# F121182 COP 528 Coursework

## Task 1. Classification of Mushroom Edibility

### I. Introduction

Mushroom, most people recognise it from our dining table, is the fruit body of mycelium. The role of mushroom is the production of large numbers of spores that are borne on the gills below the cap, and the stalk raises fruit body above the ground to facilitate spore dispersal by air currents[1]. Spores have two major roles in the life of fungus, dispersal to a new site or survival until favorable conditions return. Survival spores are liberated by the lysis of hyphae or fruiting bodies[2]. Fungi are responsible for recycling the components of dead plants, the fungal decomposition of leaves is an essential part of the carbon cycle. The nutrients taken up by fungi can be used directly in metabolism, or stored underground. Fungi are not only ingredient of our meals but also used in different medical sectors. Penicillin, one of the most widely used antibiotic agents, was discovered by Alexander Fleming in 1928, from one of the culture plate he left on the laboratory bench while he was away on holiday[3][4]. *Hericium erinaceus*, also known as Lion's Mane mushroom, had been reported to induce neuronal differentiate and promote neuronal survival[10].

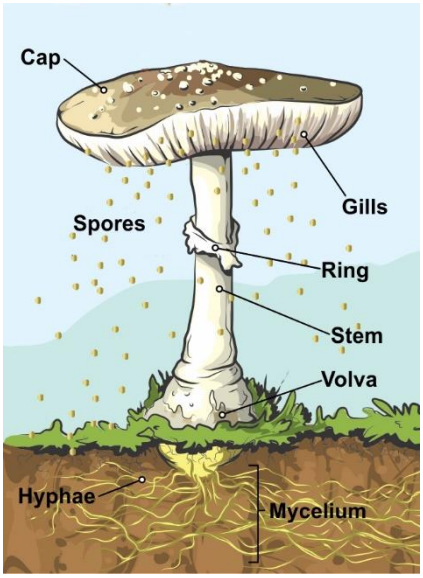


Figure. 1. Structure of mushroom[11].

We have acquired a dataset that contains information regarding mushrooms e.g. the shape of the cap, color of gill, habitat, season. The application of the dataset is to classify if the mushroom is edible or poisonous. From 1999 to 2006, 133,700 cases (7428/year) of mushroom exposure had been reported, mostly by ingestion. Misidentification of edible mushroom species appears to be the most common cause and may be preventable through education[12]. Combine Machine Learning methods with the dataset, we expect to make more accurate prediction in tending to minimise the risk of ingestion.

In comparison of three different machine learning methods, RandomForest Classifier had achieved the highest performance with 0.99 accuracy when including all features. The challenge is that our dataset only contains 173 species, which there are approximately 14,000 species of mushrooms described. This represent our model is very mush overfitted with the data we have and should not be implementing on real world application until we have a more robust dataset for training the model.

### II. Data and Preliminary Analysis

#### A. *Data*

The data is a mushroom dataset inspired by another mushroom dataset created in 1987 by Jeff Schlimmer. It contains 61,069 instances with 21 attributes retrieved from the UC Irvine Machine Learning Repository. The dataset includes hypothetical mushroom information drawn from the book "Mushrooms & Toadstools" by Patrick Hardin. Each mushroom was identified as edible, poisonous, or unknown and not recommended (the latter class was combined with the poisonous class). The attributes are both nominal and numerical, e.g. shape of the cap , stem height.

#### B. *Preliminary Analysis*

From the data we obtained, we noticed that there are many missing values in some attributes, e.g., stem-root with 51,538 missing values, veil-type with 57,892 missing values, veil-color with 53,656, spore-print-color with 54,715 missing values. Due to the amount of those missing values, we will remove some of them instead of replace them to avoid our data being skewed by the imbalance values. The other reason is that as our ultimate purpose is to differentiate edible mushrooms and poisonous mushrooms for human consumption, therefore, we need the data to be as accurate as possible to reduce the risk of wrongly classify poisonous mushrooms as edible. Detail methods used for preprocessing data will be described in *III. METHODS A. Preprocess data.*

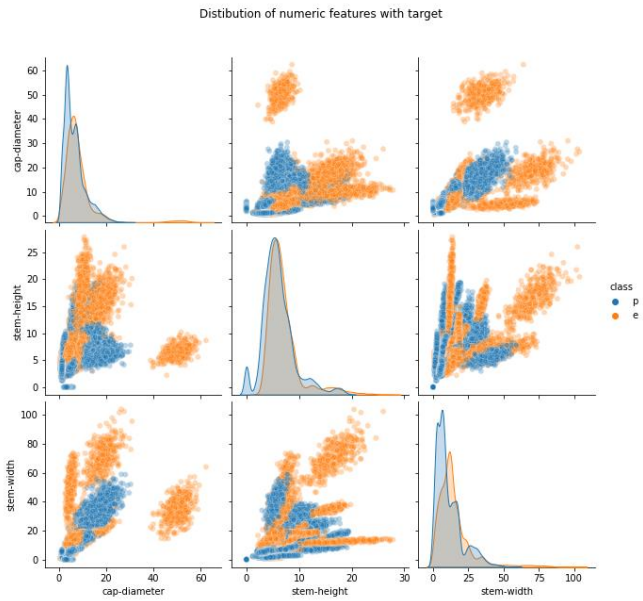


Figure. 2. Distribution of numeric features with target attributes.

