DS3002 Data Mining
Project Report
To Shroom or Not To Shroom

Aliyan Ahmed

AbdurRehman Haroon

Eesha Tariq

Javeria Shahid

1. Introduction

I. Abstract

Mushrooms appear in civilization history for their use as food, medicine, or spiritual and religious aids [1] dating as far back in time as the era of Vikings. The Egyptians regarded mushrooms as "food of the gods" [2], the Greeks, Maya, and Aztecs consumed hallucinogenic mushrooms for a myriad of purposes while others believed them to be a source of enlightenment or witchcraft. In modern times, mushrooms, wild, farmed, or commercially grown have multifaceted uses that range from culinary to pharmaceutical. The outcome of which is a high demand. Before they are to be consumed, a viral step is to determine whether it is safe to do so. Another scenario of ingestion is by mushroom hunters as a psychoactive drug. [3]. However, it is due to the lack of knowledge that renders navigating the vast array of mushroom species to distinguish between edible and poisonous ones, a challenge. According to the literature, it is possible to infer the edibility of a mushroom based on its physical traits including but not limited to gill, stem, spore, and cap properties [4]. Although this method is useful, it is by no means infallible as existing anomalies challenge its reliability.

Keywords: *Mushroom Poisoning, Edibility, Toxicity*

II. Motivation

Also known as *mycetism*, mushroom poisoning has remained a significant health concern worldwide. The symptoms of mushroom poisoning can develop as early as six hours post-ingestion and often involve gastrointestinal issues, such as nausea, stomach cramps, vomiting, diarrhea, and abdominal pain. In severe cases, it may cause injury to one's organs [5]. In western Iran, in the 4-year period starting from 2014 to 2018, a great number of 193 patients were admitted to the Poisoning Center of Imam Khomeini Hospital in Kermanshah, on the record of mushroom poisoning. Among the 193, 92.6% were people residing in cities. [6]. Therefore, it is of utmost priority that mushrooms are identified for their edibility before consumption for any purpose. In Pakistan, fifty six edible species have been reported to date [7] compared to hundreds of unidentified or toxic species. Therefore, the motivation behind this project is to investigate the correlation between the physical characteristics of mushroom species and their edibility, with the aim of shedding light on the relationship between a mushroom's physical traits and its potential toxicity or safety for consumption.

III. Dataset Description

The dataset contains 61069 instances of raw data. Each record contains the feature information of mushrooms with caps and their visually discernable traits that may contribute to the decision of whether

they are safe for human consumption represented in the attribute ‘class’ as poisonous(p) and edible(e)

In total, 20 attributes are used to describe each record: 17 nominal and 3 metrical.

The dataset from UCIML repository, Secondary Mushroom [8], is initially explored for the distribution of the target class based on features including cap diameter, shape, surface, color, bruising, gill features, stem attributes, veil properties, ring presence, spore print color, habitat, and season.

cap-diameter	cap-shape	cap-surface	cap-color	….	target
15.26	x	g	o	….	p
16.6	x	g	o	….	p
14.07	x	g	o	….	p
14.17	f	h	e	….	p
14.64	x	h	o	….	p
15.34	x	g	o	….	p
14.85	f	h	o	….	p
14.86	x	h	e	….	p

This is a representation of a portion of the dataset, not the dataset entirely.

IV. Edibility Classification

Using the dataset, our task is to predict whether a mushroom is likely to be edible or poisonous. The evaluation of the model is based on classification metrics including accuracy, precision, recall, and F-score. In the case of our project, recall takes precedence as our primary goal is to prevent mislabeling any poisonous mushroom as edible, to avoid potential accidents.

V. Analysis of Data Encoding

To determine the best encoding for the given data and task, three kinds of encoding methods including One-Hot, Hashing, and Binary were used and evaluated and the best performing one is chosen to ultimately be used further.

VI. Classifier Evaluation and Comparison

The purpose of this section is to evaluate the performance of each classifier including Logistic Regression, Decision Tree, Random Classifier, Gradient Boosting, Support Vector Machines, K Nearest Neighbors, Gaussian Naïve Bayes, Linear Discriminant Analysis and Quadratic Discriminant Analysis.

VII. Explainable AI

With the purpose of enhancing transparency and building trust in the model’s performance, the use of XAI reveals the reasoning behind predictions including the impact of each feature that aid in model improvement.

VIII. Deep Learning

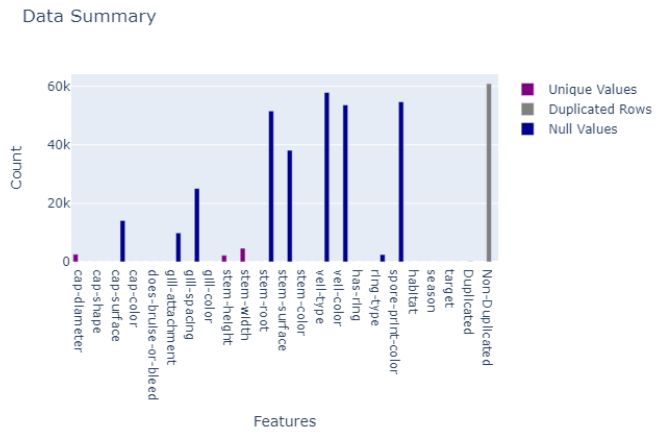
The project implements a neural network model to classify mushrooms as edible or poisonous. The model has two hidden layers with 128 and 64 neurons, respectively, and an output layer with a single neuron using the sigmoid activation function. The model is compiled with binary cross-entropy loss and the Adam optimizer

