MYCOLOGICAL RISK ASSESSMENT.

A BINARY CLASSIFICATION PROJECT

Table of Contents.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

Project Overview.

Develop a machine learning model for precise classification of mushrooms as edible or poisonous based on their physical attributes. Aimed at aiding mycologists, the model enables accurate risk assessments in fungal species identification.

Business Understanding.

Identifying edible vs. poisonous mushrooms presents challenges due to their varied traits. Leveraging machine learning on the Mushroom dataset's key features like cap shape and odor, this project aims to create a precise classification tool. It aims to boost safety and efficiency for mycologists and enthusiasts involved in mushroom analysis.

Problem Statement.

Identifying edible versus poisonous mushrooms is a critical task for mycologists, demanding precision in classification from physical attributes. To aid in this task, the project seeks to craft a robust machine learning model to serve as a reliable, real-time classification tool during fieldwork, mitigating risks linked to misidentification and bolstering safety measures for researchers and the wider community.

Objectives.

General Objective.

Developing a machine learning model to classify mushrooms as edible or poisonous based on physical attributes, aiding mycologists in reliable species identification for research and safety.

Specific Objectives.

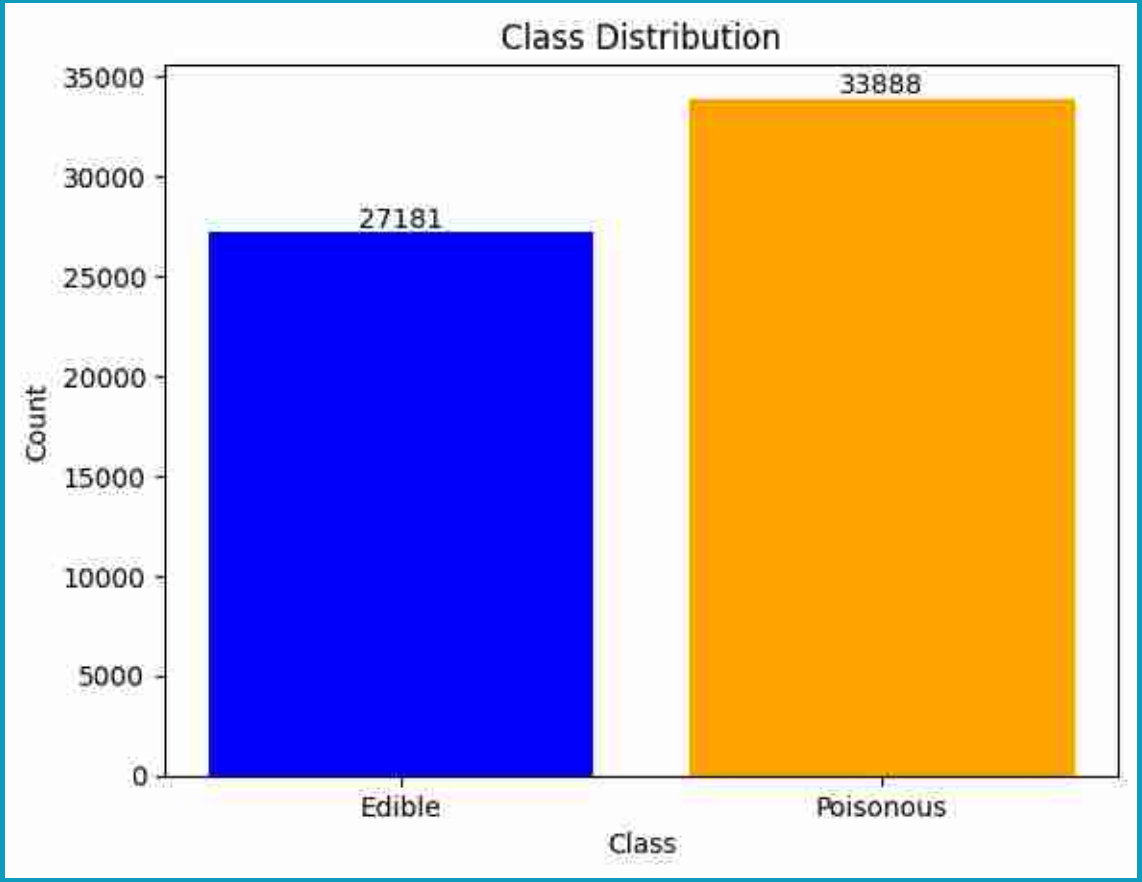
- Develop a classification model.
- Enhance model performance through algorithm selection.
- Explore ensemble learning approaches.
- Ensure model deployment feasibility.

