Is this edible? A machine Learning Project

Gabriele Maria Morello - Polina Kireeva

1 Introduction

The Machine Learning Assignment focused on the classification of mushrooms using a dataset with information about 61069 hypothetical mushrooms based on 173 species and classified as edible, poisonous or having no information about edibility.

The dataset used for the assignment is called "Secondary Mushroom Dataset" and it is derived from the one with empirical data about a variety of mushrooms.

The original dataset was not considered for the assignment because it did not allow for some of the forms of pre-processing (e.g. imputation of missing values) and it did not contain enough data (173 entries) to be used for training of many machine learning algorithms.

The dataset used for the assignment is available at: https://archive.ics.uci.edu/dataset/848/secondary+mushroom+dataset.

The assignment group is composed of the members:

- Gabriele Maria Morello;
- Polina Kireeva.

The activities were done following the structure of the Machine Learning course covering the relevant topics to simplify the final evaluation process.

All members were directly involved in defining and reviewing all the sections of the machine learning assignment providing meaningful feedback or adjustments to the analysis performed.

The group organization was structured without a rigid separation of tasks regarding sections, the idea was to adopt an agile approach with an iterative process over modular sections.

The member Kireeva supervised the definition of the problem and evaluated possible algorithms to be implemented.

The problem was to find a suitable dataset that could offer possibilities for different steps of pre-processing and provide analysis opportunities to cover almost all the CRISP-DM cycle.

The member Morello searched for a dataset suitable for the assignment and reviewed existing papers covering the topics of Machine Learning in mushroom classification (more details in the next section).

The Data understanding, Modeling and Evaluation sections cannot be attributed to a specific member because these tasks required a cooperative approach to discuss which procedures will provide the best results regarding the machine learning assignment.

November 25, 2025

2 Related Work

There were three papers analyzed to evaluate different types of methodologies and approaches to obtain useful insights for the Machine Learning assignment. The articles are:

- "Mushroom data creation, curation, and simulation to support classification tasks"[1];
- "A Comparative Study of Machine Learning Methods for Optimizing Mushroom Classification"[2];
- "IoT enabled mushroom farm automation with Machine Learning to classify toxic mushrooms in Bangladesh"[3];

The first paper analyzes topics that overlap with the assignment and could be useful to evaluate different approaches and possible future improvements.

Regarding data curation, an interesting insight provided is how the way data is collected can influence the choice of the algorithm (e.g. image or attributes based).

This also reflects on the dataset used for the algorithm training, the paper analyzed the original dataset and created a secondary one with increased number of entries while still remained balanced regarding the edible to poisonous ratio.

It may seem like an obvious conclusion but having agency over the data collection choices can produce effects on the overall cost (economic and time) of the entire process.

As for data quality and integrity there was a similar thought process regarding the use of correlation analysis in

order to remove reduntant variables.

For algorithm use, the paper evaluated five different classifiers for the predictive models with metrics like accuracy and F2 score (prioritize recall over precision) and then compared to the original dataset.

This is beyond the scope of the assignment but it provides an interesting take on the usage of more fine-tuned metrics to evaluate the training of the machine learning algorithm, so it was considered useful for the assignment purpose.

The second paper is focused on models comparison, the ones considered are:

- Random Forest;
- Gradient Boosting Machines (GBMS), a consecutive input strategy addressing mistakes of the past models by training weak learners (e.g. shallow decision trees) in a gradient descent framework;
- Ada Boost (Adaptive Boosting), consolidating a succession of feeble classifiers and changing their weight at every iteration to adjust errors. The last classifier is a weighted mix the previous ones, giving weight to the ones best performing on training data;
- Extra Trees (Extremely Randomized Trees) an ensemble technique similar to Random Forest but arbitrary (without usage of Gini or data gain);
- Bagging (Bootstrap Aggregating).

The metrics used were: accuracy, precision, recall, F1 Score, ROC AUC (probability of higher random Positive than random Negative cases), Matthews Correlation Coefficient (showing nature of binary classifications) and Cohen's Kappa (measurement arrangement between two raters ordering into unrelated categories).

The main take from this paper is the use of different models to evaluate their effectiveness by using different parameters for comparison. It provides theoretical grounds for evaluating new possible models for the assignment and new metrics to evaluate the performance of the models.

The last paper also evaluates different types of models but the main point for the assignment is the evaluation of FP and FN. It is important to properly assess them because in the case of mushroom classification a FP can pose a threat to health and safety, especially when deployed by an agricultural company.

Multiple parameters were used to account for the need to have a precise assessment on how correct the predictions on edible mushrooms are.

Another interesting analysis regards execution time. It is beyond the scope of the assignment but should be evaluated when handling projects that process enormous amounts of data. It should be important to evaluate pros and cons between more efficient models and their precision.

3 Proposed Method

The dataset (Secondary Mushroom dataset) had a problem regarding data visualization, the source of it was found in a delimiter error in the .csv file because instead of displaying a comma a semicolon was used.

The dataset was analyzed as a dataframe to have a general overview of the data. It originally contained 21 features and 61069 observations with two data types: "object" and "float64".

The data profiling step helped identify the presence of null values for each feature and their predominance over the total entries.

Then the decision was made to drop all the features with high null values, so "stem-root", "veil-type", "veil-color", "spore-print-color" and "stem-surface" were dropped.

For the features "gill-spacing", "cap-surface", "ring-type" and "gill-attachment" a grafical analysis was performed to evaluate how frequently the entries occurred for each category and how evenly distributed they were.

The possibility of imputation was considered, however with consideration to the potential bias that imputation could bring for the amount of values missing the decision to drop them was made.