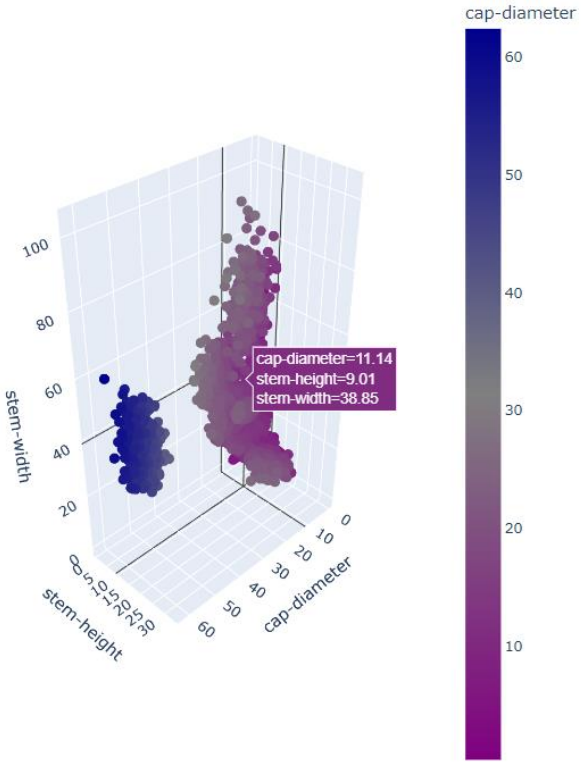
2. Edibility Classification

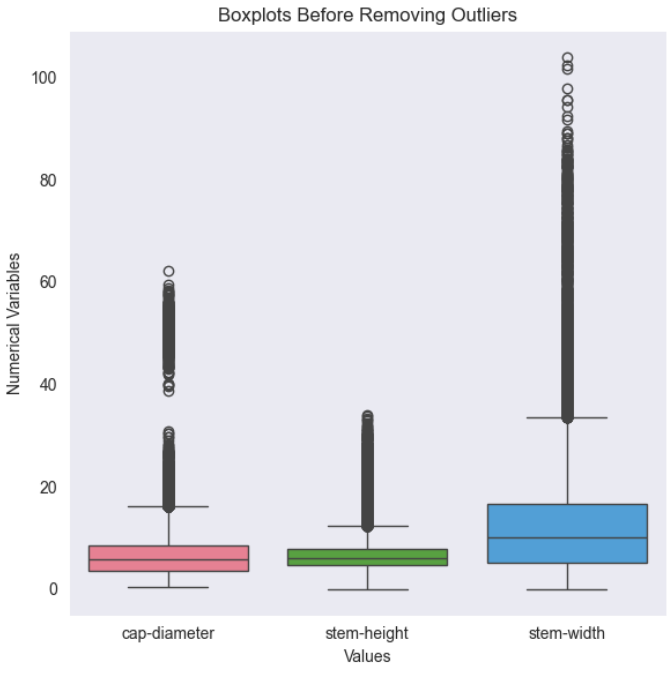
The primary objective of this project is binary classification using supervised machine learning and deep learning.

I. Exploratory Analysis

To understand the structure of the dataset by exploring its statistics, missing values, outliers, duplicate records, and unique values to prepare the data effectively [9]. This is achieved by utilizing Pandas methods in python such as head, info, describe, duplicated and more. The results are plotted in the following figures







1 Boxplots to show outliers in the data

This thorough examination sets the foundation for data cleaning, feature engineering and other preprocessing tasks for a robust and reliable model.

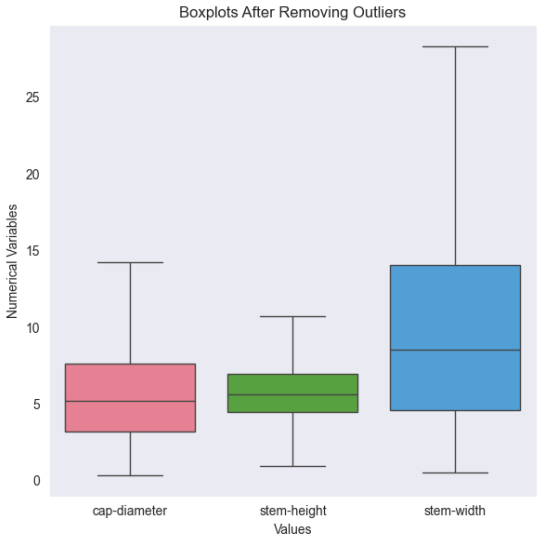
II. Data Imputation

The purpose of data imputation is to replace values that are either missing or likely to be incorrect due to human error (e.g., adding a '?' instead of 'N/A'). These values are initially replaced with NaN and later treated as null values to be filled using the **mean** for the numerical columns *{cap-diameter, stem-height, stem-width}* and **mode** for the categorical columns *{cap-shape, cap-surface, cap-color, does-bruise-or-bleed, gill-attachment, gill-spacing, gill-color, stem-root, stem-surface, stem-color, veil-type, veil-color, has-ring, ring-type, spore-print-color, habitat ,season, target}*