Data Understanding.

This project employs a Mushroom dataset from Kaggle, initially contributed to the UCL Machine Learning repository. It contains 61069 hypothetical mushrooms categorized into 173 species, with 20 variables—17 are nominal, and 3 are metrical.

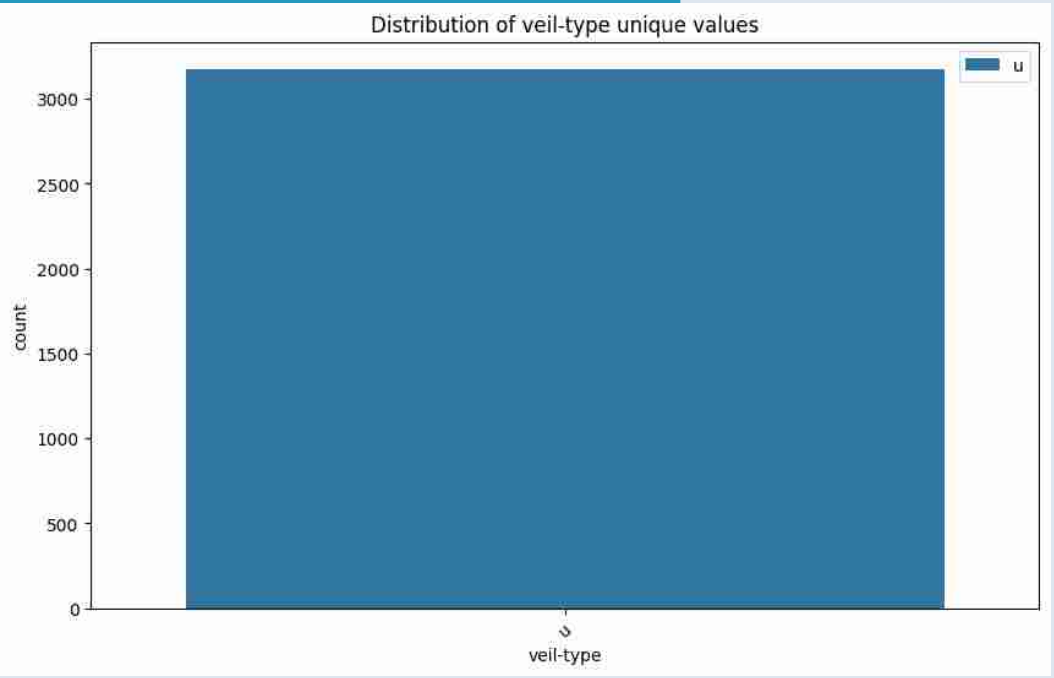
1) **Performed a series of preprocessing steps to parse the dataset to obtain insights**
2) **Univariate analysis to get an understanding of the target column**
3) **Multivariate analysis to understand the relationship of columns**

