Poisonous mushroom prediction based on observable attributes

CS982 Big Data Technologies Coursework

Computer and Information Sciences,
Livingstone Tower,
University of Strathclyde,
26 Richmond Street,
Glasgow G1 1XH

November 5th, 2024

Contents

List of figures

List of tables

Chapter 1

Introduction

Mushrooms are considered significant contributors to human nutrition and health. *Agaricus bisporus*, which is commonly known as the white button mushroom, is widely purchased in the market, accounting for 40% of total mushroom production worldwide (Shivaghami Shamugam and Kertesz, 2022).

From a nutritional perspective, mushrooms are exceptionally valuable, containing 19-35% protein on a dry-weight basis and providing all essential amino acids (Cheung, 2010). According to Kimatu et al. (2017), mushrooms contain various bioactive compounds, including polysaccharides, proteins, and secondary metabolites with significant anti-tumour properties.

On the flipside however, several species of mushrooms are known to be poisonous. *Agaricus xanthodermus* is known to cause gastrointestinal pains and various other symptoms if ingested (Özaltun & Sevindik, 2020). Some *Amanita* species are deadly poisonous, and are often mistaken for other edible mushroom species by mushroom hunters. A study by Schenk-Jaeger et al. (2012) investigated 32 amatoxin reports between 1995 and 2009 and found that 30 of them were due to accidental consumption.

Given the prevalence of mushroom-picking as a pastime in various parts of the world, it is important that the accidental risks associated with this activity be mitigated. This report analyses the UC Irvine (UCI) Mushroom dataset (Mushroom, 1981) to investigate relationships between mushroom attributes and its toxicity. Thereafter, it aims to train supervised and unsupervised models capable of properly identifying poisonous mushrooms based on its attributes, and critically evaluate the effectiveness of these models.

Chapter 2

Mushroom dataset

2.1 Dataset

The Mushroom dataset is a publicly available dataset obtained from the UCI Machine Learning Repository. It is a collection of 8,124 gilled mushroom samples from the *Agaricus* and *Lepiota* families. Each sample is characterised by 22 categorical attributes, including cap features (shape, surface, colour), stem characteristics (shape, root, surface, colour), gill properties (attachment, spacing, size, colour), and other traits such as bruises, odour, veil type, ring features, and spore print colour. The mushroom attributes are summarised in Table 2.1.

Attribute Category	Attributes
Cap Features	Shape, Surface, Colour
Stem Features	Shape, Root, Surface (above and below ring), Colour (Above and below ring)
Gill Properties	Attachment, Spacing, Size, Colour
Other Morphological Traits	Bruise, Odour, Veil Type, Veil Colour, Ring Type, Ring Number
Reproductive Features	Spore Print Colour
Ecological Factors	Population, Habitat

Table 2.1: Summary of mushroom attributes in the Mushroom dataset

The dataset was imported as a two-dimensional Pandas dataset, with the first column being the target class (‘e’ for edible and ‘p’ for poisonous) while the remaining columns are made up of the 22 categorical features. Each category is represented by a single alphabet, with the exception of the missing-value stalk-root category denoted by a ‘?’ symbol. Each of the 8,124 rows represents a single mushroom sample and its corresponding features.

This Mushroom dataset was ideal for our report as it contains a rich set of observable features, which is particularly relevant to mushroom foragers who pay attention to these physical characteristics in the wild.

2.2 Exploratory Data Analysis

Class distribution between edible and poisonous mushrooms is relatively balanced (around 51.8% edible, 48.2% poisonous).

A major observation is that many features (mainly those pertaining to colours and shapes) contain multiple categories that make them good candidates for one-hot encoding. As for other features (such as 'gill-spacing' and 'population'), an ordered relationship may exist among the categories and hence are better suited for ordinal encoding.

A custom ordinal encoder was applied to the 'bruises?', 'gill-spacing', 'gill-size', 'ring-number' and 'population' columns to preserve the hierarchical relationships between the categories. For example, mushrooms with an 'abundant' population feature were encoded with a high value of '5', while 'solitary' mushrooms were encoded with a low value of '0'. This also allows a correlation heatmap (Figure 2.1) to be plotted among these features and the toxicity of the mushrooms. Some notable correlations include the presence of bruises and the gill size of the mushrooms that exhibit moderately negative correlations with toxicity.

