation method is better than a simple accuracy based measure because they give a better idea of a model's performance than classification accuracy does.

5.Compare the advantages and disadvantages of eager classification versus lazy classification.

The following benefits of eager classification:

Speed: The categorization process is quick and effective once the model has been trained.

Better generalization: Typically, eager classifiers are better at classifying and generalizing new data points that they have never seen before.

More precise: Since eager classifiers were trained on a specific dataset and discovered patterns and correlations within the data, they are frequently more precise than lazy classifiers.

The following are disadvantages of eager classification:

Prone to overfitting: If the model is too sophisticated or the dataset is small, eager classifiers may overfit the data.

Lack of adaptability: After the model is trained, adding new data to it requires retraining the model from scratch.

Large datasets are less suitable for: Large amounts of memory are necessary for eager classifiers.

Advantages of lazy classification include:

Flexibility: Lazy classifiers can be updated easily with new data without having to retrain the entire model.

Less prone to overfitting: Lazy classifiers are less likely to overfit the data since they do not make any assumptions about the underlying distribution of the data.

Better suited for large datasets: Lazy classifiers do not require a lot of memory since they do not store the entire model.

Disadvantages of lazy classification include:

Slower prediction: Lazy classifiers can be slower to make predictions since they require searching through the entire dataset to find similar instances.

More sensitive to noise: Lazy classifiers may be more sensitive to noisy data or outliers since they are based on the similarity of instances.

May require more data: Lazy classifiers may require more data to make accurate predictions since they rely on the similarity of instances.

6.Briefly explain overfitting and underfitting. Which of these are important issues in classification?

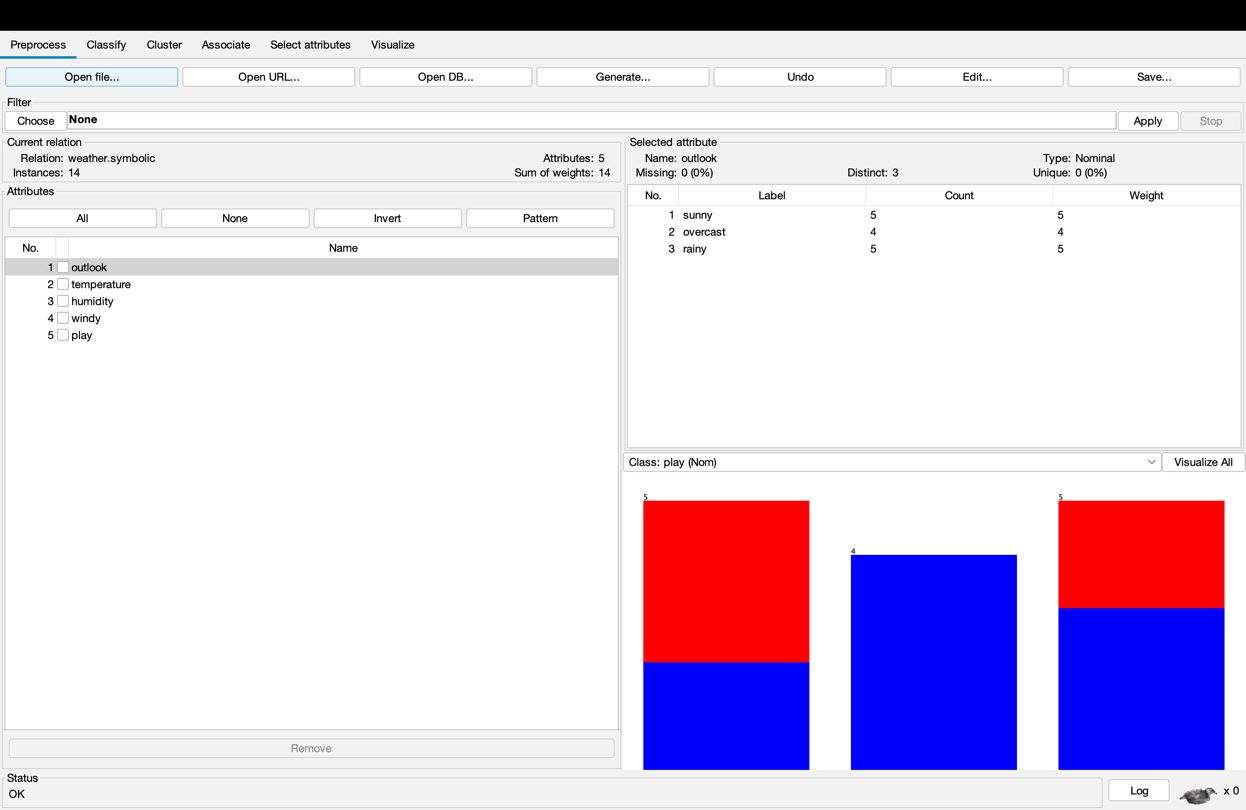
Overfitting happens when a model is too complicated and starts to fit the training data too closely, leading to poor generalization to new, unforeseen data.