III. Outliers and Duplicates

a. The function to detect and remove **outliers** uses Interquartile Range. The IQR, calculated as the difference between the third and first quartiles, is robust to outliers, making it suitable for outlier detection. Additionally, it offers a simple, non-parametric approach that works well with skewed distributions, providing a clear threshold for identifying outliers. The records containing these outliers are removed from the dataset to improve its quality.

The resulting data is visualized in the following boxplots :



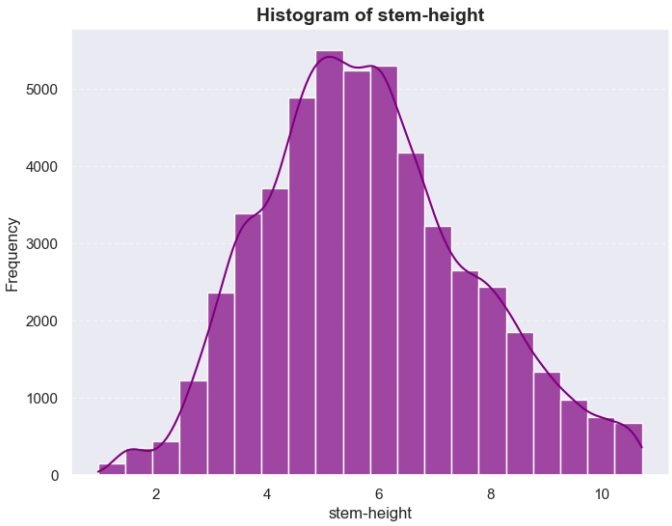
b. Since the presence of **duplicate** records in the data may give way to

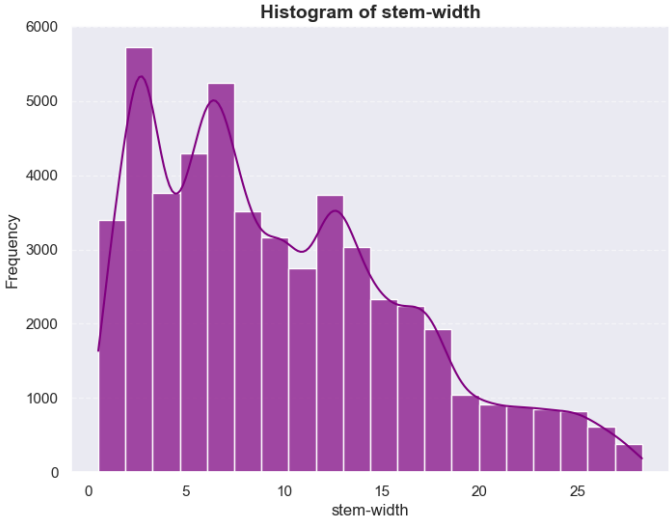
bias in the model, we removed the 146 records out of the 61069 instances in the dataset in question which amounts to <1% of the entire dataset. For this reason, no crucial information is lost, and the size and variation of dataset is largely preserved.

IV. Plots for Visualization

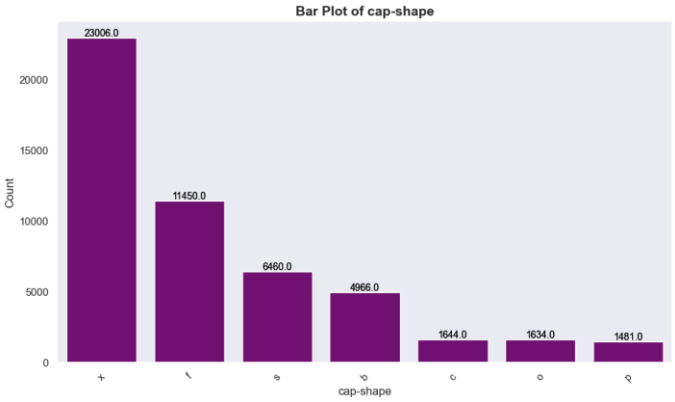
For the purpose of discovering and inferring relationships among the features of the data, they are plotted as histograms, bar charts, correlation matrix, and scatter plot. These figures provide valuable insight such as redundant features and those that provide minimal information along with the presence of any skewness in data that would otherwise go unnoticed and worsen the learning capabilities of the model.

