Mushroom Classification

MUSHROOM CLASSIFICATION – POISONOUS OR EDIBLE?

Dataset: https://www.kaggle.com/datasets/uciml/mushroom-classification

Intro

I chose this dataset because of my newfound love of mushrooms (the secret is to cook them in soy sauce). In my free time, I have been researching growing mushrooms at home and though there is very little chance I will accidentally grow something poisonous from a grow kit, perhaps one day I will want to venture outdoors for some mushroom foraging. I can use this model to help me decide if they are safe to eat.

I feel the benefits of this dataset for this exercise lie in its huge size. With 8,124 rows and 23 features there is lots to dive into here. Each feature is labelled with a letter to describe the label so I will have to refer to the Datasheet to decode the information. The target variable will be if the mushroom is poisonous or not and with such a vast array of features available, I will experiment with the most effective way to build this model – I have doubted all features will be needed!

Question: *Which features are most indicative of a poisonous mushroom?*

Process

Using the code from the week 6 notebook, I followed through this using the mushroom dataset.

Creating Data Frame

The data contains various features with each feature having various labels. To use the provided code, I needed to give each label a numeric representation. With the help of my good friend Stack Overflow I learnt about the Pandas factorize function[1], which with a simple for loop allowed me to create a numerical representation of the data with ease. Initially I only used an even split of 3,000 mushrooms as per the original code, but I re-ran the same tests again with a split of 8,000 to utilize the size of my data.

Decision Trees:

I ran the code with my new data set and experimented with various features. I had a range of results, the lowest being 55% accuracy but mostly sitting in to 70%-90% range, which is very high! Some brief research on mushrooms suggested checking the gills and the cap colour[2], which gave 85% accuracy – pretty good!
When I ran the code with all the features it gave me 100% accuracy. Whilst this could be true, I want to explore the data further to evaluate if the model is correct or there is an error somewhere.

When examining the plots, their sparsity suggests that either the model is not plotting all the data, or numerous data points are stacked on top of each other.

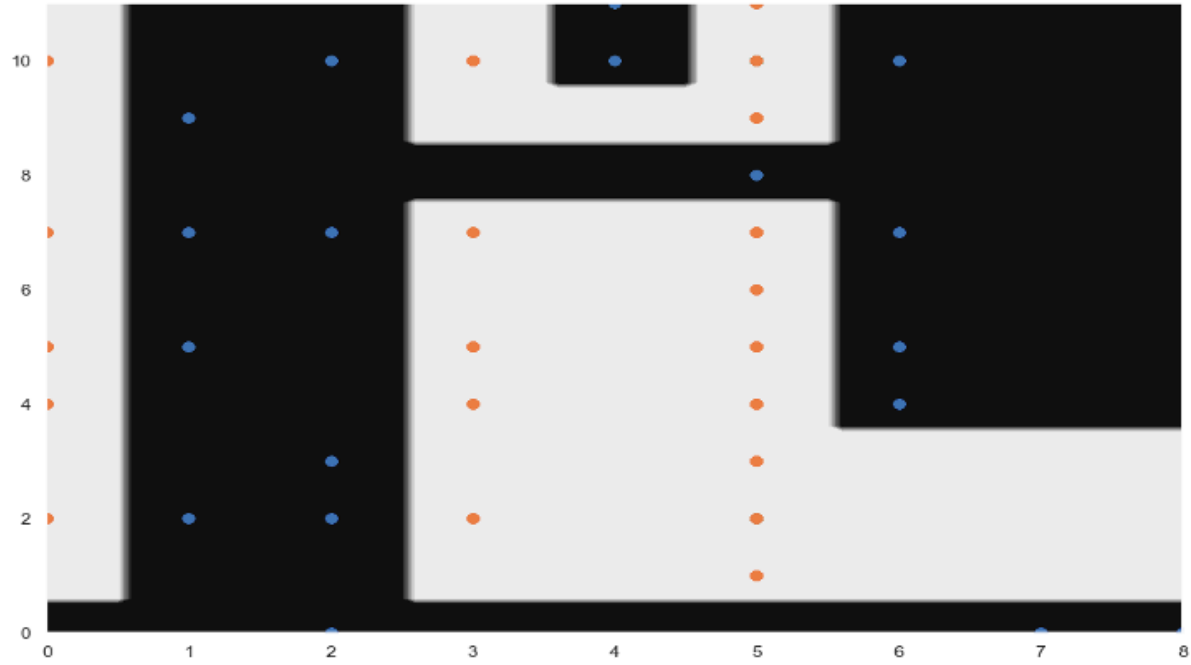

Figure 1: Decision Plot of 'Odour' and 'Gill Colour' Features. Accuracy = 98.91%

Talking through the problem with my classmate she suggested I look at a heatmap of my data and see if there are any high correlations. If I find any, I can try removing them from the data set and seeing if this made any difference.