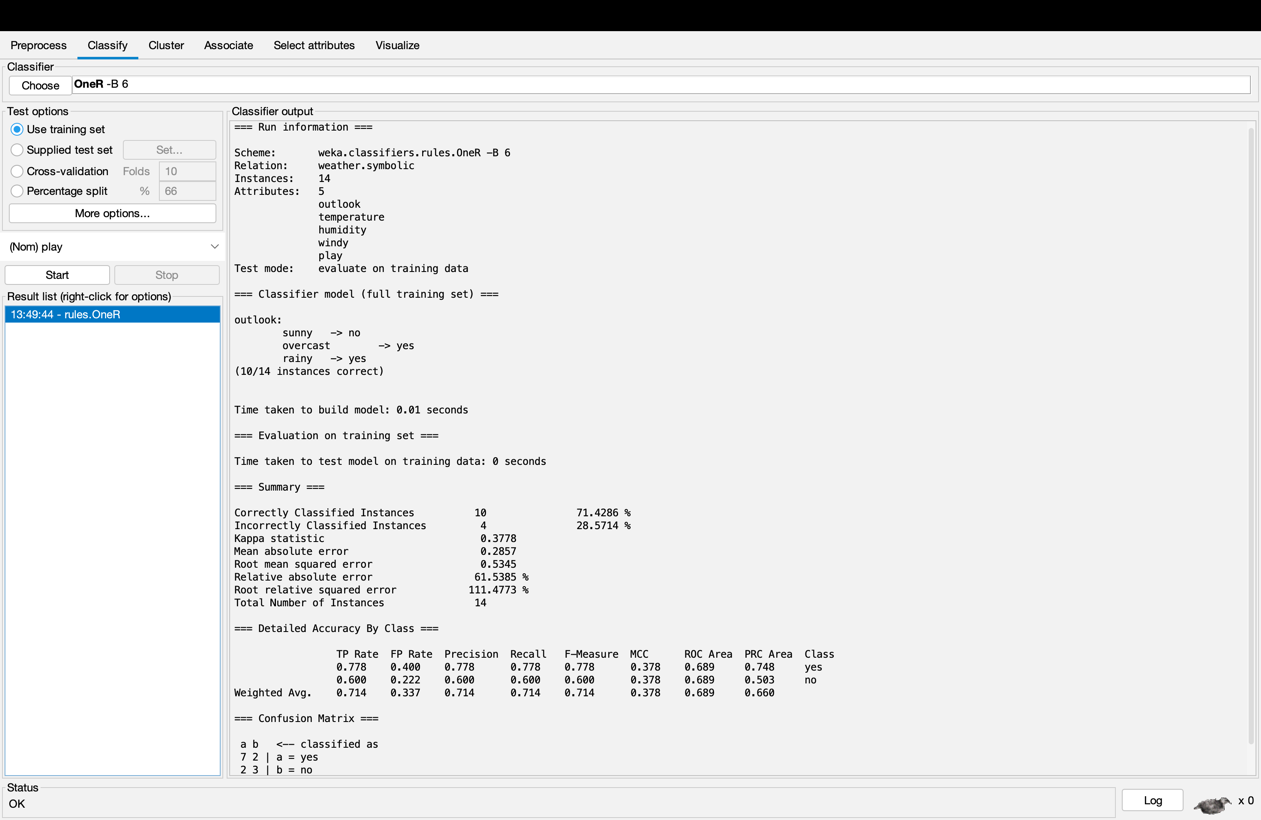
On the other side, underfitting happens when the model is too straightforward and doesn't adequately capture the fundamental patterns and connections in the data, leading to subpar performance on both the training and test sets of data.

