Classifying Poisonous Mushrooms using Supervised Learning Algorithms

Mikayla Narothan[1], Oluwademilade Osikoya[1], and Kabelo Komape[1]

[1]University of the Witwatersrand, School of Computer Science and Applied Mathematics, South Africa
*these authors contributed equally to this work

ABSTRACT

Mushrooms are part of the elite super-food group, this is a group of select foods which are loaded with vitamins, while also boosting your bone health and promoting a healthy immune system. A well known fact is that mushrooms can be poisonous depending on its species. While these mushrooms are not edible, there has been recent interest in its medicinal properties for treating diseases. This paper analyzes which is the most suitable classification learning algorithm when grouping mushrooms into poisonous and edible categories. In order to distinguish between these categories a mushroom dataset from the UCI Machine Learning Repository is used, this dataset contains key physical features helpful in classifying mushrooms.

Introduction

Classification of mushrooms is not a new thing, in fact it is a widely popular topic due to various reasons. Mushrooms are part of the elite super-food group but not all mushrooms are a part of the elite super-food group as some mushrooms are poisonous and the toxicity of the poison is dependent on the actual species of mushroom. There has also been some research into mushrooms due to the fact that edible mushrooms are very rich in nutrients, including potassium, vitamin D, proteins, and fiber while being low in fats, carbohydrates and sodium.

Mushroom hunting has started gaining traction due to the fact that mushrooms are very healthy and can grow anywhere so there is no need to pay for it as healthy foods are expensive. Due to this it is imperative that we have a method to classify mushrooms as either poisonous or non-poisonous.

Just a few years ago machine learning algorithms required the processing power of a super computer but thanks to technological advancements an average computer is able to run powerful model. We can implement various supervised learning algorithms on a mushroom dataset from the UCI Machine Learning Repository in order to train a model that is able to accurately predict if a mushroom is poisonous based on physical properties that do not require any special methods or equipment's to measure.

The mushrooms are classified into two classes namely poisonous and non-poisonous so it is a binary classification. We will use a random forest classifier for our primary classification and make a compartive analysis to a neural network model and a logistic regression model.

The reason for using a random forest classifier for our primary model is because due to the way the algorithm is set up the default hyperparameters often produce good prediction results so most times there is no need to tweak the hyperparameters. A big problem with machine learning is over-fitting, but most of the time a random forest classifier will not over-fit especially if there are enough trees in the forest then the model will not overfit. Random forests are also hard to beat performance wise and usually models that can do better performance wise require more time to implement and are usually complex. Logistic regression is another useful algorithm for binary classification and a neural network can be architectured to be a binary classifier, we will implement both algorithms for a comparative analysis with our primary algorithm.

Overall a random forest classifier is a safe bet as it is fast, simple and flexible.

Methodology

This section provides the aim of the research conducted in a formal manner. Each classification algorithm used is then very briefly explained as well as the motivation for implementing each of these algorithms.

Aim of Research

Due to the various reasons, explained in the introduction, mushroom classification is an important task. This research aims to determine the most suitable classification algorithm to classify a mushroom, by comparing performance and speed. The proposed algorithms is an Artificial Neural Network (ANN), Logistic Regression and a Random Forest.

Implementations

Artificial Neural Network

An ANN is based on the biological neural network of the brain. It processes and analyzes information using the structure shown in 1. The network consists of multiple layers: The first layer is the input layer and the last layer is the output layer, any layer in between is part of the hidden network. The input layer takes in the data being modeled. The hidden layers take a summation of the outputs from the previous layer and its weights and transforms this using the most suitable activation function. This is the output that is taken to the next layer and the process is repeated. The two processes used to determine the weights for the model is forward and back propagation. For the purposes of this research a built-in python library is used to implement this ANN.

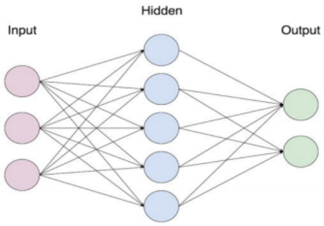


Figure 1. ANN Architecture

Motivate ANN here

Logistic Regression

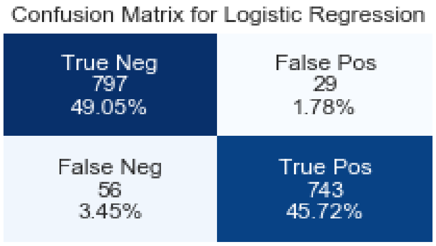
Logistic regression is a probabilistic approach to classifying discrete values. In most cases, it is used to classify binary values which is the case for this research project. A hypothesis function (h_θ) is used to predict future outcomes, this function is:

$$h_\theta(\mathbf{x}) = \frac{1}{1+\exp -\theta^T \mathbf{x}}$$

In the hypothesis function, **x** represents the input, in this case the physical features of mushrooms, and θ represent the weights of each feature. During the learning process, the weights are updated using gradient descent to find the optimal values.

Random Forest

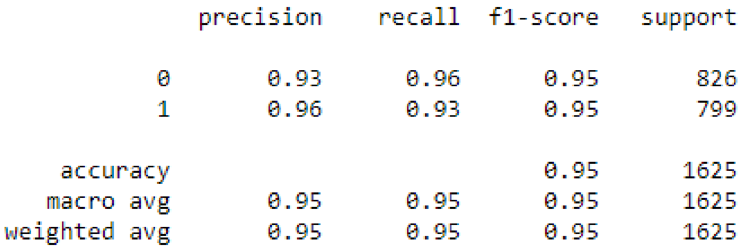
Results



Confusion Matrix for Logistic Regression

True Neg 797 49.05%	False Pos 29 1.78%
False Neg 56 3.45%	True Pos 743 45.72%

Figure 2. Confusion Matrix for Logistic Regression with Dummy Encoding



```
              precision    recall  f1-score   support

           0       0.93      0.96      0.95       826
           1       0.96      0.93      0.95       799

    accuracy                           0.95      1625
   macro avg       0.95      0.95      0.95      1625
weighted avg       0.95      0.95      0.95      1625
```

Figure 3. Classification report for Logistic Regression with Dummy Encoding

Conclusion

References

1. Figueredo, A. J. & Wolf, P. S. A. Assortative pairing and life history strategy – a cross-cultural study. *Hum. Nat.* **20**, 317–330, DOI: https://doi.org/10.1007/s12110-009-9068-2 (2009).

2. Hao, Z., AghaKouchak, A., Nakhjiri, N. & Farahmand, A. Global integrated drought monitoring and prediction system (GIDMaPS) data sets. *figshare* http://dx.doi.org/10.6084/m9.figshare.853801 (2014).

Author contributions statement