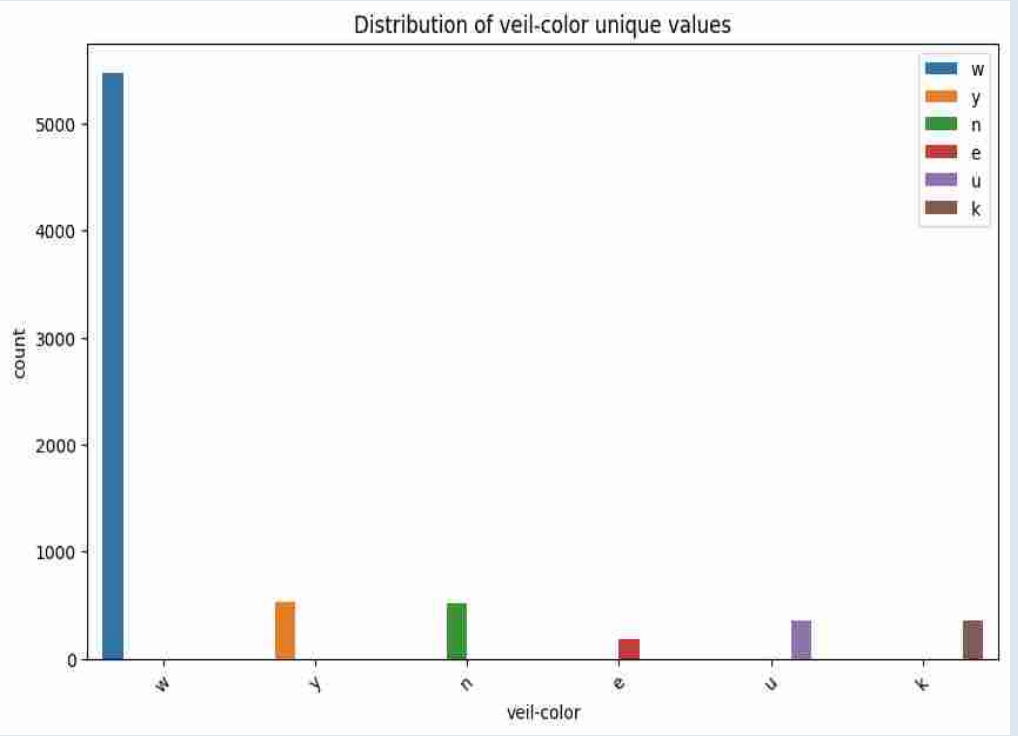
Visualizations



This graph shows a slight class imbalance in our dataset.
We are sure to use SMOTE to mitigate this.