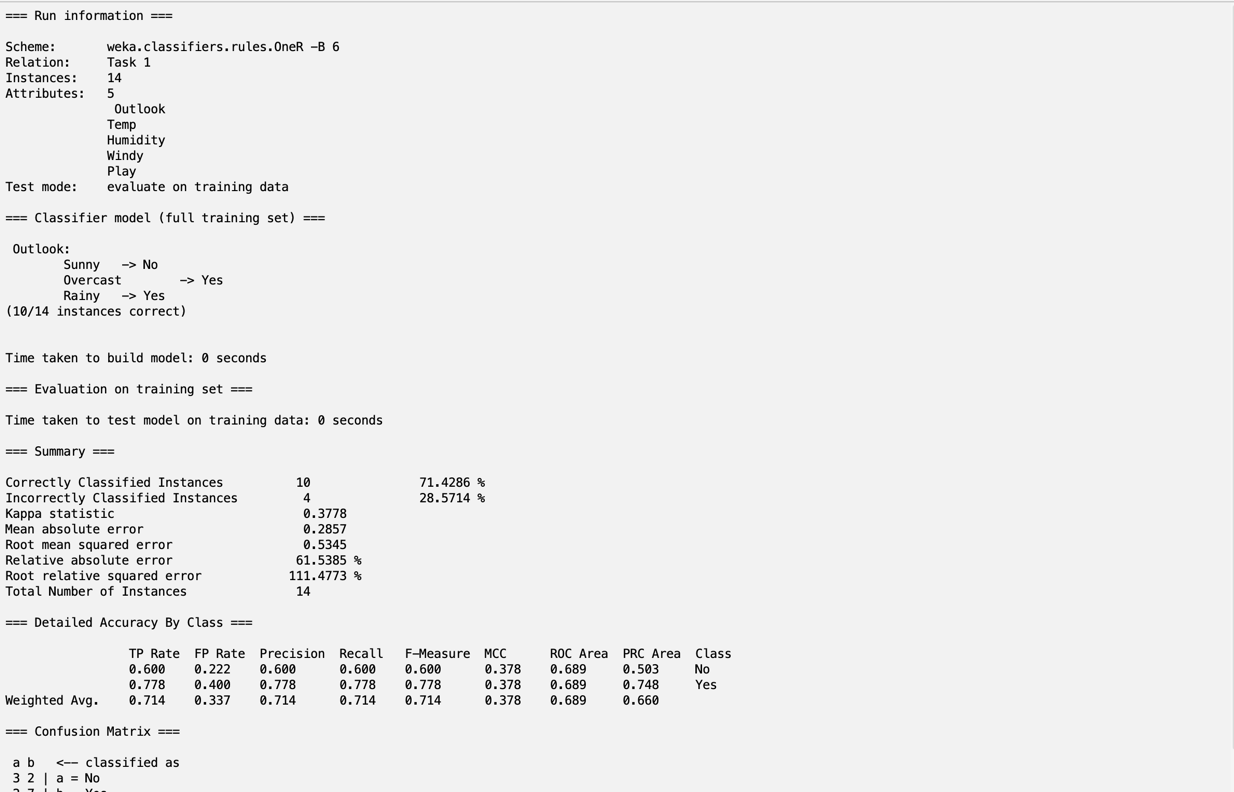
In classification issues, overfitting and underfitting are possible, so it's crucial to strike a balance between the model's complexity and the amount and complexity of the dataset.

