Is this edible? A machine Learning Project

Gabriele Maria Morello - Polina Kireeva

Abstract—The abstract goes here.

Index Terms—Machine Learning, Mushroom, Edible.

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 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 different variety of mushrooms.

The original dataset was not considered for the assignment because it did not allow various forms of pre-processing (e.g. imputation of missing values) and it contained too few variables (173 entries) to be used for training of the machine learning algorithm.

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 trying to cover all the topics and simplify the final evaluation process.

The group organization was structured without a rigid separation of tasks regarding sections. All members were directly involved in defining and reviewing all the section of the machine learning assignment providing meaningful feedback or adjustment to the analysis performed.

The member Kireeva supervised the definition of the problem and evaluated possible algorithm to be implemented. The problem was to find a suitable dataset that could offer possibility of pre-processing and provide analysis opportunity to cover the entire CRISP-DM cycle.

The member Morello supervised the dataset search finding it suitable for the assignment and reviewed existing paper 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 have a clear picture of the type of methodologies and different approaches that can provide useful insights for the Machine Learning assignment. The articles are:

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

The first paper analyzes various topics that are present in the assignment, creating parallels that are useful to understand for comparisons and future improvements.

Regarding data curation, an interesting insight provided is how the way data is collected can influence the algorithm preference (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 edibitily to poisonous ratio.

It may seem 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.

Regarding data quality and integrity there was a similar thought process regarding correlation analysis 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 of all the previous ones, giving weight to the best performing on training data;
- Extra Trees (Extremely Randomized Trees) an ensemble technique similar to Random Forest but arbitrarely (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 P than random N 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 a theoretical ground for evaluating new possible models for the assignent 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 it because in the case of mushroom classification a FP can cause problem to health and safety, especially when deployed for 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. That is beyond the scope of the assignment but it should be evaluated with bigger projects and enormous amounts of data. It should be important to evaluate pros and cons between better performing models and their precision.

3 Proposed Method

The dataset (Secondary Mushroom dataset) originally contained 21 features (of "object" and "float64" data types). Through data profiling 9 underrepresented (with a lot of NA values, too many to allow for imputation without the risk of significantly distorting the data) features were identified. Then the label (poisonous/edible classification) as well as some true/false features ("does-bruise-or-bleed" and "has-ring") were encoded using 0 and 1 (binary encoding).

Six of the features ("cap-shape", "cap-color", "gill-color", "stem-color", "habitat" and "season") had their values represented as letters - since most algorithms only support numerical values, 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 the one-hot encoding was chosen as the appropriate encoding approach. However, applying it to all six features increased the amount of features to 51 (with an increase of 30 from the original 21, more than doubling it). This amount will, however, be later reduced to simplify computations and save resourses.

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

The next preprocessing step was the creation of a correlation matrix, which would help identify redundant features - if any two or more features were highly correlated, that would mean 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. Features "season_a_encoded" and "season_u_encoded" showed a correlation of over 0.75, so one of the features ("season_a_encoded") was discarded.

First, a shallow decision tree with the 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 desicion tree with the depth of 13 showed 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 0.996 (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 reverse. Feature importance for 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 was kept to simplify computations.

4 Results

This is the result.

5 Conclusion

This is the conclusion.

Appendix A

This is the appendix.

References

[1] This is the bibliography.