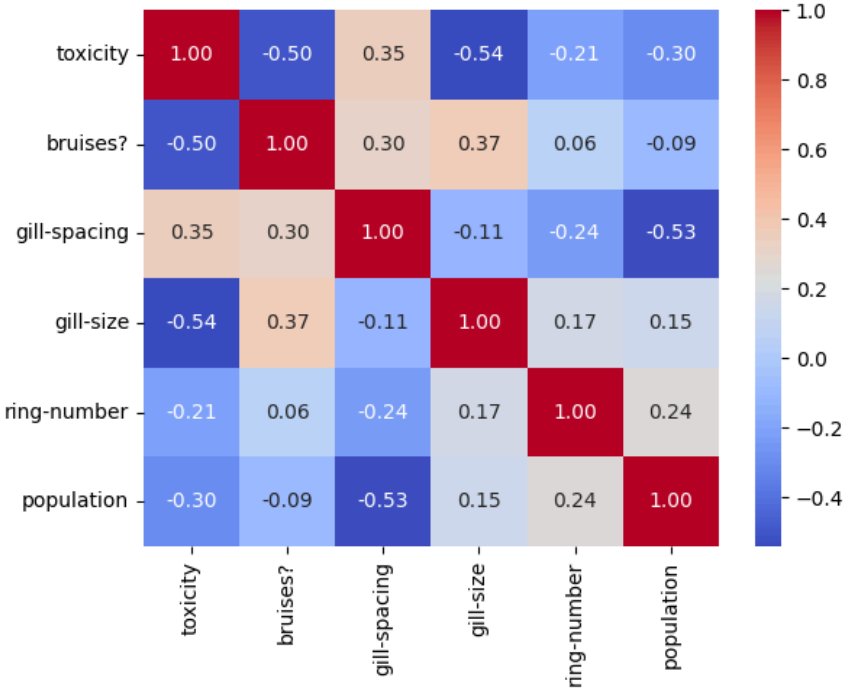

Figure 2.1: Correlation heatmap of toxicity and ordinally-encoded features

Count plots were used to investigate the relationships between each categorical feature and toxicity. Odour stood out as a feature with categories that almost exclusively contained either edible or poisonous mushrooms (Figure 2.2). This suggests that decision tree algorithms may

perform well with such features, as there would be very high purity within each tree node after a split.

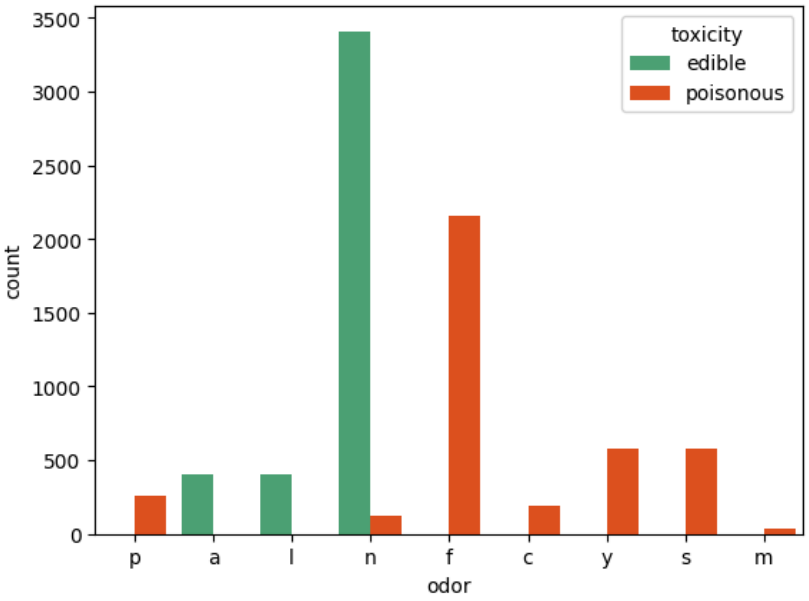

Figure 2.2: Count plot of various odours by toxicity

Similarly, Figure 2.3 shows that mushrooms with a buff ('b') gill colour are almost always poisonous.