# Practical Tasks: Classification through WEKA

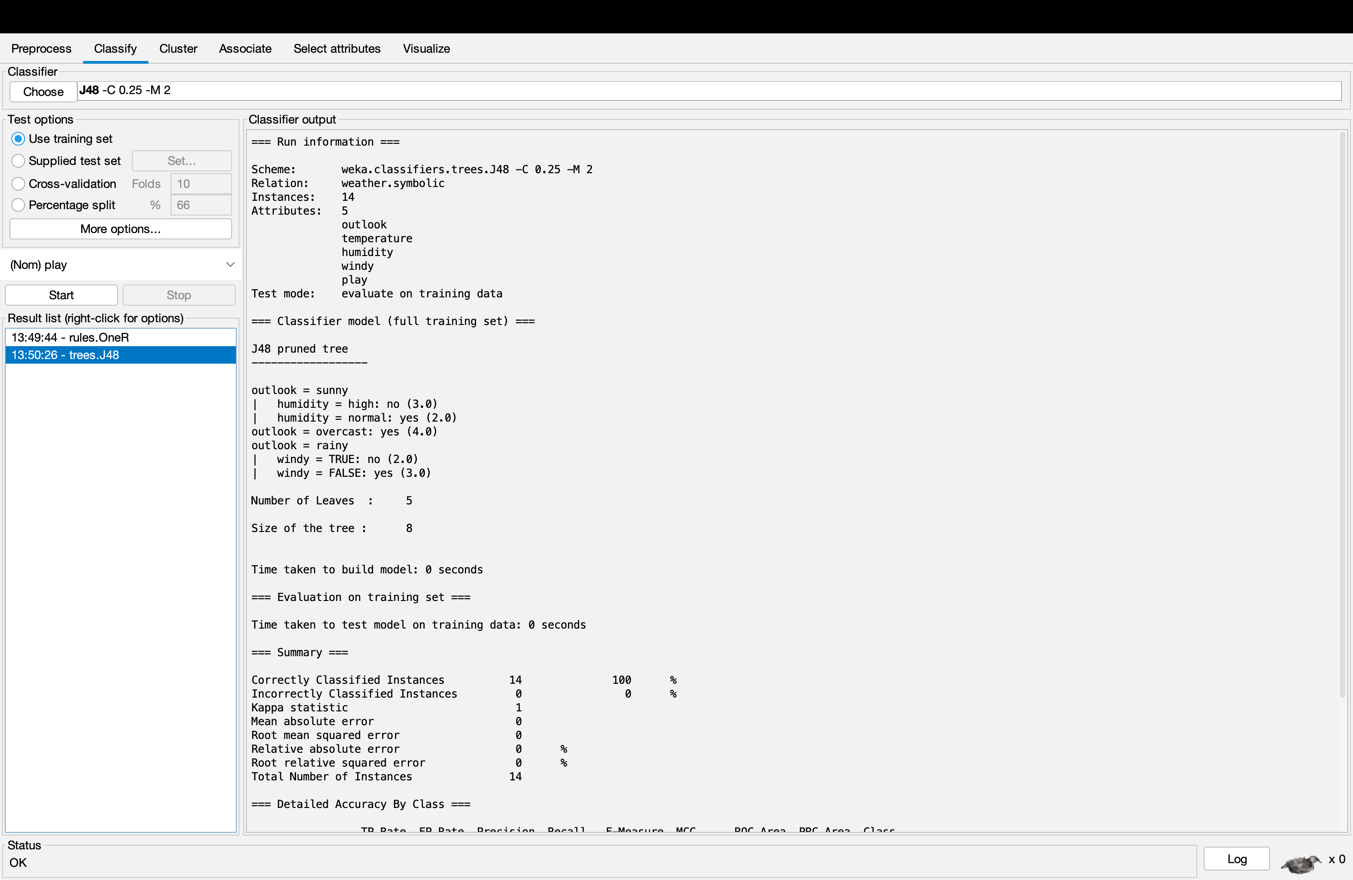
**Task 1: Using OneR (1R) Classification**