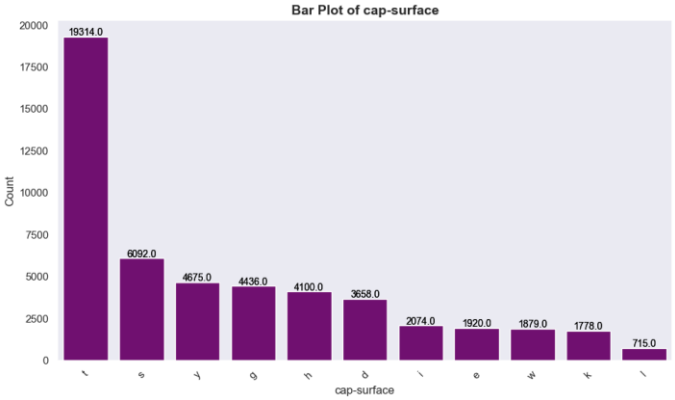
a. Histograms



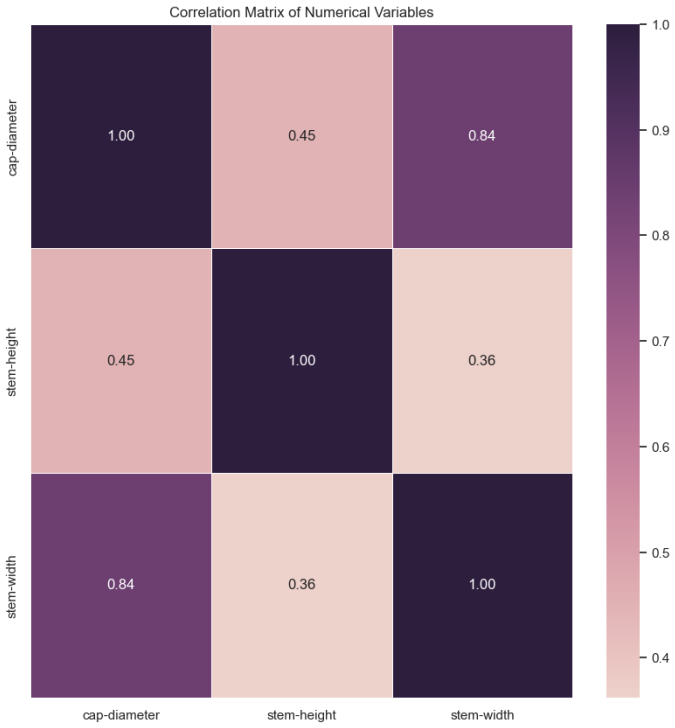


b. Bar Charts

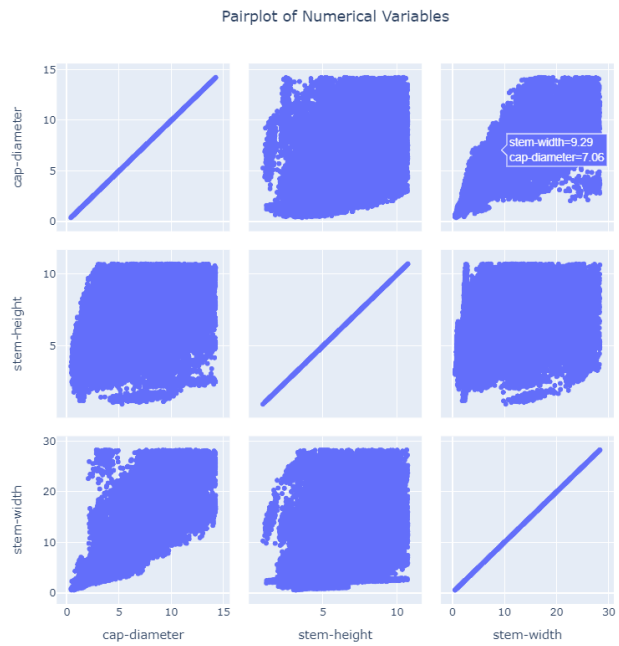




c. Correlation Graph



d. Scatter plot



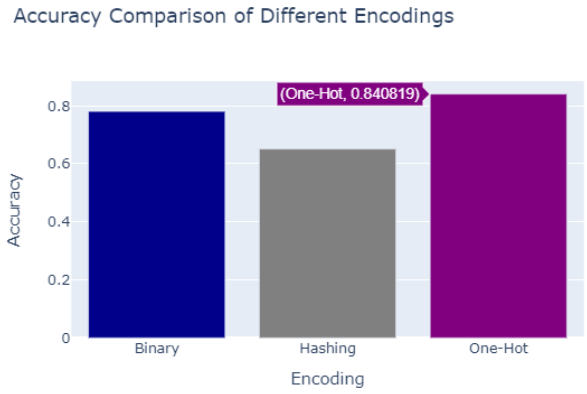
From the above plots, we concluded that there are no features that can be labeled as irrelevant as they are do exhibit properties and patterns that are essential to out model.

V. Class Imbalance

To mitigate the issue of class imbalance, we used RandomOverSampler object of the class imblearn [10] which increases the instances of the minority class (e) by randomly selecting examples with replacement and adding them to the dataset until the class distribution is balanced. By decreasing class imbalance, we prevent the model from being biased towards any majority class and learn better patterns for a more accurate and reliable classification model.