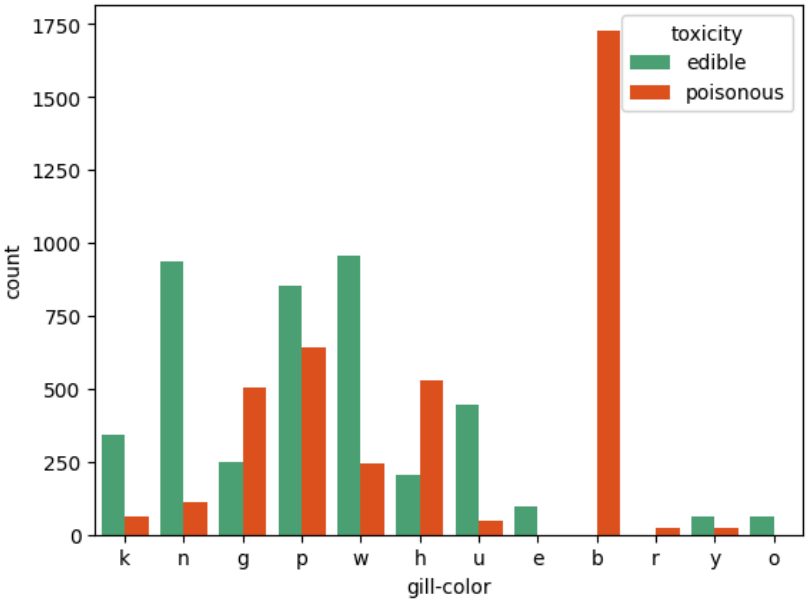

Figure 2.3: Count plot of various gill colours by toxicity

Likewise, Figure 2.4 reveals several distinct correlations between certain colours and toxicity.

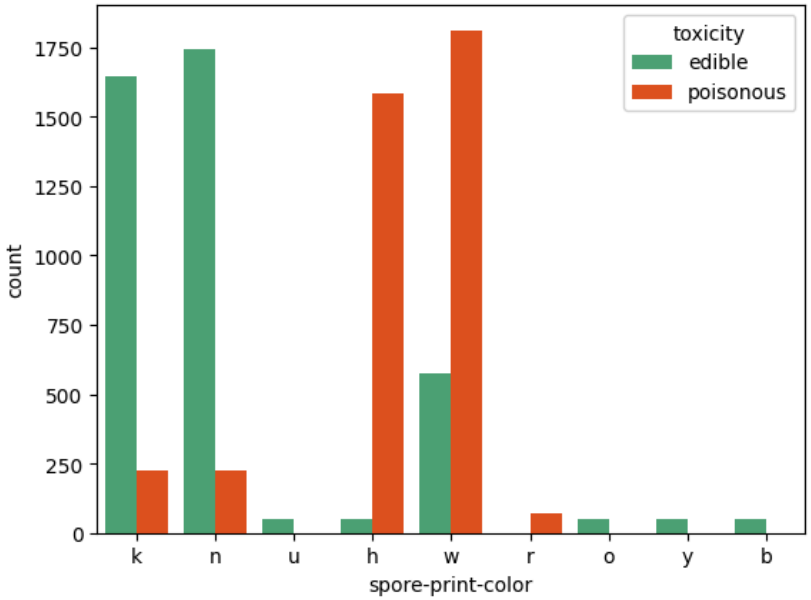

Figure 2.4: Count plot of various spore print colours by toxicity

Finally, correlations were investigated among multiple categorical features through the use of parallel set plots (Figure 2.5). The broader the lines between the different features, the stronger the connection. It is observed that most categories in 'odor' and 'spore-print-color' can classify the mushrooms into edible or poisonous with a high degree of certainty, while other features like 'veil-color' lack categories with the same discernibility.