Visualizations

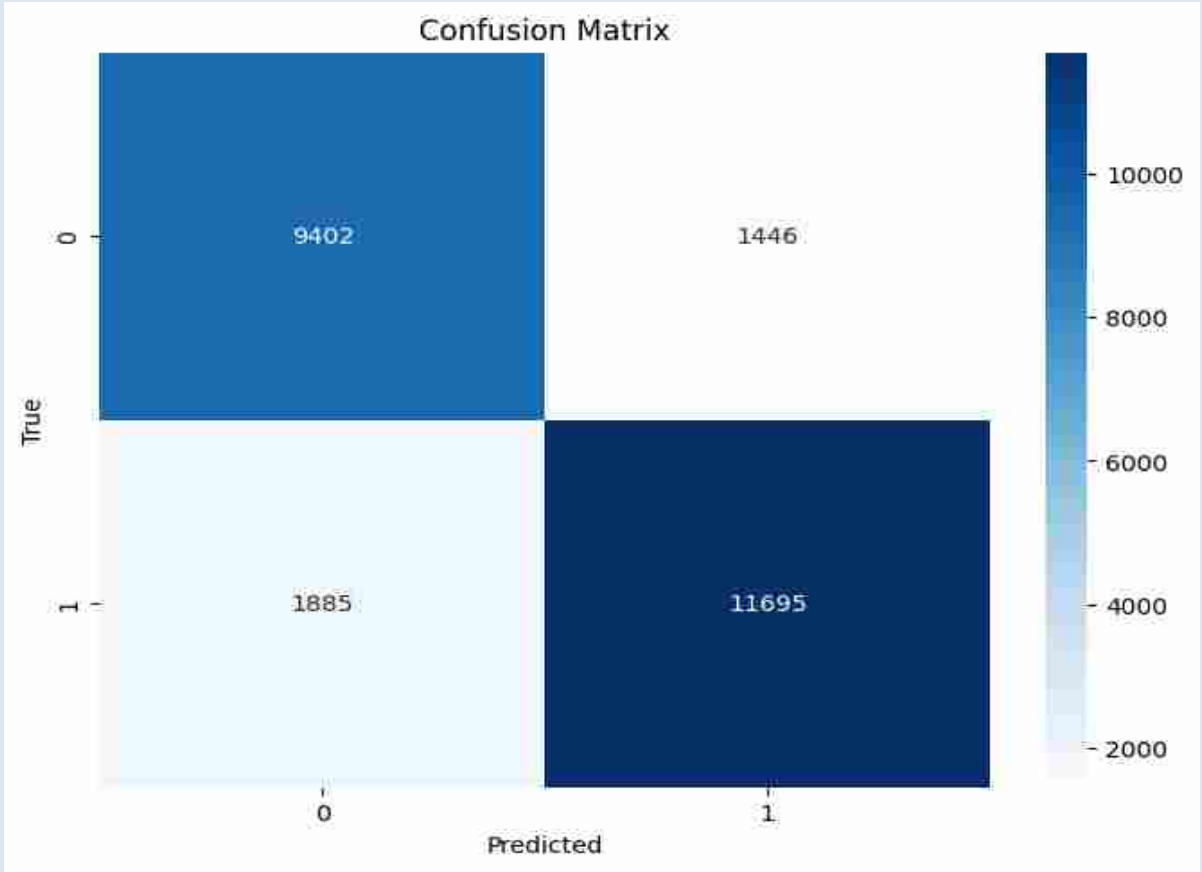




Veil-type and veil-color are flagged for dropping.

Modeling & Evaluation.

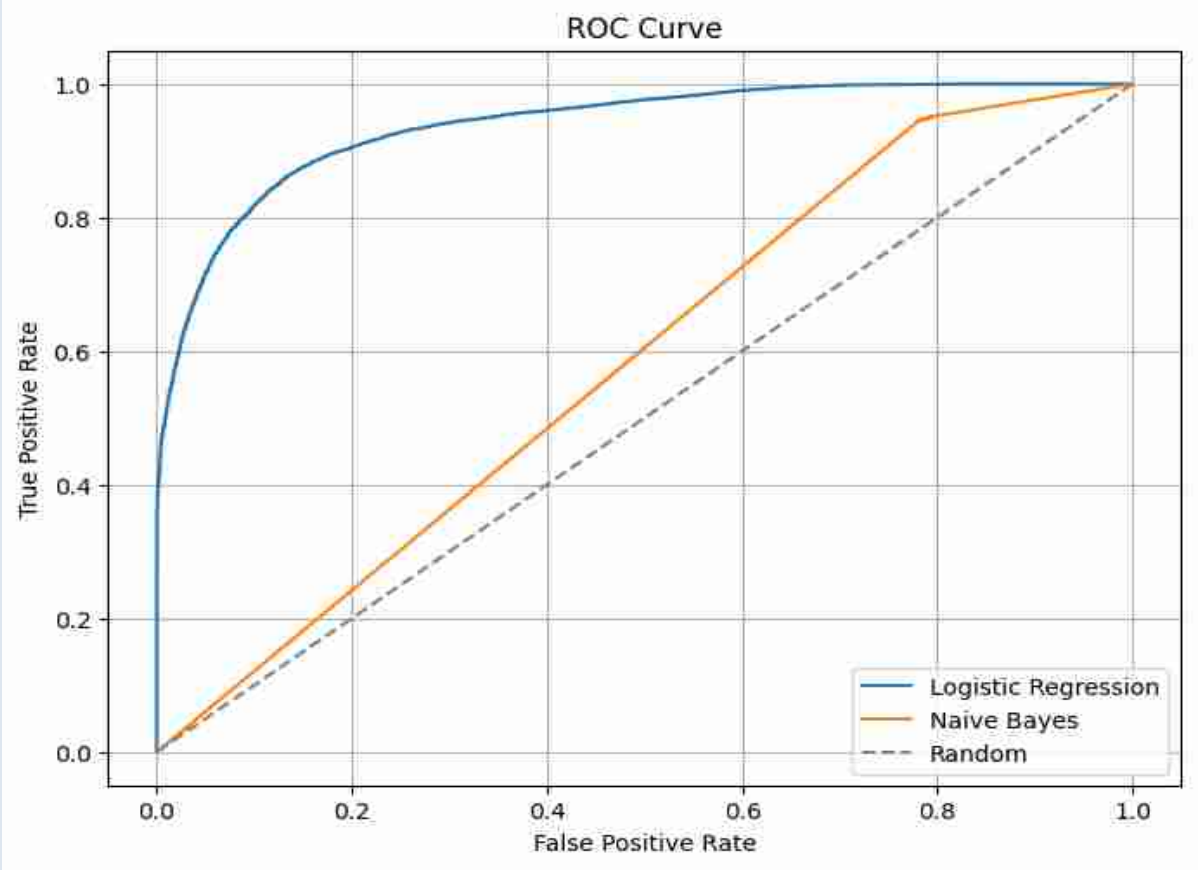
Logistic Regression (Baseline Model)



The confusion matrix highlights that the model misclassified 1446 edible mushrooms as poisonous and 1885 poisonous mushrooms as edible.