VI. Analysis of Data Encoding

For the categorical features, the three types of encodings used include One-Hot, Binary, and Hashing. All three perform adequately on the chosen dataset. However, in order to create a quality product, we analyze the performance of each encoding method to choose the one that performs better than the rest on a logistic regression model. To further improve this analysis, the model is evaluated for its best parameters for the logistic regression model among **{'C': [0.001, 0.01, 0.1, 0.05] , 'max_iter': [300, 400, 500, 600], 'penalty': ['l1','l2'] , 'solver': ['saga']}** for each encoding. This information is to be used in the implementation of the machine learning models later on. the performance measure for this analysis is accuracy and the comparison results in One-Hot encoding performing better than the other two with the parameters *{'C': 0.1, 'max_iter': 300, 'penalty': 'l2', 'solver': 'saga'}*.



Hashing on the other hand, performed the worst with an accuracy of 60% with the parameters *{'C': 0.1, 'max_iter': 300, 'penalty': 'l2', 'solver': 'saga'}*.

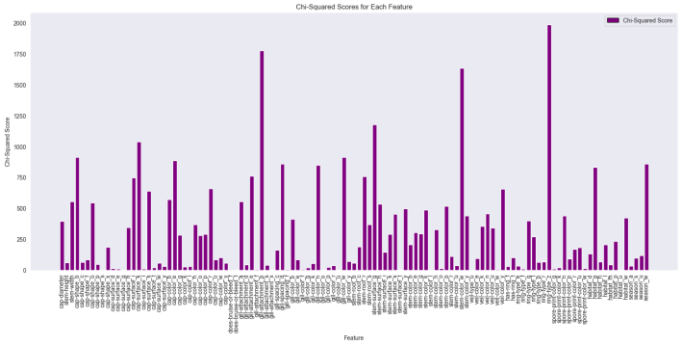
Lastly, Binary encoding resulted in an accuracy percentage of 78 using the parameters *{'C': 0.1, 'max_iter': 300, 'penalty': 'l2', 'solver': 'saga'}*.

Therefore, from this point forth, the dataset used is One-hot encoded.

VII. Feature Extraction

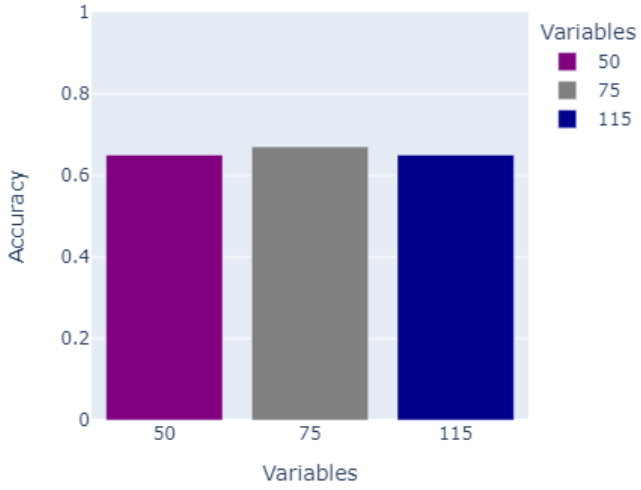
a. **Filter method**: KBest features

By using the feature selection technique, SelectKBest, from scikit-learn library, the most informative features are extracted based on the chi-square test. Initially, all the features are evaluated and the transformed dataset will retain its number of features. The chi-square scores are then visualized in a bar plot as follows



Since most of the features exhibit moderate to high impact, the choice to narrow down the number of features is restricted between 50%-100% of the data.

To find a precise number, k is manually set to [115, 75, 50] to find out the optimal value based on their support score from the SelectKBest library. The results of each are then compared



b. Wrapper method

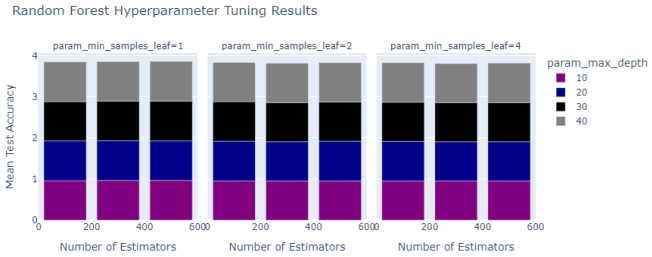
Using Logistic Regression as the estimator, Recursive Feature Elimination (RFE) is implemented. It also employs GridSearchCV to find the most optimal number of features using 3-fold cross-validation. The features selected are ones where the information is concentrated. the 14 best features are evaluated. However, for the purpose of this project, wrapper method not declared reliable and we proceed using the results from KBest method.

VIII. Machine learning

To determine the best hyperparameters, nine different machine learning models are applies:

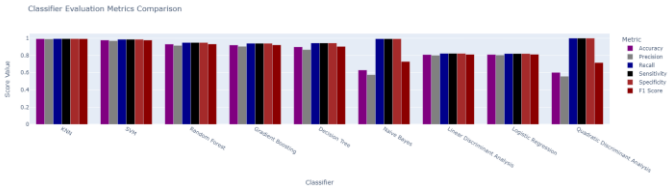
1) Logistic Regression
2) Decision Tree
3) Random Classifier
4) Gradient Boosting
5) Support Vector Machines
6) K Nearest Neighbors
7) Gaussian Naïve Bayes
8) Linear Discriminant Analysis
9) Quadratic Discriminant Analysis.