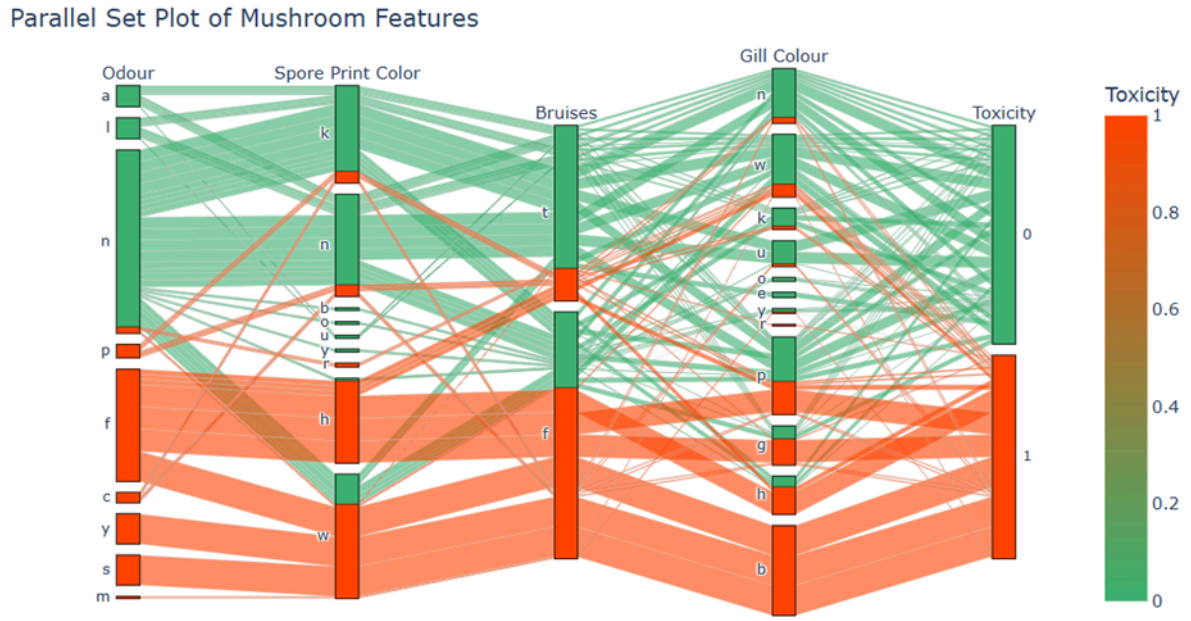

Figure 2.5: Parallel set plot

Chapter 3

Unsupervised Analysis

Unsupervised learning methods are used to identify patterns in datasets in the absence of labelled data. In the case of the Mushroom dataset, all observations have already been labelled ‘poisonous’ or ‘edible’. Ostensibly, it would be more appropriate to apply supervised learning methods that can deliver more accurate predictions based on ground-truth labels. Nevertheless, unsupervised learning methods may be useful in highlighting hidden relationships in the data, drawing their own conclusions from raw data without being influenced by assigned class labels. As such, the true label column was dropped from the dataset before training, and only used for comparison and evaluation purposes.

3.1 Application

As the dataset is categorical and comprises lettered values, the data is first pre-processed to represent each category using numerical values. This is done through ordinal encoding and one-hot encoding.

We experimented with two unsupervised learning methods – agglomerative hierarchical clustering and K-Means clustering. Agglomerative clustering was tested using two different distance measures – Euclidean and Hamming. As for K-Means clustering, the dataset was pre-processed by scaling the features to ensure that no one feature dominates the rest. As there were 107 features in the one-hot encoded dataset, Principal Component Analysis (PCA) was performed to reduce the high-dimensional feature space into its principal components.

We observed that the results obtained using agglomerative clustering were somewhat comparable to those obtained using K-Means clustering. Hence, we have selected K-Means clustering for our analysis, taking into consideration the easier interpretability and better scaling.

3.2 Results

The quality of clustering was evaluated using the following four Scikit-Learn metrics: Silhouette Score, Calinski-Harabasz Index, homogeneity, and completeness. The Silhouette score and Calinski-Harabasz index both measure the compactness within and separation

between different clusters, while homogeneity and completeness measure how well the clusters match the ground-truth labels. The results obtained through K-Means clustering are tabulated in Table 3.1 below.

	Before PCA	After PCA
Silhouette Score	0.10	0.59
Calinski-Harabasz Index	720.96	6932.83
Homogeneity	0.59	0.57
Completeness	0.61	0.60

Table 3.1: K-Means Clustering before and after performing PCA

3.3 Findings

Following the use of PCA, it can be observed that the Silhouette score for K-Means clustering increased from 0.10 to 0.59, while the Calinski-Harabasz index increased from 720.96 to 6,932.83. The substantial increases in both suggest that PCA has helped to separate the data into more distinct and compact clusters.

However, the use of PCA does not come without its drawbacks, as observed from the drop in homogeneity and completeness by 1% and 2% respectively. This could be attributed to the loss of information in feature-compression, whereby only the largest variances in the data were retained. In spite of this, the loss of information is minimal and far outweighed by the significant improvement in clustering quality.

