

Data Partitioning and Modeling

The data was partitioned into train and test datasets.

The training data set was used to create the decision tree model.

The trained model was then applied to the test dataset.

This is important because it could avoid overfitting. If we train all data to create the model and test it using the same data. The model is to memorize the data and unable to handle the unknown situation, which leads to overfitting.

When partitioning the data using sampling, it is important to set the random seed because we need the modeling results to be reproducible so that our conclusion could be persuasive.

A screenshot of the resulting decision tree can be seen below:

