December 6th, 2024

CS-505-740

Binary Prediction of Poisonous Mushrooms Report

Colton Starkey, Kyle Schwartz, Sean Durbin,

Thabo Adams, Wiley Hartzog

1.0 INTRODUCTION

1.1 Project Overview

Mushrooms are an extremely common fungus with over 14,000 unique species across the globe. While mushrooms are a very common food, that much biodiversity there are many that are poisonous as well. Much of the illness caused by mushrooms is from those picked in the wild that people might not be educated on (Colorado State University, 2024). This leads us into the current project.

We were provided with .csv training and test files that contain data on various types of mushrooms. For these mushrooms, we were tasked with making predictions on whether they were poisonous, or edible given various traits. These range from things like cap diameter to color that we will go over in greater detail throughout later sections in this document. We did this by creating machine learning models that process and interpret the data to make accurate predictions on this classification of mushroom.

1.2 Importance of Results in a Business Context

If we are looking at this from a business perspective and trying to sell these models to a consumer, this could be very beneficial to someone that is getting into picking wild mushrooms or just wants some general guidance on what to look for in certain types of mushrooms. Although for liability purposes since this is not 100%

accurate would not want someone to put their own live in the hands of our model and fully trust it (much like how Chat GPT leaves a disclaimer that they are only a model), it could help someone make a decision on the fly and build an understanding.

Besides the average consumer, identifying what trends lead to different types of mushrooms could aid in scientific research if the model was sold to an institute that studies them. If for example it is determined that many of the poisonous mushrooms have cap diameters between an x and y value, this could provide insight to research further into what the cause of this could be and save lives in the future if the knowledge is leveraged properly.

2.0 Dataset Overview

2.1 Dataset Description

Our dataset is sourced from Kaggle as part of a competition series they did for machine learning (Reade, 2024). It is a fairly large dataset with a little over 1,000,000 columns of data in both the test and training sets. In each of these columns, we have rows that contain attributes for the following:

- Class
- Cap diameter
- Cap shape
- Cap surface area
- Cap color
- Does bruise or bleed

- Gill-attached
- Gill spacing
- Gill color
- Stem height
- Stem width
- Stem root
- Stem surface
- Stem color
- Veil type
- Veil color
- Has ring
- Ring type
- Spore print color
- Habitat
- Season

The class refers to whether the mushrooms are edible or poisonous and was only included in the training file that was provided.

2.2 Key Features and Attributes

All of the data in each of the above columns besides the actual id of each mushroom was stored as either a float or character/string, or boolean. For example, cap diameter was a float value such as 8.64 for ID 3116945 of the test data. For the "does bruise or bleed" column, that was a Boolean value with "f" for false or "t" for true.

Although they might not mean much when reading in plain English, characters like "a" for autumn and "w" for winter were used for the season attribute

which are still used by the model to identify trends and make predictions.

2.3 Data Preprocessing Steps

The data could not be simply uses as provided and get accurate results. We had to encode the data and many of the columns contained null values and with models like Random Forest, lead to errors when processing. To overcome this, some amount of data preprocessing had to be completed.

We first began by using pandas to pull in data from the files we downloaded locally. Then the category columns were identified, and data was mapped in and encoded as integer values.

After this, data was split into training and test but we also normalized by dropping the less important or redundant columns to reduce unneeded noise in the dataset such as veil-color and veil-type that we determined through a separate random forest model.

As mentioned, there were also null fields in the data we needed to clean up. We did this this using a simple imputer that would check for NaN values and replace them with the mean value of whatever column they were found in. After all of these steps were taken, models were ready to train and test.

3.0 Model Development

3.1 Machine Learning Algorithms Implemented

The first model in our file used was made using XGBoost and uses a gradient boosting algorithm. XGBoost was great as it featured support for handling categorical features like what we have and is generally considered quicker with training. Considering most members of our group do not have a dedicated GPU to run these computations and have a dataset of over 1,000,000 items, this was important for us. It also has some features to natively handle imputation, although that was less necessary here since we handled ourselves.

A logistic regression algorithm was implemented which is designed specifically for binary classification, which is what we were doing here. Since there were only two dependent variables here it would be binomial logistic regression *Logistic Regression in Machine Learning*. GeeksforGeeks.).

The last model we used was a random forest classifier. This model as the name suggests, is again great for classification and best suited for larger datasets like we have here. Another strength it has is giving some more unique insight into feature importance as feature selection is actually imbedded within decision trees. The way it works is by creating a set of decision tress from a randomly selected subgroup of the training set. Then, uses the collection of outputs from the different decisions to output a final class prediction. The randomness of the decision trees and this process is known as “feature bagging” and helps to prevent overfitting and reduce dominance of any single feature (*Random Forest classifier using Scikit-learn*. GeeksforGeeks.).

3.2 Criteria for Model Selection

The final model we chose to go with was our random forest classifier. It seemed to provide the most accurate results and feature importances we could gain insight from. It was also one of the 3 models listed in the final project outline to include in our documentation and file.

4.0 Performance Evaluation

4.1 Evaluation Metrics

We used multiple different metrics to gauge performance of the model. These are accuracy, precision, recall, F1 score, and ROC-AUC.

- Accuracy: Percentage of total predictions made that were correct.
- Precision: measured by true positives / (true positives + false positives)
- Recall: measures by true positives / (true positives + false negatives)
- F1 Score: used in binary classification and is derived from both precision and recall. Generally considered a better metric than accuracy.
- ROC-AUC: stands for receiver operator characteristic area under curve. Represents the probability of a model given randomly picked positive and negative examples that it will rank positive higher than negative (*Classification: Roc and AUC | machine learning | google for developers*.).

Additionally, to ensure that our Random Forest model was not overfit to the training data, we made use of k-fold cross-validation. By splitting the training

dataset into 5 folds, the model was trained and validated on different subsets of the data. This ensured that the evaluation metrics reflected how well the model generalized to unseen data. The results of this process confirmed that the high accuracy observed, 99%, was not due to overfitting but rather to the model's ability to generalize effectively.

4.2 Results Comparison (Table of Metrics)

Below is a table of the results we found for our models:

Test Results Comparison with Detailed Metrics:

	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
0	XGBoost	0.981167	0.98	0.98	0.98	0.994639
1	Logistic Regression	0.630299	0.63	0.63	0.63	0.683139
2	Random Forest	0.990122	0.99	0.99	0.99	0.995617

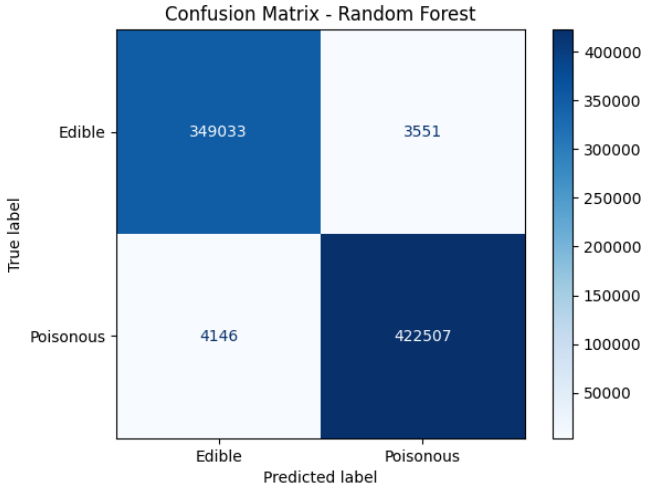
As shown, we saw the best performance in all five of the listed evaluation metrics with the random forest classifier with our XGBoost (gradient boosting model) a close second. Logistic regression had by far the weakest performance with an accuracy of approximately 63% and similar numbers for all other evaluation metrics.

4.3 Confusion Matrix for Final Model

Below is the confusion matrix for our final random forest model. It is color coded with lighter shades meaning fewer values in whatever matrix. The top left is true positive. Top right is false negative. Bottom left is false negative and bottom

right is true negative.

This means we accurately predicted 349,033 true edible values and 422,507 true poisonous giving the 99% rate we saw in section 4.2 for random forest.



5.0 Model Insights

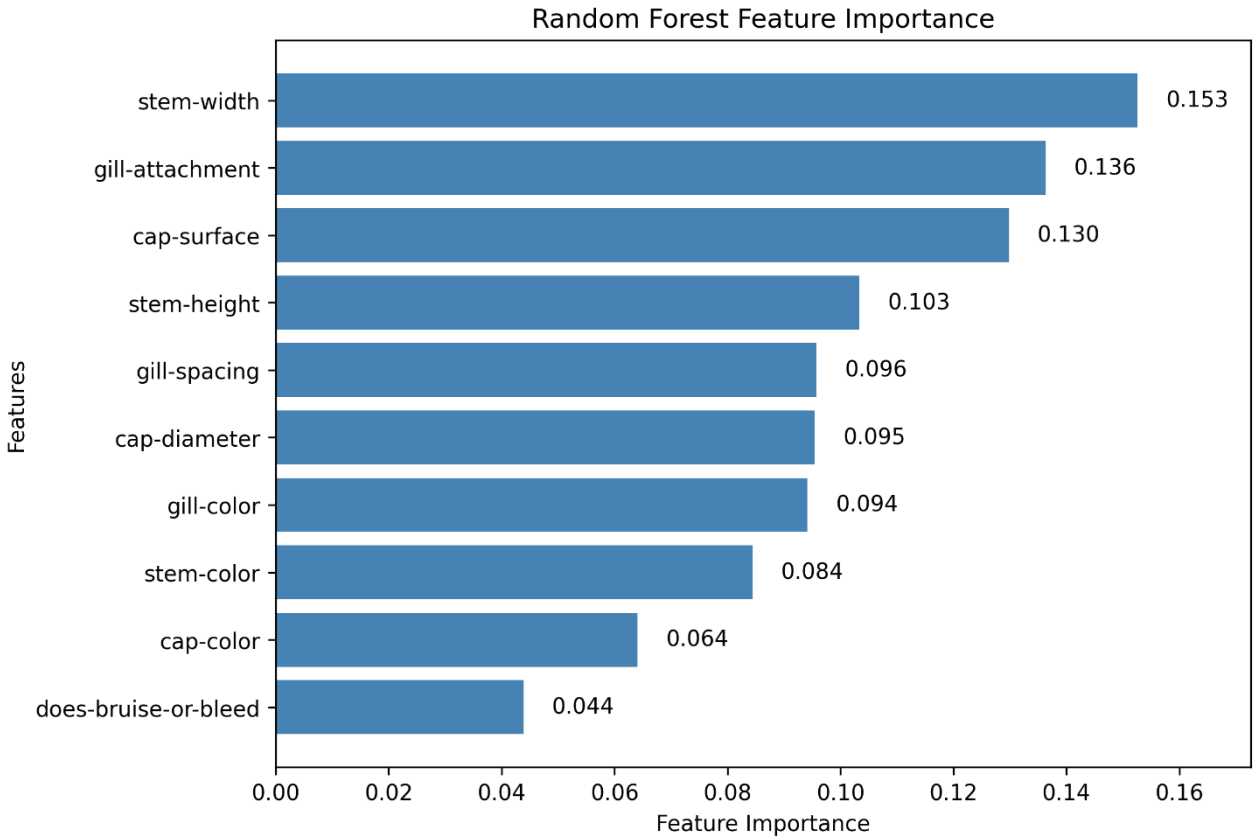
5.1 Feature Importance Analysis

The feature importance analysis of the Random Forest classifier highlighted that certain features had significantly influenced the model's predictive accuracy. Among these features, stem width was the most crucial factor for determining whether a mushroom is edible or poisonous. This is followed closely by gill attachment, cap surface and stem height which all played a significant role in the model's decisions.