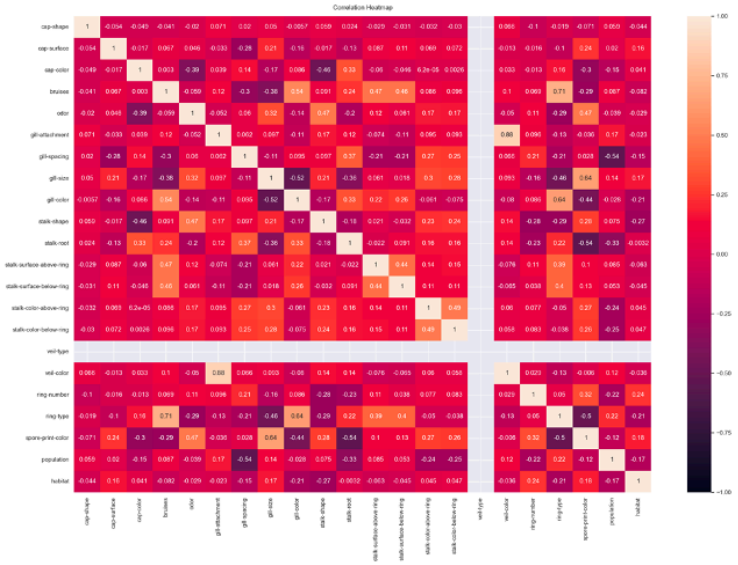

Figure 2: Heatmap showing the correlation of each feature across the data set.

After looking at the map I decided to drop ‘Gill Attachment, Veil Type and Bruises’ from the data. Data that is highly correlated could imply that it is giving the model similar information and including both features could lead to overfitting. It did not appear to make a large difference to the accuracy scores however, I decided to leave these features out of the model.

The next step I took was to look at the precision and recall scores alongside a confusion matrix across all the features. The accuracy score came in at 100% with precision and recall scores of 1. These scores show the accuracy matrix of the positive predictions (precision score) and the actual positives (recall score). These scores suggest the model is very good. Still, I want to make sure this is true and not something strange happening within the data.

Further sanity tests:
I used a counter to ensure I was getting a good mix of data in the test and train sets, and they look good.

Class distribution in training set:
Poisonous: 2731 samples
Edible: 2869 samples
Class distribution in test set:
Poisonous: 1167 samples
Edible: 1233 samples

There is too much data to go through every single feature combination (that would be around 2.082e+31 combinations) so instead I calculated the individual feature accuracy.

Odour has 98% accuracy at a max depth of 2 – madness! Let’s look at the tree:

Figure 3: Decision Tree with one Feature, 'Odour’, and a max depth of 2.

The model reveals an early split based on odour, implying that a reliable odour rating alone might suffice without the need for a complicated model. This insight might have remained undiscovered if the model were constructed through a black box machine learning approach. The decision tree, by showcasing the choices being made, underscores its power in providing valuable insights[3].

I dove in a little further and had a look at the feature importance on the model that included all the features and found that I was right in thinking not every feature would be needed! In fact, I could remove 7 of them without changing the model at all.