The results for each one is then analyzed by mean test accuracy. The results for Decision Tree classifer are represented below:



a. Classifier Evaluation and Comparison

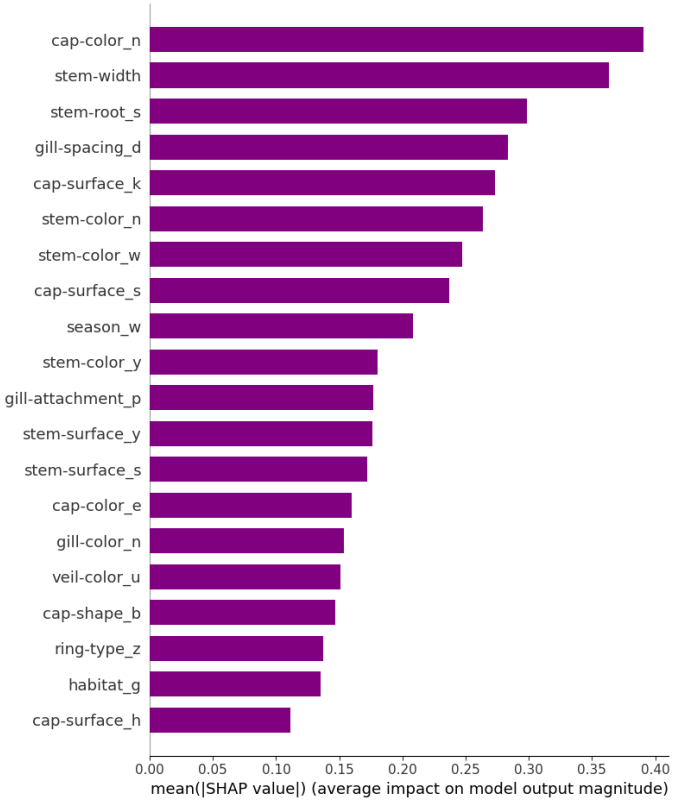
Using the best parameters determined for each model previously, each one is evaluated based on classification metrics(Precision, Recall, Sensitivity, F1 Score, Specificity, Accuracy) and represented in the form of bar charts



IX. Explainable AI

a. SHAP

To implement XAI, SHAP (SHapley Additive exPlanations) is used which is an XAI techniques that assigned each features an importance value for the predictive models [11]. In this section, we use the previously trained machine learning models to interpret the relationships between feature values and their impact on the target variable. For the Logistic Regression model, the results for feature values are represented in the following plot

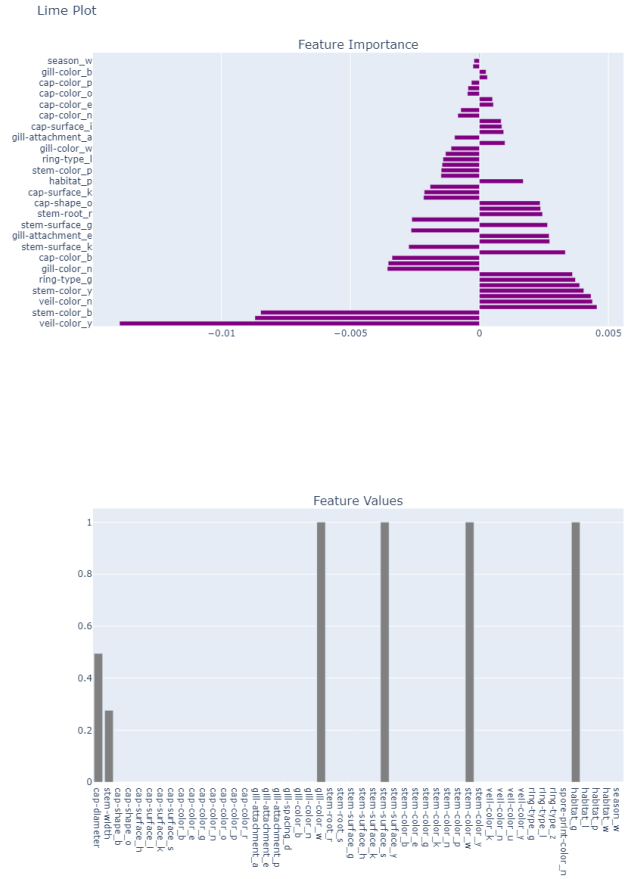


2 feature values for logistic regression

As the value increases, we find that the impact of that feature also increases. From the figure, we find out that cap-color_n is the most impactful feature while cap-surface_h is the of the least importance.

b. LIME

The second method used is LIME (Local Interpretable Model-agnostic Explanations) on each of the previously trained machine learning models. The purpose of using LIME is to interpret individual predictions by approximating the model's behavior around a specific instance. This allows user to understand how and why the prediction was made which ultimately improves the user's trust in the model.