The Random Forest model makes use of feature importance as part of its decision-making process and leverages its tools of decision trees. For instance, the stem width likely correlates strongly with characteristics specific to poisonous

mushrooms. This makes it a reliable differentiator. Similarly, cap diameter provides insight into species-level differentiation and gill color offers visual cues related to mushroom toxicity.

The feature importance values are illustrated in the chart below. This visualization provides a clear perspective on the contributions of each feature and showcasing how the model prioritizes these factors during classification. By focusing on the most impactful features, the Random Forest model achieves a high degree of accuracy.



5.2 Model Interpretation

The Random Forest model works by combining the predictions of many decision trees to classify mushrooms as either edible or poisonous. Each tree is trained on random parts of the data and look at different features to make its decisions. The final prediction is based on a majority vote from all the trees, which makes the model more accurate and less likely to make mistakes.

The model makes use of conditions in the decision trees to classify mushrooms. For example, if a mushroom has a stem width in a certain range and specific gill attachment traits, then the tree might classify it as poisonous. By combining the results of many trees, the model can consider different feature combinations to make reliable predictions.

A key strength of the Random Forest is that it's easy to understand how it works. Tools like feature importance charts show which features mattered most in the predictions. The confusion matrix revealed a small number of false negatives, where the model incorrectly classified poisonous mushrooms as edible.

Overall, the Random Forest model is accurate and provides clear explanations of its decisions. This has made it a useful tool for us when attempting to predict mushroom toxicity and the key traits involved.

6.0 Business Impact

6.1 Steps to Solve the Problem

As shown in the confusion matrix, there was still 4146 cases in our testing

where we falsely identified mushrooms as edible when they were in fact poisonous. While this comes out to less than 1% of cases, this is still more room for error than we would feel comfortable with putting full trust in considering that could be a major safety concern.

It is not possible to ever achieve 100% accuracy but if we were to share this data with others, it would be good to consider bringing a subject matter expert on board to further validate data. This would also help give insight into where we could improve the model and potentially add more independent variables for the binary classification.

6.2 Projected Outcomes on Business Metrics

If we are to improve overall performance and accuracy of our random forest model, we could further increase safety and reduce number of false negatives. False negatives are problematic too but also less concerning since that would not be a life-threatening mistake.

If we consistently produce accurate results and this proves to be an effective tool for educating general people on what to look for in an edible mushroom, this will help with customer churn as it would lead to better retention.

7.0 Limitations

7.1 Challenges and Constraints

To run each mushroom through the model, it must have certain features characterized. Those features are subject to interpretation themselves and therefore can only be classified as accurately as they are described in the data.

Using this tool requires experienced cross-referencing to ensure safety. Since it cannot be safely used as a standalone tool, its user base is still somewhat limited, although it does help create a good foundation.

The legality of this could be tricky, as there are food safety implications. Creating a model that attempts to identify if a mushroom is toxic or not, could place the creators at risk of liability in the event of misclassification. Any further distribution will need to be screened by a legal authority.

7.2 Impacts on Model Performance and Generalizability

Not every mushroom that is found in the wild is described in our data. This can affect the model's ability to generalize data, as well as its ability to identify certain species that a user may encounter. However, it will still classify that mushroom as described, so we must be aware of that possibility.

By using only tabular data, our model is untrained on image analysis. Therefore, it is required for a user to identify certain features and correctly input the data to be classified. Therefore, it is not as user efficient as taking a picture and getting a result quickly. However, it is highly accurate and requires low computing power, which are two areas where image analysis can struggle.