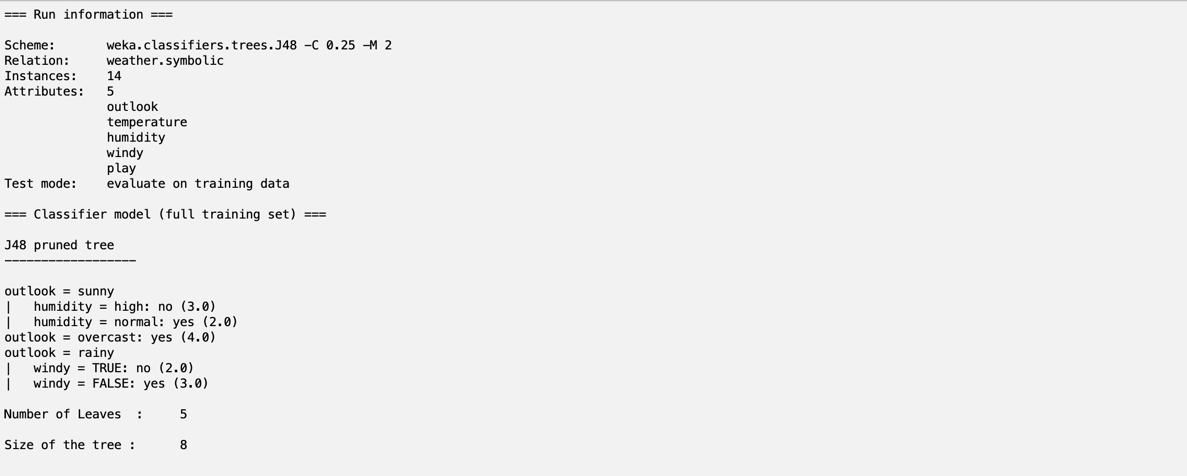
****

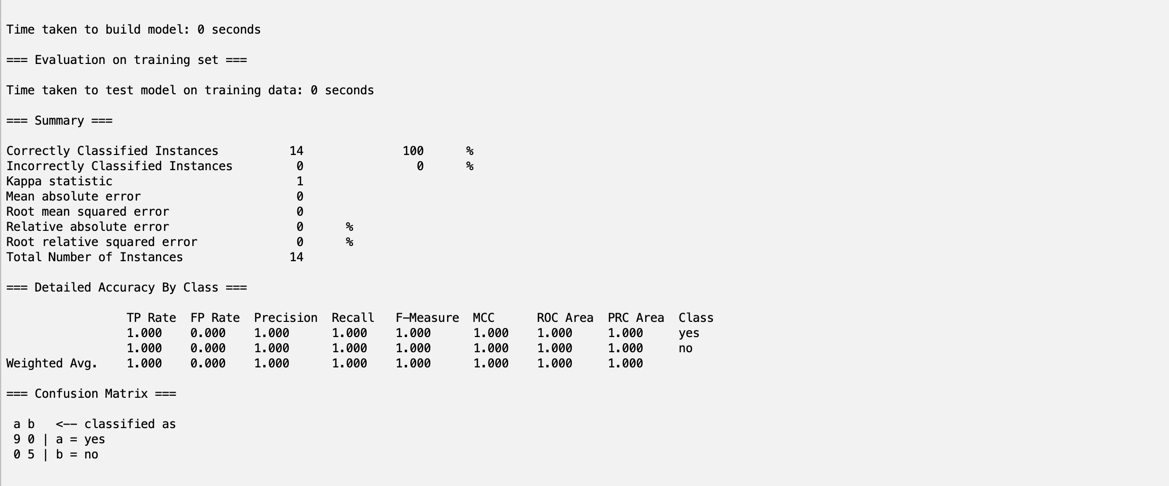
****

****

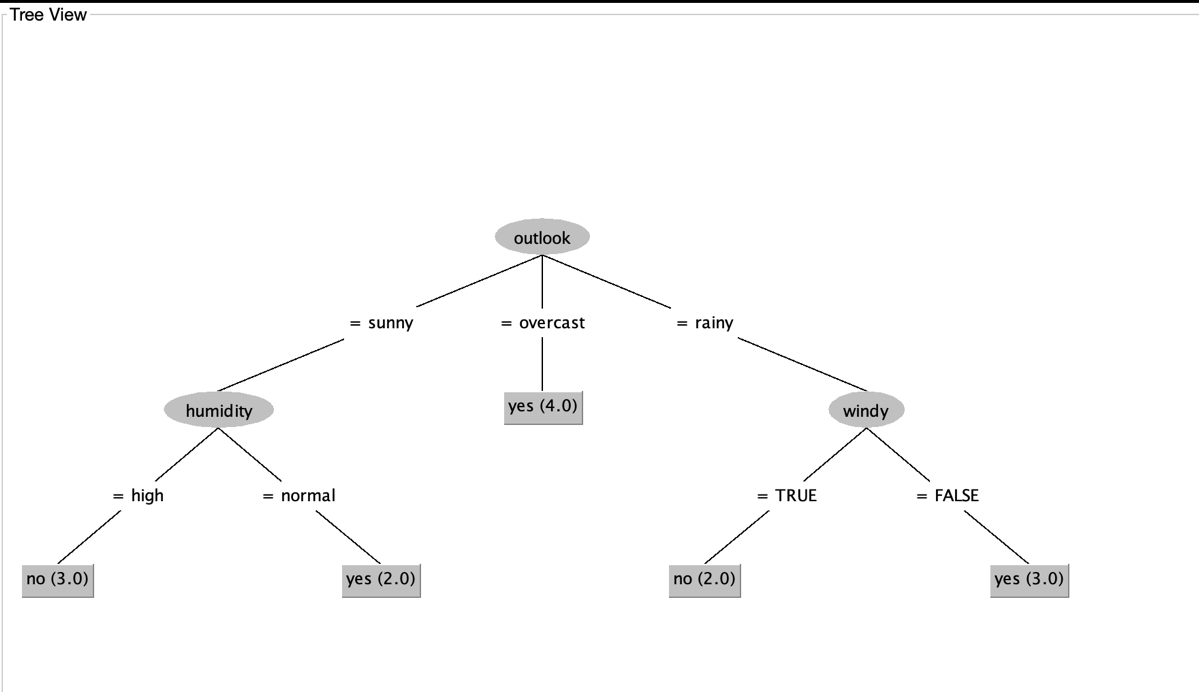
**Task 2: Constructing a decision tree with the weather data**