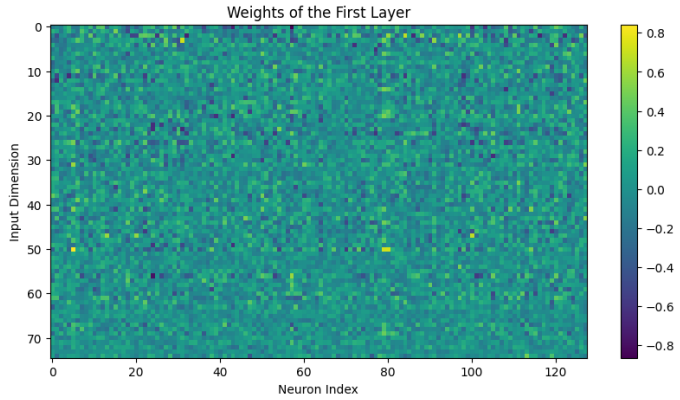
3LIME representation of NB classifier

X. Deep Learning

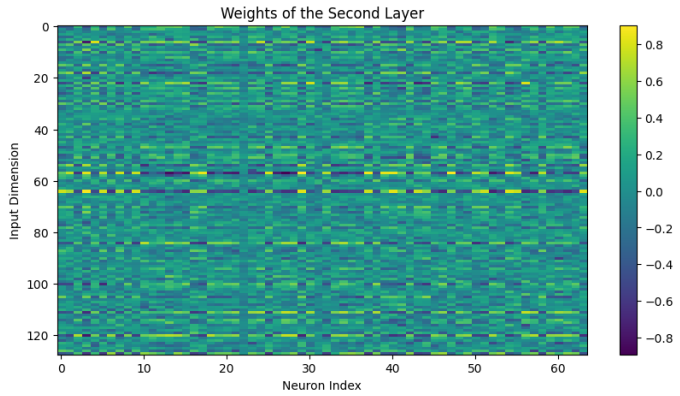
Finally, a neural network model is designed to classify mushrooms into edible or poisonous categories. The model consists of two hidden layers with 128 and 64 neurons, respectively, and an output layer with a single neuron using the sigmoid activation function. The model is compiled with the binary cross-entropy loss function and the Adam optimizer.

A grid search is performed to find the optimal combination of neurons in the first hidden layer and optimizer. The model is trained on the training data, and its performance is evaluated using accuracy as the metric. The best combination of parameters is selected based on the highest accuracy score.

The weights of the first and second layers are visualized as heatmaps to gain insights into the model's learned representations.



4 Weights of the first layer



5Weights of the second layer

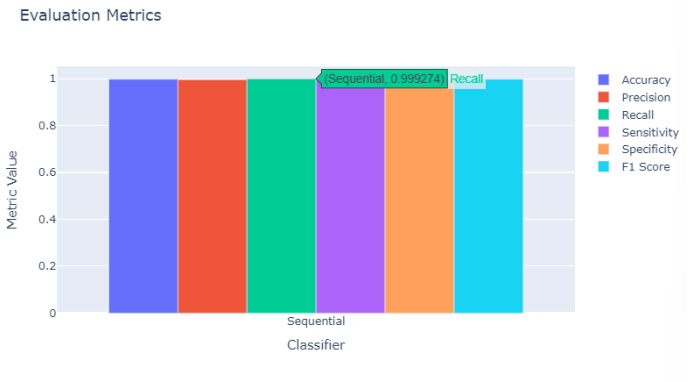
For the mushroom classification dataset, this model aims to learn patterns and relationships between the input features and the target variable (edible or poisonous). By analyzing the weights, we can understand which input features are most important for the model's predictions and how they interact with each other.

The visualization of weights can help identify:

1) Which input features are most influential in the model's predictions
2) How the model combines features to make predictions
3) Potential biases or correlations in the data that the model has learned

By interpreting the weights and evaluating the model's performance, we can refine the model, select the most informative features, and improve its accuracy in classifying mushrooms.

The current model is evaluated based on classification evaluation metrics:



As previously mentioned, the most important metric for this task is recall. Through the plot, we can determine the value for recall to be 0.99. Since the model is to be used for a critical task, high metric values are desirable.

3. Sample predictions

To test the dataset for specific instances, four sample predictions were made for different classifiers.

Classifier Name	Prediction	Actual
Gradient Boosting	Poisonous	Poisonous
Decision Trees	Edible	Poisonous
Random Forest	Poisonous	Poisonous

4. Conclusion and Future Direction

This project, with the main task of classification, has successfully demonstrated the application of various machine learning techniques and lastly a deep learning model. Through meticulous exploratory data analysis, feature engineering, model selection, and evaluation, we have developed a robust and reliable model that can accurately classify mushrooms as edible or poisonous.

However, the potential for this project's capabilities to be improved persists. There are several potential directions for future work of the mushroom classification system:

1) **HTML Parsing**: With the use of BeautifulSoup and/or Selenium to parse HTML data from various online mushroom

databases [12]. This advancement will enable us to extract pertinent information such as physical traits, names, and images. The implication of this is a more comprehensive and accurate identification of mushroom species, enhancing the depth and breadth of our dataset.