Reducing the feature space into two principal components has enabled the clustering to be visualised using a 2-dimensional scatter plot. Figure 3.1 and Figure 3.2 compare the results of the labels discovered through K-Means clustering against the ground-truth labels. On the whole, the K-Means-assigned clusters appear visually similar to the true-labelled clusters, with exceptions mainly noted in the bottom-centre region of the scatterplot.

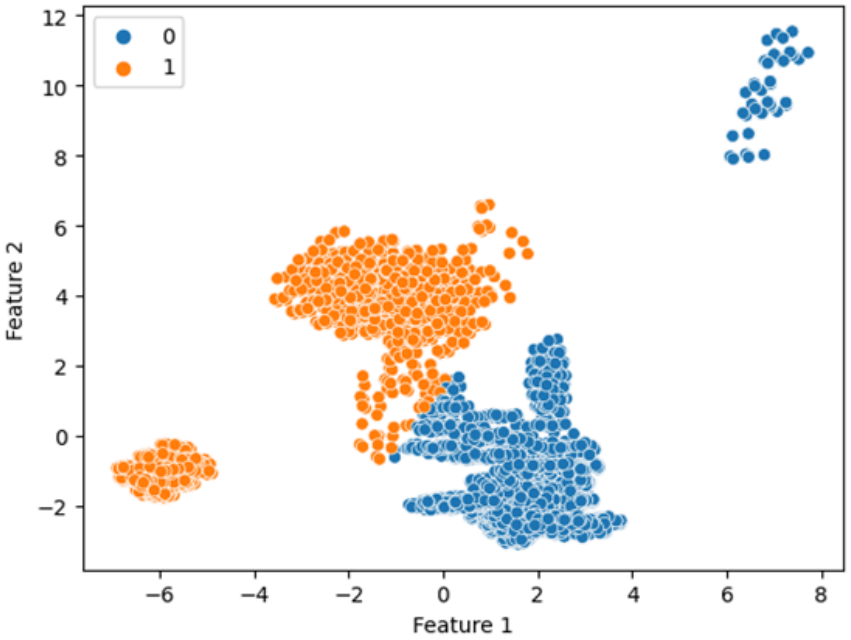

Figure 3.1: Scatterplot based on K-Means clustering

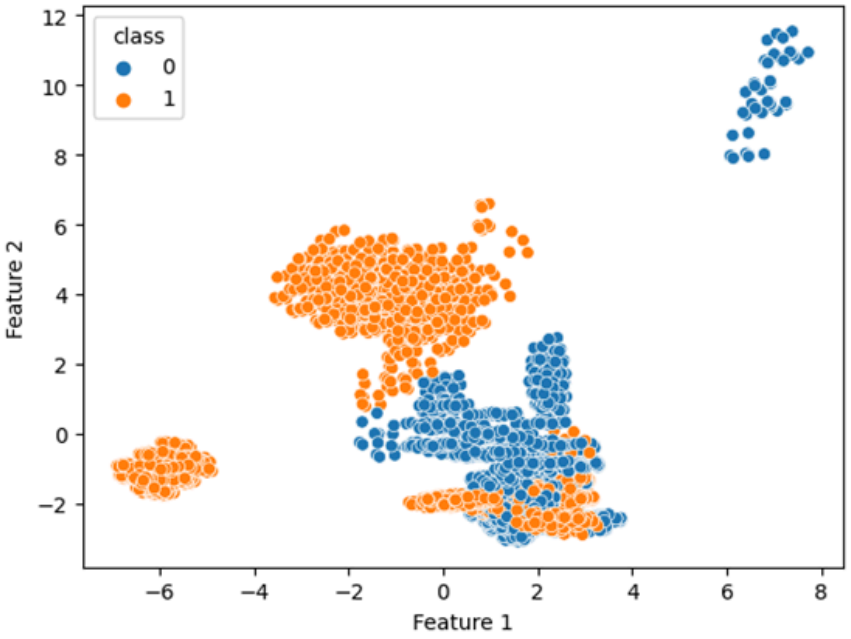

Figure 3.2: Scatterplot based on ground-truth labels

It is worth noting that these data points have been misclassified as ‘edible’ when they are actually ‘poisonous’, which may pose serious health ramifications if the results of this clustering technique were used in a real-world mushroom classification problem.