The features "cap-shape", "cap-color", "gill-color", "stem-color", "habitat" and "season" had their values represented as letters so, since most algorithms only support numerical values, the use of encoding was needed.

Simply mapping the letters as numbers could mislead some algorithms to see the values as a scale, which is not true in this case of colors, seasons, shapes and habitats, so one-hot encoding was chosen as the appropriate encoding approach for this assignment.

The downside of applying it to all 6 features was that it increased the amount of features to 51, more than double compared to the original ones (an increase of 30 from the initial 21). This amount will, however, be later reduced to simplify computations and save resourses.

The numerical features "cap-diameter", "stem-height" and "stem-width" had values using different scales (0.38 to 62.34, 0.0 to 33.92 and 0.0 to 103.91), so they were all normalized to the scale of 0 to 1.

The next step was the creation of a correlation matrix, which would help identify redundant features - if any two or more features were highly correlated, it would mean that having both present would not be highly beneficial for the

model's accuracy.

A correlation matrix heatmap was used for a visual representation of the level of correlation between all features, then a filter was applied so only highly correlated (both positive and negative correlation) features would be displayed for clarity.

The features "season_a_encoded" and "season_u_encoded" showed a correlation of over 0.75. This was considered to be above the threshold of highly correlated features and so "season_a_encoded" was discarded.

This concluded the data understanding phase. All the previous steps were done to ensure a proper set of data that could be used to train the machine learning algorithms.

The next phase was modelling the data, the first step consisted of splitting the data into a training set (data from which the model would learn to find patterns) and a test set used to verify how the model would perform.

After the splitting was completed, tree classification algorithms were considered, with an accuracy score to evaluate how well the model is able to identify edible mushrooms.

In addition to the accuracy score, a graphical representation of how the accuracy of the classification changes based on the depth of the decision tree was modelled. An optimal depth was chosen based on where the steep incline in accuracy met the plateau, to avoid overfitting and to simplify computations.

Feature importance was also calculated to see if all of the features selected were relevant in the modelling of the algorithm.

In view of the decision tree's good performance, another algorithm of the same family, the random forest classifier, was deployed to evaluate the difference in accuracy.

Lastly, AutoML was used to evaluate the best models and hyperparameters for the Secondary mushroom dataset.

4 Results

The results of the assignment review different types of tree algorithms used, their accuracy and an AutoML analysis to verify the existence of better algorithms.

A shallow decision tree with a maximum depth of 2 was used to see the baseline accuracy of a simple model, which was 0.63. Then decision trees were trained with increasing depth and the results were displayed in a graph.

The graph showed a significant decrease in accuracy gain after the depth of 13, and a plateau after 16. A decision tree with the depth of 13 showed an accuracy of 0.92.

Since the decision tree showed good results, it was decided to apply a more complex algorithm of the same type - the random forest. The accuracy achieved was 0.996 (rounded up for simplicity).

Then, in an attempt to push the efficiently achieved accuracy even further, AutoML was used. With a time budget of 60 seconds, the best algorithm was determined to be XGBClassifier with the accuracy of 0.995 (rounded up for simplicity).

Considering the relatively large number of features, feature importances were evaluated - the ones with a 0 value for the deeper decision tree were removed.

As a result, the decision trees' accuracies didn't change and the AutoML result showed a tiny accuracy improvement (approximately 0.0004).

As for the random forest, a tiny drop in accuracy was observed (approximately 0.00008) - this is insignificant, as such changes could be attributed to the randomness of the random forest - meaning, with a different seed the difference could have been reversed.

Feature importance for the random forest was also reviewed to make sure the 10 removed features did not hold much importance to the algorithm, and they did not: the one with the highest importance out of the removed ones had 0.0026 (less than 1%).

Since no considerable negative results were observed, the removal of features was kept to simplify computations.

5 Conclusion

The chosen Secondary Mushroom dataset provided ample opportunities for preprocessing, allowing the members to use encoding (binary and one-hot encoding) and normalization.

Through data profiling, correlation heatmaps and feature importance evaluation the members were able to evaluate the potential impact of different features on further algorithm training, and make decisions on selecting the features to be used.

In view of the relatively high amount of features to be considered, most of which were one-hot encoded (categorical in their nature), for the classification problem of this assignment the members ended up leaning towards tree algorithms, namely the Decision Tree and Random Forest as a more complex algorithm.

This selection of algorithms was further supported by AutoML, which suggested XGBClassifier as the optimal algorithm (within a set budget) - the eXtreme Gradient Boosting model that uses a tree-boosting framework with the approach of training an ensemble of weak tree classifiers

to achieve a strong predictive model.

The selection of tree algorithms also aligns with the literature reviewed in the ”Related Work” section, where all of the considered algorithms work with (or can work with, and commonly do) decision trees.

Both the assignment and the reviewed article implemented the Random Forest, and the XGBClassifier proposed by AutoML uses the same principle as the reviewed GBMS, as it follows the principle of gradient boosting, they are similar but differ in modelling details.

Within the scope of our assignment, both of these more refined algorithms achieved a very high accuracy (over 99.5%).

REFERENCES

[1] Mushroom data creation, curation, and simulation to support classification tasks. https://www.nature.com/articles/s41598-021-87602-3.
[2] A Comparative Study of Machine Learning Methods for Optimizing Mushroom Classification. https://www.jcbi.org/index.php/Main/article/view/726.
[3] IoT enabled mushroom farm automation with Machine Learning to classify toxic mushrooms in Bangladesh. https://www.sciencedirect.com/science/article/pii/S2666154321001691.