8.0 Future Work

8.1 Areas for Further Research

Further research in this field likely comes attached to image analysis. Although it requires vastly more computing power, and it likely would sacrifice a slight amount of accuracy, incorporating mushroom images alongside the tabular data could greatly expand the market for this tool. This could lead to mobile and web app development, as well as the ability to potentially classify new species that have little to no data implemented in the training model.

We could further improve the generalizability and usefulness of this model by incorporating image analysis into the model. It could also potentially incorporate identifying the species itself alongside its edibility. Submitting an image that is backed by accurate, tabular data could combine the high accuracy of our model into a package that can be run more easily utilized by all. This could lead to development across many fields, including medical, food safety, forensic, and even conservation efforts.

8.2 Recommendations for Improvement

Although not currently available, to improve our model's training data, we would be wise to include more species and features in different environments and growth stages, as these can vary the characteristics that the model relies on. Furthermore, exploring additional features not included in our data could uncover an even more important classifying feature. This might include a crowdsourcing effort

from other mycology groups that would be willing to submit accurate data for model training. More data will always create more opportunities to produce an exceptional classification tool.

Additionally, further improvement could possibly be gained by other hybrid models that experiment past the three models we created. Our combination of a random forest feature selection model and separate random forest training model could be less optimal than other combinations.

9.0 Conclusions

9.1 Implication for Business Decisions

It could impact business decisions outside of our model as well. If this data to distributors of edible mushrooms as well as an example, if a trend is seen where green mushrooms for example are poisonous, they might steer clear of those to avoid concern from the public over safety.

For mushroom hunters using this tool, considering stem-width was the most important feature according to our random forest model that could lead to more of an emphasis being put on that when searching for edible mushrooms. Could also help in scientific research as there are certain traits of poisonous and edible mushrooms that have medicinal properties.

9.2 Summary of Key Findings

Looking at the results of the project in hindsight, several key findings stand out. Among the three models tested, the Random Forest model was the strongest performed and achieved an overall accuracy of 99%. This result was validated through k-fold cross-validation, which demonstrated that the model was not overfit and performed reliably across different subsets of the data. By comparison, the Logistic Regression model only achieved an accuracy of 63%.

Feature importance analysis highlighted stem width as the most critical factor in determining mushroom toxicity, followed by gill attachment and cap surface. These features not only guided the model's predictions but also offered valuable insight for identifying key traits that distinguish poisonous mushrooms from edible ones.

From a business perspective, the Random Forest model has the potential to be a valuable educational tool. It could assist users in making informed decisions about mushroom edibility and provide a foundation for further research on toxicity trends. However, the presence of false negatives (where a poisonous mushroom might be classified as edible) highlights the importance of using it as an advisory tool rather than a definitive source of classification.

Looking to the future, we could focus on enhancing accessibility and functionality. Incorporating image analysis capabilities would allow users to classify mushrooms through photographs rather than intensely detailed datasets. This would broaden the tool's usability and appeal greatly amongst food safety, conservation, and education efforts.

In conclusion, the Random Forest model demonstrated exceptional accuracy and was support by robust validation methods like cross-validation. These findings highlight the model's potential to serve as a reliable and insightful tool for predicting mushroom toxicity.

10.0 References

10.1 List of Cited Sources

Colorado State University. (2024, June 19). *Mushrooms*. Food Source Information. https://www.chhs.colostate.edu/fsi/food-articles/produce/mushrooms/#:~:text=There%20are%2C%20however%2C%20many%20morphological,on%20some%20type%20of%20substrate

GeeksforGeeks. (2024b, June 20). *Logistic Regression in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/understanding-logistic-regression/

GeeksforGeeks. (2024a, January 31). *Random Forest classifier using Scikit-learn*. GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/

Google. (n.d.). *Classification: Roc and AUC | machine learning | google for developers*. Google. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=The%20area%20under%20the%20ROC,random%20positive%20and%20negative%20example.

Reade, W., & Chow, A. (2024). *Binary prediction of poisonous mushrooms*. Kaggle. https://kaggle.com/competitions/playground-series-s4e8