Finally, the features were ranked by their contribution to each principal component (PC). Based on the results in Table 3.2 and Table 3.3, no single feature strongly contributed to the variance of either principal component, with the biggest contributions being only 11.9% and 12.7% within PC1 and PC2 respectively.

Feature	Contribution
stalk-surface-above-ring	0.119042
stalk-surface-below-ring	0.114969
ring-type	0.112940
odor	0.104064
spore-print-color	0.099790

Table 3.2: Top Features Contributing to PC1

Feature	Contribution
stalk-root	0.127291
spore-print-color	0.117306
gill-color	0.111140
habitat	0.073949
ring-type	0.069528

Table 3.3: Top Features Contributing to PC2

The overlap in the clusters shown in Figure 3.2, coupled with the mediocre initial results of K-Means clustering before PCA in Table 3.1, suggests that the data may not contain intrinsic patterns that clearly distinguish poisonous mushrooms from edible ones. Moreover, since no single feature strongly contributed to the separation of the clusters, the clustering result is not easily interpretable. Thus, K-Means clustering is not an appropriate method to accurately identify poisonous mushrooms.

Chapter 4

Supervised analysis

Supervised learning functions through the use of labelled data. In the case of the Mushroom dataset, the input data are the physical attributes with the output being the class. A supervised learning model aims to understand the relationship between the inputs and outputs to make accurate classifications of new or unseen data.

4.1 Application

A decision tree classifies data through creating a set of rules based on the features provided. Starting at a 'root' node and working its way down, these rules separate each node based on the features that best separate the classes. These nodes continue to split, forming a tree-like structure until it reaches the leaf nodes in which a prediction can finally be made.

Other algorithms, including logistic regression, K-nearest neighbours, random forest, and Naïve Bayes, were also tested on the dataset. All methods, except Naïve Bayes, provided optimal classification results. The decision tree model was chosen due to its ability to visually represent the classification process and identify the most important features contributing to edibility. In the context of this dataset, these attributes provide a high level of transparency and insight that other methods cannot replicate.

To minimise overfitting, a grid search cross-validation is applied, which controls the maximum depth of the tree and also sets the minimum number of samples needed to split a node. Performance metrics such as accuracy, precision, recall and F1 score are also used to evaluate the best model. After the optimum model is identified, it is applied to the test split for evaluation. In this case, a train-test split of 70/30 has been chosen. The performance metrics, along with a confusion matrix, are finally determined.

4.2 Results

Figure 4.1 displays the most optimal decision tree found by the algorithm. Evaluation metrics indicated are displayed in Table 4.1, with the final tree having the most optimum parameters of a maximum depth of 10 and a minimum sample split of 2. Figure 4.2 displays the confusion matrix created from the algorithm.