Modeling & Evaluation.

Naive Bayes vs baseline model.

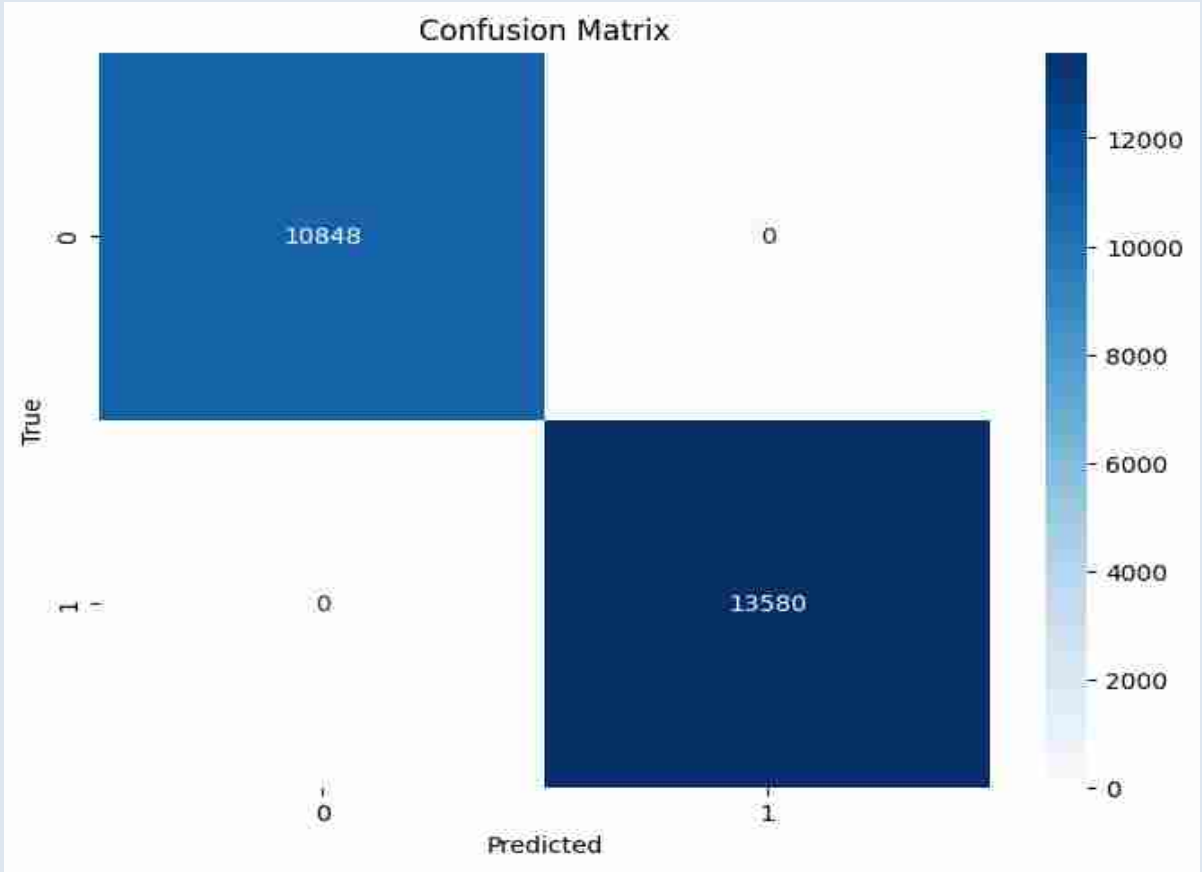


Based on the ROC curve, the Logistic Regression performs better than the Naive Bayes.

- **Baseline model is closer to the left-hand border.**
- **Naive Bayes is closer to the diagonal (random guessing) suggesting a less effective model.**

Modeling & Evaluation.

