The Use of Machine Learning Algorithms to Classify Mushrooms as Either Poisonous or Edible, Given A Dataset Consisting of Different Attributes of Mushrooms Varieties

MACHINE LEARNING – MUSHROOM CLASSIFIER

An AI Driven Poisonous Mushroom Detector

A Final Course Project
Submitted in Partial Fulfilment of the Requirement for the Degree
ofMaster of Science in Applied Data Science

FACULTY OF SCIENCE & TECHNOLOGY
THE UNIVERSITY OF THE WEST INDIES

Mona, Kingston, Jamaica W.I.

May 04, 2022

ABSTRACT

Introduction:

Mushroom poisoning happens when a person consumes mushrooms that contain poisons while looking for benign, similar-looking mushrooms. Mushrooms are the fruiting bodies of a group of higher fungus that have developed for millions of years alongside plants (Horowitz et. al., 2019). According to Horowitz (2019), they are widely disseminated all across the globe. Although there are thousands of types of mushrooms, only around 100 elicit symptoms in humans when consumed, and only 15-20 mushroom species are potentially fatal. There is no easy way to tell the difference between edible and toxic mushrooms. Poisoning occurs in more than 95 percent of mushroom toxicity instances as a result of misidentification (Horowitz et. al., 2019). This predicament presents the perfect opportunity for machine learning algorithms to be used to make the distinction between poisonous and edible mushrooms.

Methodology:

For this machine learning project to classify mushrooms as either edible or poisonous, tree-based machine learning algorithms were utilized. The models used were a Decision Tree, Random Forest, and a Gradient Boosted Tree. The training dataset is multivariate in nature, and consists of 22 categorical attributes, having 6498 instances. The training dataset also consisted of a ground truth category column. The data set was split in the ratio 70% for training and 30% for testing during the training process.

Results:

In this machine learning project, 30% of the training dataset was used to internally validate/ test the following models listed below:

Decision Trees:

- Accuracy : 0.9958656331
- AUC : 0.9957939012
- F1 Score : 0.9958652707
- Recall : 0.9958656331
- Precision : 0.9958989747

Random Forrest:

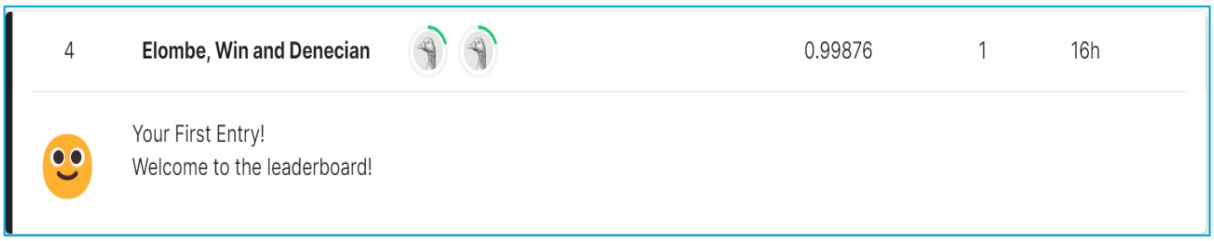
- Accuracy : 0.9989637306
- AUC : 0.9989258861
- F1 Score : 0.9989637306
- Recall : 0.9989637306
- Precision : 0.9989658010

Gradient Boosted Tree:

- Accuracy : 0.9994821155
- AUC : 0.9994450610
- F1 Score : 0.9994821155
- Recall : 0.9994821336
- Precision : 0.9994826359

Conclusion:

Based on the above performance metrics, the best performing machine learning model was the **Gradient Boosted Tree**, with an accuracy of **0.9994821155**, thus it was chosen for the final predictions. Also, the Gradient Boosted Tree performed the best on all the other performance metrics evaluated for such as, AUC, F1 Score, Recall and Precision.



Group Member Contributions

Elombe Calvert: **620087323**

- Build ML Pipeline
- Train ML Models
- Evaluate Models

Win Phyo: **620092749**

- Data Cleaning & Processing
- Evaluate Models
- Project PDF Write-Up

Denecian Dennis: **620062729**

- Evaluate Models
- Train ML Models
- Project PDF Write-Up