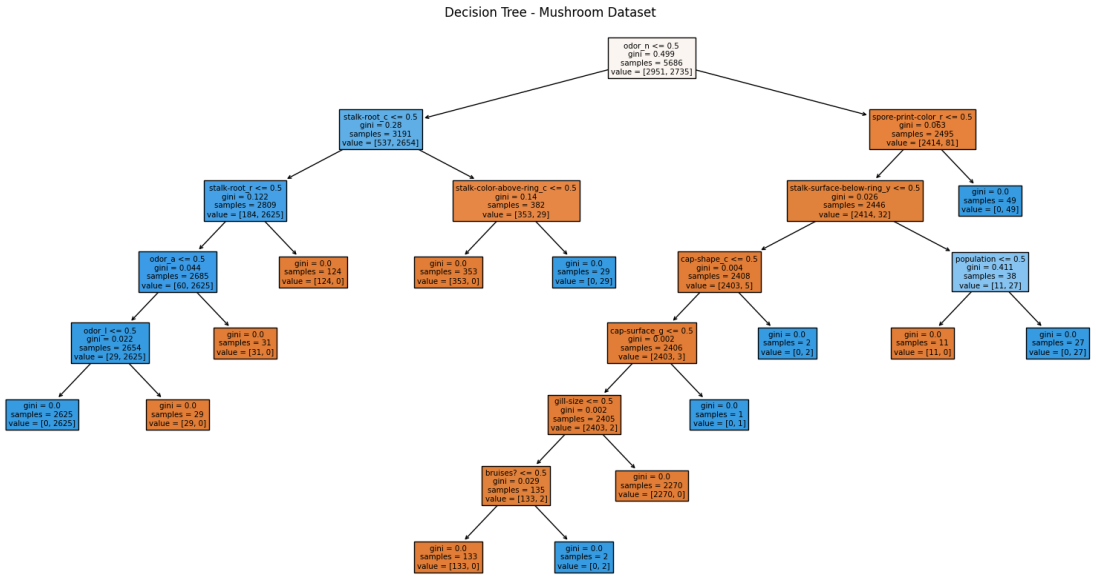

Figure 4.1: Plot of the most optimum decision tree

Performance Metric	Score
Accuracy	100%
Precision	100%
Recall	100%
F1 score	100%

Table 4.1: Evaluation metrics obtained from the decision tree classifier.

$$\begin{bmatrix} 1257 & 0 \\ 0 & 1181 \end{bmatrix}$$

Figure 4.2: Confusion matrix obtained from the decision tree classifier algorithm.

Figure 4.3 further shows a graph ranking the most important features for determining the mushroom's edibility, with the results finding that the most optimal feature is its odour. Table 4.2 further specifies the importance of each feature.

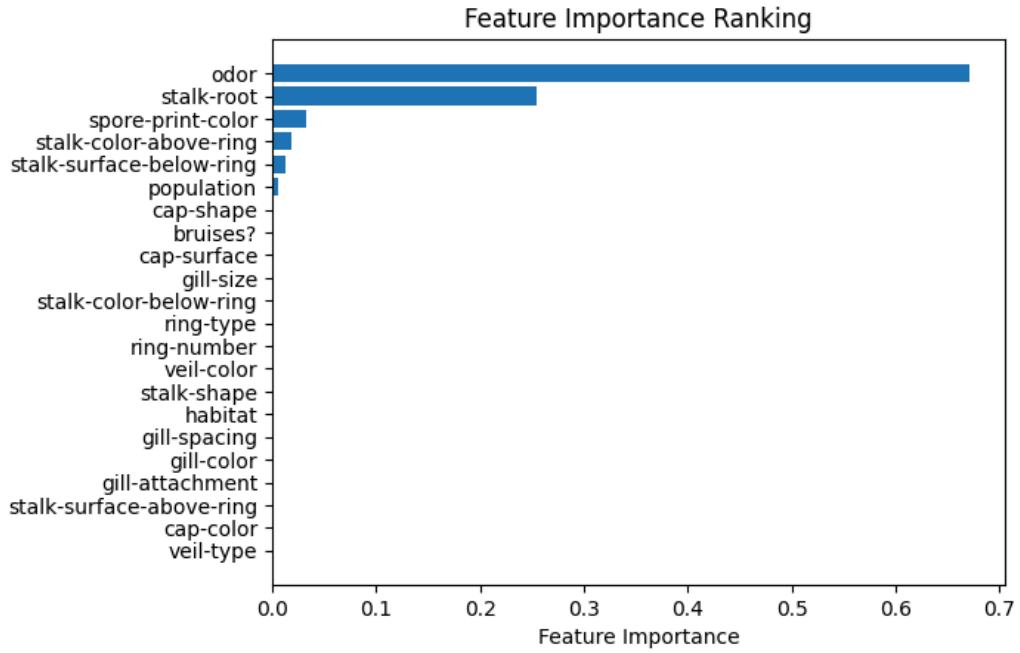

Figure 4.3: Horizontal bar plot ranking features by contribution to the edibility of the mushrooms

Feature	Feature Importance
odor	0.671463
stalk-root	0.254446
spore-print-color	0.032963
stalk-color-above-ring	0.018879
stalk-surface-below-ring	0.013228

Table 4.2: Top five most important features

4.3 Findings

The decision tree model yielded a highly accurate and reliable prediction of mushroom classification, achieving an optimal performance across all metrics. It was found that the optimal hyperparameters were a maximum depth and minimum sample split of 10 and 2 respectively shown in Figure 4.1. When applied to the test data set, the model achieved accuracy, precision, recall, and F1 scores of 100% as shown from Table 4.1.