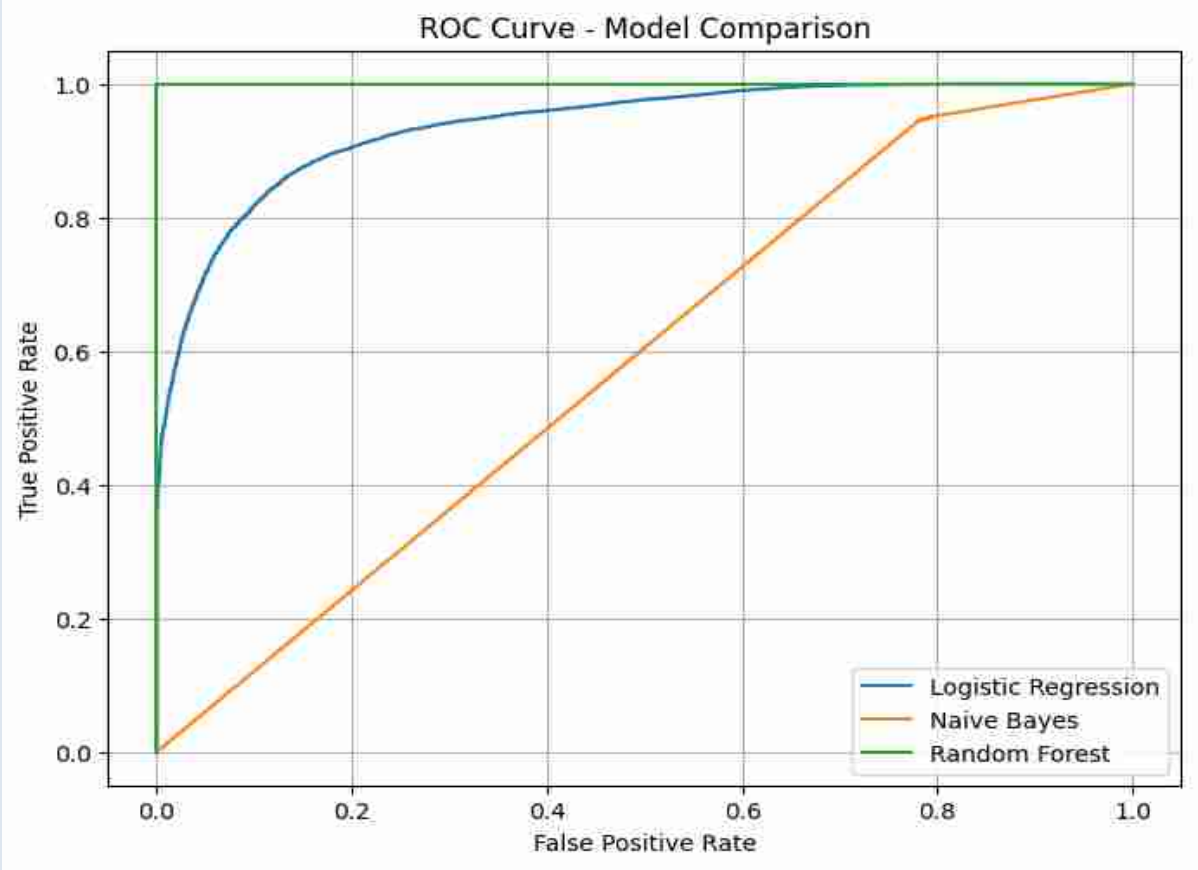
Random Forest (Advanced Model)



The confusion matrix highlights that classified all the classes perfectly.

Modeling & Evaluation.

Random Forest vs Naive Bayes vs baseline model.



Based on the ROC curve, the Random Forest performs better than the others.

- **Random forest model is closest to the left-hand border.**
- **Its indicating a better balance between high sensitivity and low false positive rate**

Challenges

- **Imbalanced dataset:** some of the methods used to preprocess because of imbalance required a lot of processing power and time.
- **Hyperparameter tuning:** techniques used here like GridSearchCV are computationally expensive.

Conclusion.

The project successfully built an effective model for classifying mushrooms as edible or poisonous using their physical traits. This model helps mycologists and enthusiasts quickly identify mushrooms, improving safety in mushroom hunting and research.


Thank you