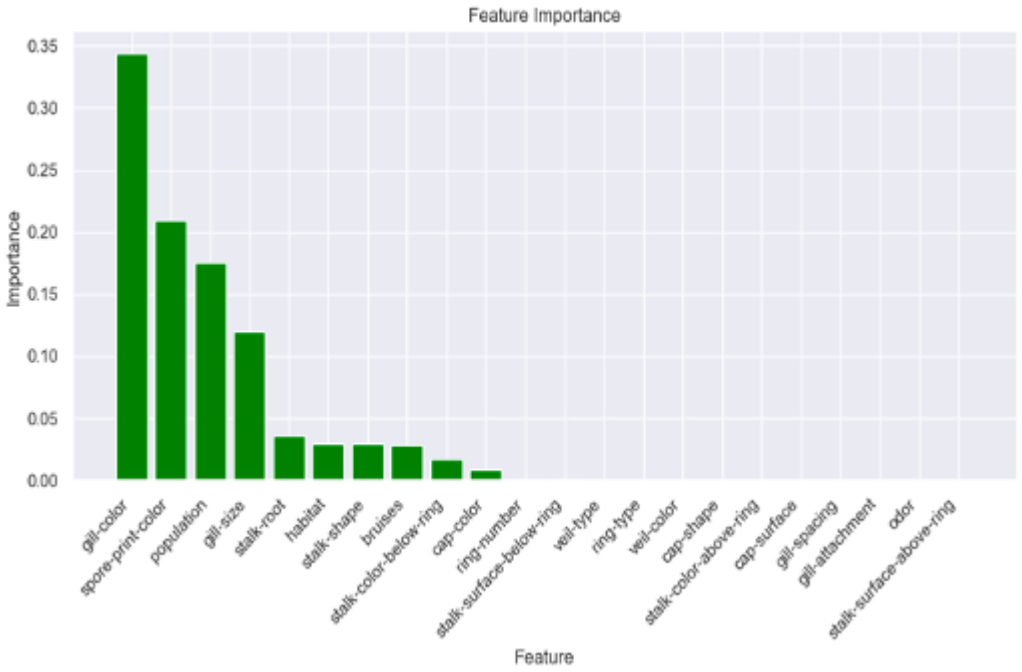

Figure 4: Feature Importance on Decision Tree Model.

It is also interesting to note how Odour is highlighted as having very low importance on the model. The features accuracy is measured on how well it predicts the target variable directly whereas the importance of the feature is calculated by how much the feature contributes to the overall reduction in Gini impurity, the bigger the reduction the more important the feature[3]. The Gini impurity is a number that represents the likelihood of random data being misclassified[4]. By calculating the accuracy of the features, it can help streamline the model complexity and help uncover patterns and relationships in the data[5]. This suggests that the 'Odour' feature is not as important as the interactions between the other features and that its contribution to the overall model is less significant. It could also be due to decision trees being 'greedy' algorithms[6], and the first best step does not include odour causing it to subsequently be disregarded throughout.

Reflection

This exercise has shown me the intuitive processes needed to apply data science techniques in a 'real-world' application. Exploring the idea of using Decision Trees as a potential tool for mushroom classification showed me the power of Decision Trees in such that they can reveal patterns within the data that may not be discovered with black box machine learning. For a task such as mushroom classification I feel understanding what is happening within the model is important as you may not always have viable readings for every single feature and so understanding how the model prioritises these allows users to decide how much they can trust the model.

The ethics to think about behind this predictive tool and its use could relate to public safety well-being. If this model were to be used outside of academic exploration and implemented in real-life, it should always be used alongside an expert. Machine Learning and its uses can be hugely beneficial however their results should never go unquestioned.

In conclusion, this exercise underscored the pivotal importance of balancing model complexity with interpretability in the field of data science. Through exploring Decision Trees as a tool for mushroom classification, I gained insights into their power to unveil hidden patterns, which is often not possible with black box machine learning models. Understanding the inner workings of the model proved essential, especially when dealing with incomplete or unreliable data. Ethically, it's important to acknowledge the limitations of the model and encourage expert involvement in real-world applications. There are huge potential risks involved in solely relying on a machine learning model for critical decisions.

---

[1] Stack Overflow. (n.d.). *Pandas: convert categories to numbers*. [online] Available at: https://stackoverflow.com/questions/38088652/pandas-convert-categories-to-numbers [Accessed 1 Mar. 2024].

[2] Lark, R. (2023) *How to identify poisonous mushrooms*, *Environment Co*. Available at: https://environment.co/how-to-identify-poisonous-mushrooms/ (Accessed: 07 March 2024).

[3] Krisalay (2024) *Feature importance in decision tree*, *Medium*. Available at: https://medium.com/@krisalay/feature-importance-in-decision-tree-8e60f2174717 (Accessed: 07 March 2024).

[4] Author: Fatih Karabiber Ph.D. in Computer Engineering and Fatih Karabiber Ph.D. in Computer Engineering (no date) *Gini impurity*, *Learn Data Science - Tutorials, Books, Courses, and More*. Available at: https://www.learndatasci.com/glossary/gini-impurity/ (Accessed: 07 March 2024).

[5] Krisalay (2024) Feature importance in decision tree, Medium. Available at: https://medium.com/@krisalay/feature-importance-in-decision-tree-8e60f2174717 (Accessed: 07 March 2024).

[6] Grus, J. (2019) *Data Science from scratch: First principles with python*. Sebastopol, CA: O'Reilly Media.