The analysis also revealed that certain features bore significant contribution towards the model predictions. As shown in Figure 4.3 and Table 4.2, odour emerged as the most

influential feature, contributing to 67% of the model's decision-making process, followed by stalk-root (25%), spore-print-colour (3.3%), stalk-colour-above-ring (1.9%) and stalk-surface-below-ring (1.3%). Further features showed a negligible influence on the model, indicating that these primary attributes are the most influential in differentiating between edible and poisonous mushrooms. The feature prioritisation is further backed up by Figure 2.2, showing consistent classifications with specific odours like almond, anise, and no scent.

Finally, the confusion matrix displayed in Figure 4.2 confirms that all mushrooms were correctly classified, showing no false positives or false negatives. This further validates the effectiveness of the decision tree model at distinguishing between classes. These findings highlight the capability of the decision tree model in identifying the classification of mushrooms based on some primary characteristics, especially odour. The high accuracy, prediction and feature interpretability make this a highly effective tool for identifying mushroom edibility.

Chapter 5

Reflections

The K-Means clustering method showed significant improvement in clustering quality after applying PCA dimensionality reduction. However, when compared to the ground-truth labels, the clusters only attained mediocre homogeneity and completeness scores, highlighting the limitations of unsupervised approaches for critical classification tasks.

In contrast, the decision tree classifier achieved perfect accuracy using its optimal hyperparameters. Despite the empirical success of the chosen supervised method, it is not without its limitations. Decision trees are known to overfit the data they are trained on, which means the prediction model fails to generalise to unseen data. This is compounded by the limited data variability of the Mushroom dataset, as 8,124 observations can be considered quite small, and the samples were obtained from only two mushroom families out of many others. Furthermore, many values are missing from its second-most important feature, stalk-root. As a result, even a few errors can get extrapolated if the model were relied upon at scale, which poses dire real-world consequences if people mistakenly ingest mushrooms misclassified as edible.

In a serious mushroom classification context, we would strongly recommend using ensemble tree models, such as random forest, which are less prone to overfitting. In addition, a much larger, more representative and complete dataset should be obtained to train the model.

Chapter 6

Conclusion

The results of this investigation have demonstrated the remarkable ability of machine learning techniques in uncovering hidden patterns and relationships in data.

Although the unsupervised learning method did not provide much utility in addressing this problem, the inherent structures revealed through the K-Means clusters may be valuable for educational and research purposes.

While the results of the supervised model were highly promising, the prediction model should not replace, but only complement, more stringent methods of testing and verification. Given the critical nature of keeping false-negative errors to an absolute minimum, we recommend that mushrooms identified by the model as edible should be double-checked and approved by certified experts as safe for consumption.

Future research may focus on expanding the dataset to more mushroom samples, thereby ensuring broader applicability and reliability in practical settings such as food safety and mycological research.

Needless to say, there's still *much-room* for exploration.

Appendix A

Environment

Language: Python 3.12.4

IDE: Jupyter Notebook 7.0.8

Bibliography

Cheung, P.C.K. (2010) 'The nutritional and health benefits of mushrooms', *Nutrition Bulletin*, 35(4), pp. 292–299. https://doi.org/10.1111/j.1467-3010.2010.01859.x.

Kimatu, B.M. et al (2017) 'Antioxidant potential of edible mushroom (Agaricus bisporus) protein hydrolysates and their ultrafiltration fractions', *Food Chemistry*, 230, pp. 58–67. https://doi.org/10.1016/j.foodchem.2017.03.030.

Özaltun, B. and Sevindik, M. (2020) 'Evaluation of the effects on atherosclerosis and antioxidant and antimicrobial activities of Agaricus xanthodermus poisonous mushroom', *The European Research Journal*, 6(6), pp. 539–544. https://doi.org/10.18621/eurj.524149.

Schenk-Jaeger, K. M. et al. (2012) 'Mushroom poisoning: A study on circumstances of exposure and patterns of toxicity', *European Journal of Internal Medicine*, 23(4), pp. e85–e91. https://doi.org/10.1016/j.ejim.2012.03.014.

Shivaghami Shamugam and Kertesz, M.A. (2022) 'Bacterial interactions with the mycelium of the cultivated edible mushrooms Agaricus bisporus and Pleurotus ostreatu', *Journal of Applied Microbiology*, 134(1), pp.1-10. https://doi.org/10.1093/jambio/lxac018.

UCI Machine Learning Repository (1987) Mushroom Data Set. Available at: https://archive.ics.uci.edu/dataset/73/mushroom (Accessed: 19 October 2024).