# Experiment

***Treatment:***

| Treatment | Decision Tree | Train Quantity | Test Quantity | Accuracy |
|---|---|---|---|---|
| 1 | | | 4 | |
| 2 | | | 40 | |
| 3 | | | 400 | |
| 4 | | 4 | 4000 | |
| 5 | | | 4 | |
| 6 | | | 40 | |
| 7 | | | 400 | |
| 8 | | 40 | 4000 | |
| 9 | | | 4 | |
| 10 | | | 40 | |
| 11 | | | 400 | |
| 12 | | 400 | 4000 | |
| 13 | | | 4 | |
| 14 | | | 40 | |
| 15 | | | 400 | |
| 16 | Original | 4000 | 4000 | |
| 17 | | | 4 | |
| 18 | | | 40 | |
| 19 | | | 400 | |
| 20 | Library | 4 | 4000 | |

| | | | | |
|---|---|---|---|---|
| 21 | | | 4 | |
| 22 | | | 40 | |
| 23 | | | 400 | |
| 24 | | 40 | 4000 | |
| 25 | | | 4 | |
| 26 | | | 40 | |
| 27 | | | 400 | |
| 28 | | 400 | 4000 | |
| 29 | | | 4 | |
| 30 | | | 40 | |
| 31 | | | 400 | |
| 32 | | 4000 | 4000 | |

*1000 repetitions per treatment.*

**Experimental Unit:**

- Decision Trees

**Response Values:**

- The accuracy of the decision tree classification

**Experimental Factors:**

- **Studied:**

- Training set data quantity.

- Testing set data quantity.

- Decision tree variant.

- **Not studied:**

- Training set data values.

- Testing set data value.
- Decision tree internal structure.

**Observational Factors:**

- Random selection of an invalid case.

**Factor Levels:**

- **Decision Tree:** Original, Library
- **Train Quantity:** 4, 40, 400, 4000

**Test Quantity:** 4, 40, 400, 4000

**Results:**

The results of the experiment are available here:

https://github.com/Esarac/FungiParadise/blob/master/Experiment/results.csv

**Analysis of the results:**

Given the results, we want check whether there is a statistically significant difference between our original decision and the decision tree from the Accord library. Also, we want to check if there is a statistically significant difference between the Train Quantity groups. We will use the Two-Way ANOVA statistical tool to accomplish these objectives. We will analysis the Decision Tree variable groups, the Train Quantity variable groups, and a combination of Decision Tree and Train Quantity variable groups through ANOVA.

To do the Two-Way ANOVA analysis, we created a program in Python using the libraries pandas and statsmodels available in the latest Anaconda release. The program source code is available here:

https://github.com/Esarac/FungiParadise/blob/master/Experiment/experiment.py

We will use an alpha value $\alpha = 0.05$

Hypotheses for the Decision Tree variable:

$H_0$ (Null hypothesis): The decision trees are equally accurate in their prediction accuracy

$H_a$ (Alternate hypothesis): The decision trees are not equally accurate in their prediction accuracy

Hypotheses for the Train Quantity variable:

$H_0$ (Null hypothesis): All the train quantity groups yield the same prediction accuracy

$H_a$ (Alternate hypothesis): All the train quantity groups yield a different prediction accuracy

Hypotheses for the Decision Tree & Train Quantity combination variables:

$H_0$ (Null hypothesis): All the decision tree and train quantity combinations yield the same prediction accuracy

$H_a$ (Alternate hypothesis): All the decision tree and train quantity combinations yield a different prediction accuracy

The program gave us the following tables after analysis the experimental data:

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| C(Decision_Tree) | 0.011755 | 1.0 | 1.182378 | 2.769546e-01 |
| C(Train_Quantity) | 85.485049 | 3.0 | 2866.116064 | 0.000000e+00 |
| C(Decision_Tree):C(Train_Quantity) | 0.459460 | 3.0 | 15.404644 | 5.985984e-10 |
| Residual | 31.734965 | 3192.0 | NaN | NaN |

***Table 1****: Two-Way Anova*

| | Decision_Tree | Train_Quantity | Test_Quantity | Repetition | Accuracy |
|---|---|---|---|---|---|
| 0 | Library | 4 | 1111.0 | 50.5 | 0.625749 |
| 1 | Library | 40 | 1111.0 | 50.5 | 0.933595 |
| 2 | Library | 400 | 1111.0 | 50.5 | 0.992890 |
| 3 | Library | 4000 | 1111.0 | 50.5 | 0.999995 |
| 4 | Original | 4 | 1111.0 | 50.5 | 0.584793 |
| 5 | Original | 40 | 1111.0 | 50.5 | 0.959644 |
| 6 | Original | 400 | 1111.0 | 50.5 | 0.992714 |
| 7 | Original | 4000 | 1111.0 | 50.5 | 0.999744 |

***Table 2****: Accuracy means*

Important values for the Decision Tree variable (**Table 1**):

- F-value: 1.182378
- P-value: 0.2769546

Since our p-value is greater than our $\alpha$ $(0.05)$, we will not reject the null hypothesis.

Important values for the Train Quantity variable (**Table 1**):

- F-value: 2866.116064
- P-value: 0

Since our p-value is less than our $\alpha$ $(0.05)$, we will reject the null hypothesis and accept the null hypothesis.

Important values for the Decision Tree & Train Quantity variables (**Table 1**):

- F-value: 15.303644
- P-value: $5.985984 \times 10^{-10}$

Since our p-value is less than our $\alpha$ $(0.05)$, we will reject the null hypothesis and accept the null hypothesis.

In addition, we can see in **Table 2** that as the Train Quantity increases, the mean of the prediction accuracy increases.

Evaluation and Conclusions:

From the ANOVA analysis and the data obtained through the Python program we can draw the following conclusions:

1. There is no significant statistical difference between our original decision tree implementation and the Accord library decision tree implementation
2. The prediction accuracy changes depending on the train quantity (as the train quantity increases, the prediction accuracy generally increases).
3. The increase in prediction accuracy that occurs as the train quantity increases can be observed in both decision tree implementations (our original implementation and the Accord library implementation).

This is great news for our team, since this means that we somehow managed to create a decision tree implementation that, while not as good, is actually comparable to the decision tree implementation of a trusted third-party library (the Accord library).