****

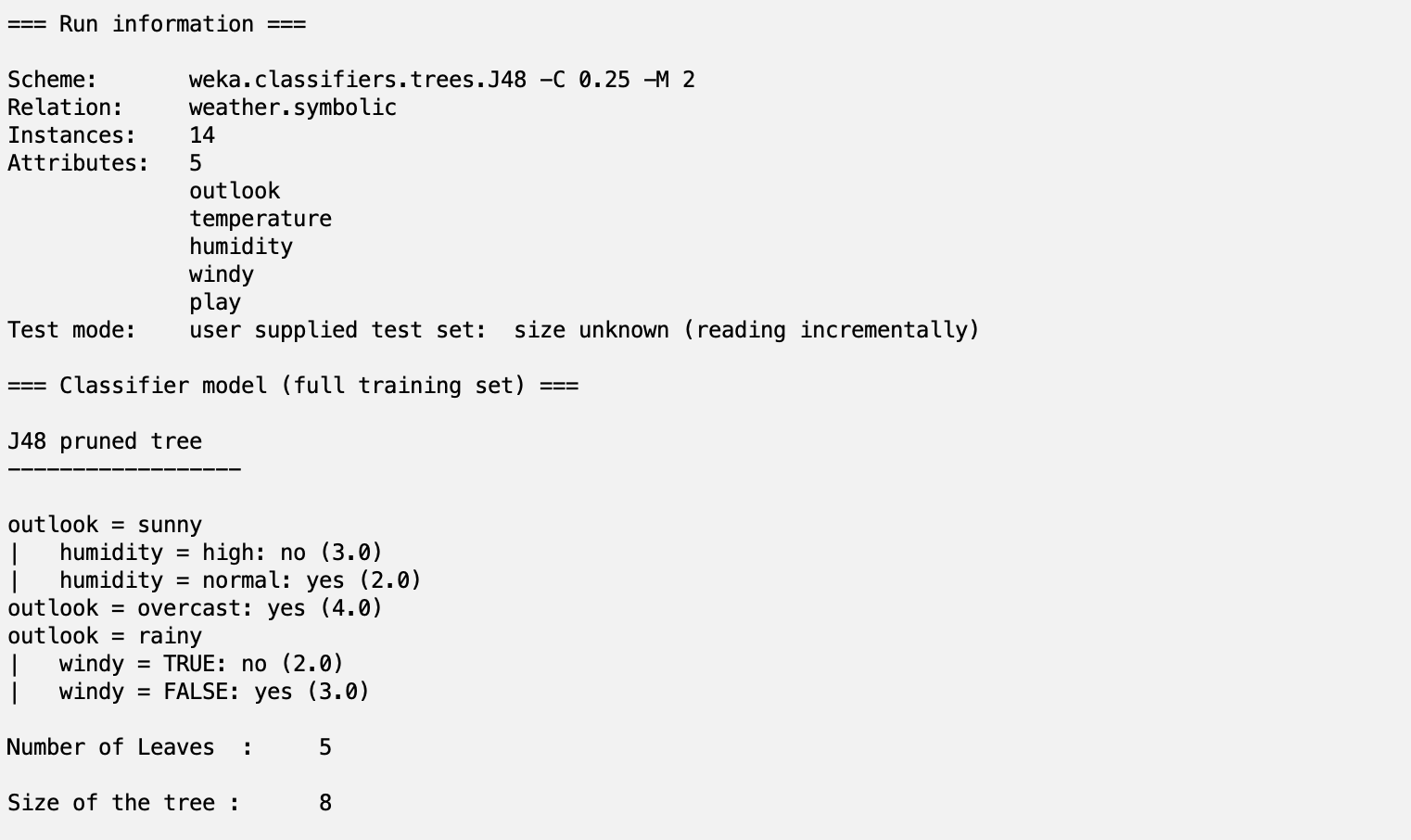
****

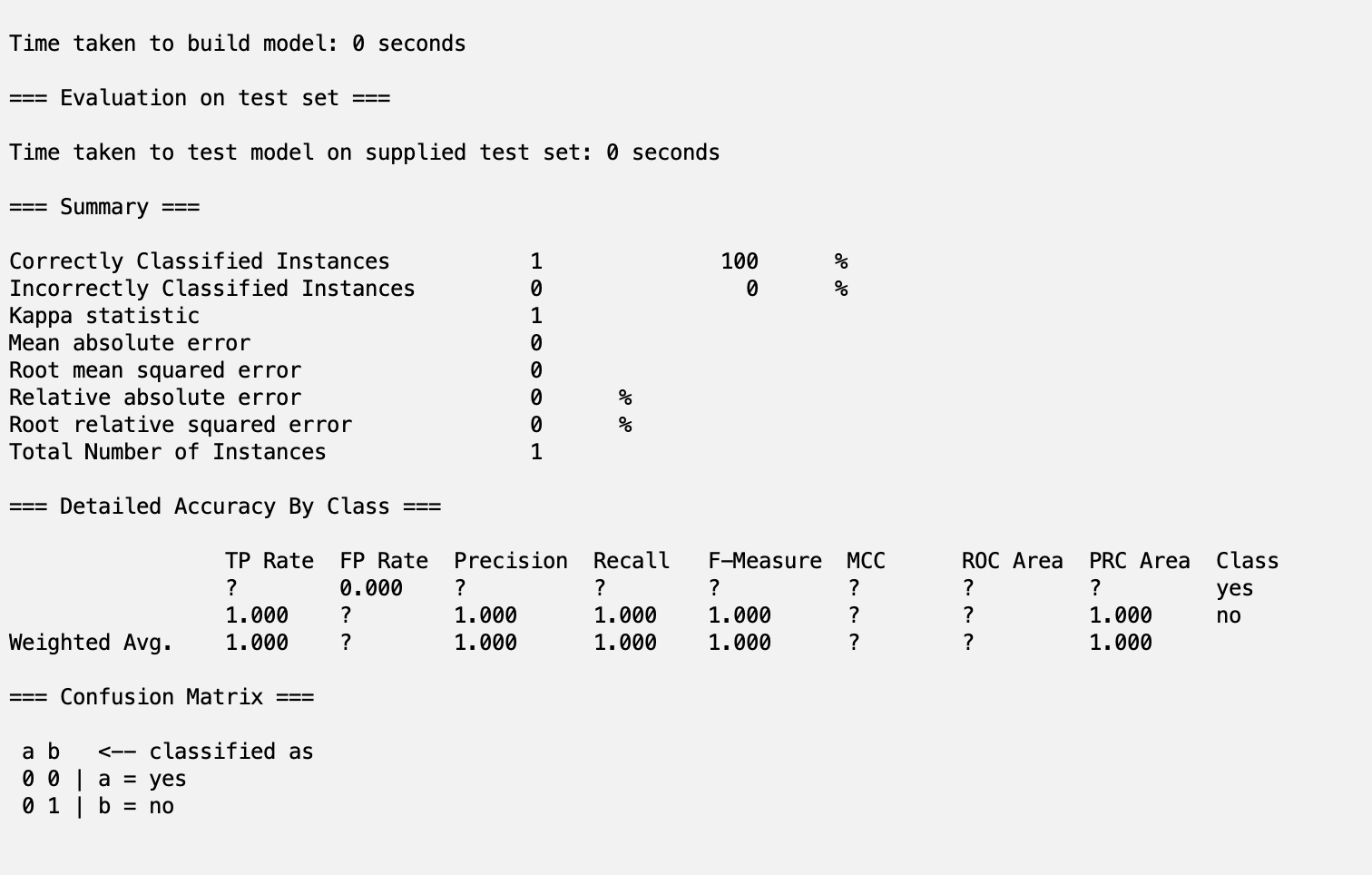
****

**Tree visualization of weather.symbolic after using J48 algorithm**

****

**Task 3: Testing the decision tree classifier constructed**

****

****