2) **Image Recognition Enhancement**: Implementing image recognition techniques, such as convolutional neural networks (CNNs), can enhance the model's ability to classify mushrooms based on images. By leveraging CNNs, the model can learn intricate patterns and features from mushroom images, improving classification accuracy as we introduce a new aspect and use for the model.
3) **Interactive Web Application**: Developing an interactive web application that allows users to upload mushroom images for classification can enhance user engagement and accessibility. This application can provide real-time predictions and explanations, making it a valuable tool for mushroom enthusiasts, foragers, and researchers.
4) **Deployment on Mobile Platforms**: Adapting the classification model for deployment on mobile platforms, similar to the smartphone application mentioned in the sources, can make mushroom classification more accessible and convenient for users in the field. This mobile deployment can enable real-time mushroom identification and enhance user safety during foraging activities.
5) **Overfitting**: Currently, the model exhibits the possibility of overfitting as inferred by the high accuracy values. It is desirable to introduce more variation in the data to reduce the error in testing. Additionally, methods such as early stopping can be used to improve performance on real world data.

Individual Contributions

Aliyan Ahmed: Led the exploratory analysis, data imputation, and handled outliers and duplicates to ensure data quality and integrity.

Eesha Tariq: Spearheaded the creation of visualization plots, addressed class imbalance issues, and managed data encoding for enhanced data representation and understanding.

AbdurRehman Haroon: Conducted an in-depth grid search to identify the best parameters for the classifier, evaluated model performance, and compared different classifiers to optimize model selection.

Javeria Shahid: Implemented explainable AI techniques and neural networks to enhance model interpretability and prediction accuracy, contributing to the project's overall success.

References

[1] N. Robinson. [Online]. Available: https://rrcultivation.com/blogs/mn/exploring-the-role-of-mushrooms-throughout-history#:~:text=Many%20cultures%20that%20ate%20mushrooms,ceremonies%20in%20the%20Middle%20Ages..

[2] S.-t. C. John A. Buswell. [Online]. Available: https://www.taylorfrancis.com/chapters/edit/10.1201/9780203753682-15/edible-mushrooms-attributes-applications-john-buswell-shu-ting-chang.

[3] P. D. F. E. P. D. m. A. S. P. D. m. T. Z. P. (. D. m. a. H. A.-S. P. D. r. n. Robert Wennig, "Mushroom Poisoning," PMC, October 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7868946/.

[4] N. S. W. Alexander Hanchett Smith, "The Mushroom Hunter's Field Guide," [Online]. Available: https://books.google.com.pk/books?hl=en&lr=&id=TYI4f6fqrfkC&oi=fnd&pg=PA3&dq=can+you+determine+mushroom+edibility+based+on+features&ots=ZHAYnt3qfR&sig=YJZEWWf35hzH6IoSi4EjMT5q9qI&redir_esc=y#v=onepage&q=can%20you%20determine%20mushroom%20edibility%20based.

[5] S. Y. S. ,. M. A. A. ,. R. .. ,. A. B. K. Hameed Ur Rahman, "Acute Liver Injury From Mushroom Ingestion: A Timely Intervention in Mushroom Poisoning," [Online]. Available: https://www.cureus.com/articles/184368-acute-liver-injury-from-mushroom-ingestion-a-timely-intervention-in-mushroom-poisoning.

[6] M. R. G. T. A.-J. S. R. M. G. &. A. K. Maryam Janatolmakan, "Demographic, clinical, and laboratory findings of mushroom-poisoned patients in Kermanshah province, west of Iran," BMC Pharmacology and Toxicology, 13 September 2022. [Online]. Available: https://bmcpharmacoltoxicol.biomedcentral.com/articles/10.1186/s40360-022-00614-1#:~:text=The%20actual%20annual%20rate%20of,mushroom%20poisoning%20cases%20%5B12%5D..

[7] Z. K. S. F. I. F. I. Authors: Kishwar Sultana Kishwar Sultana, "Diversity of edible mushrooms in Pakistan.," [Online]. Available: https://www.cabidigitallibrary.org/doi/full/10.5555/20093347429.

[8] [Online]. Available: https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset..

[9] E. Anello, "7 Steps to Mastering Data Cleaning and Preprocessing Techniques," [Online]. Available: https://www.kdnuggets.com/2023/08/7-steps-mastering-data-cleaning-preprocessing-techniques.html.

[10] [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html.

[11] Z. Kelta, "Explainable AI - Understanding and Trusting Machine Learning Models," [Online]. Available: https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models.

[12] "Mushroom World," [Online]. Available: https://www.mushroom.world/.

[1] M. a. S. M. Hossin, "A_REVIEW_ON_EVALUATION_METRICS_FOR_DATA," [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/37219940/5215ijdkp01-libre.pdf?1428316763=&response-content-disposition=inline%3B+filename%3DA_REVIEW_ON_EVALUATION_METRICS_FOR_DATA.pdf&Expires=1713458786&Signature=IhDRdXJJDJrh35X6YEJUqTx6H~R7tr2RHvokaV9DAYOp6NrfmH0.

[3] D. H. a. H. (. S. M. U. M. L. R. h. Wagner. [Online]. Available: https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset.

[4]