In preliminary analysis, from Fig. 2. we can see the distribution of numerical features with the target attribute, and we assumed that none of the numerical features are very indicative of separating edible and poisonous mushrooms. In Fig. 3. are the relation of three features we assumed with common knowledge that will be good indicator of whether a mushroom is poisonous or not. Habitat *w*(waste) and habitat *u*(urban) do not have any poison mushrooms. Gill color *u*(purple) mushrooms are all poisonous. And whether a mushroom has ring or not does not indicate well on our target.

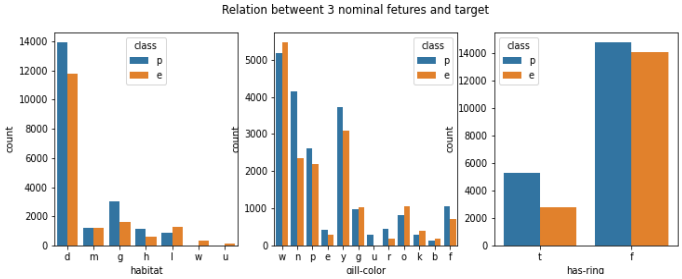


Figure. 3. Relation between 3 features and target.

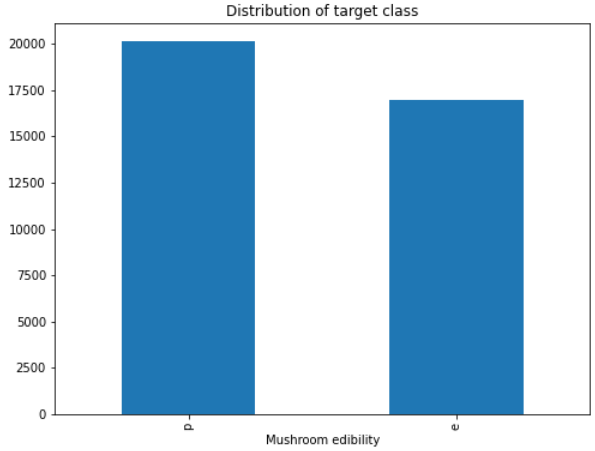


Figure. 4. Distribution of the target class

## III. METHODS

The pipeline of methods that will be implemented can be found in Fig. 5. The methods pipeline is mainly separated to 4 parts, Preprocess data, Build and train models, Evaluation and Prediction.

In preprocessing data, check for missing values, decide if they should be replace with other methods or should remove them.

There are both nominal and numerical features in the dataset. Most of the machine learning methods require input and output to be numeric, therefore, we will use different encoders to convert nominal features. With numeric features, it will be better to apply standardisation or normalisation before training the model, as different features might have different scale and vary a lot. This might skew our model and make inaccurate predictions.

We need to choose which model to use before building and training it. By consider whether we have the target of the data or not, we can decide to use supervised learning method or unsupervised learning methods. In our case we do have the target of the data, so we will use supervised learning to train

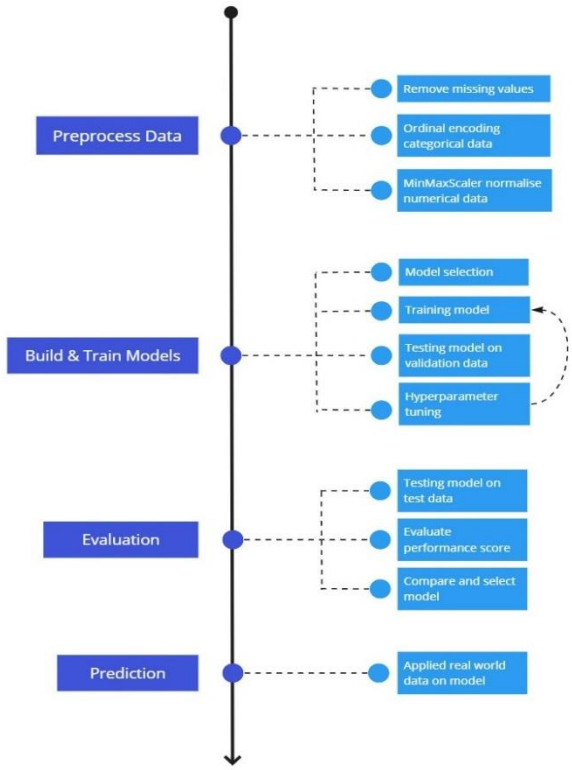


Figure. 5. Methods pipeline

our model. We had chosen Logistic Regression model, RandomForrest Classifier and DecisionTree Classifier to compare. Depend on the result of each prediction, we will choose the model with the highest accuracy score as our model, and tune the hyperparameters of the model to improve performance.

At last, we test our model with unseen test dataset to evaluate the final performance. If our score is high during training but perform poorly on testing data, this might indicate that our model is overfitting with the training data. This could happen when we have a small dataset that can not represent the overall population.

## IV. EXPERIMENTS

### A. *Preprocess data*

Within the dataset, there were 20 features in total, which 9 of them contained missing values. To maintain the integrity of the dataset, we removed features with more than 25000 missing values and the remaining instances that still contained any missing values were also removed. After removing the missing values, we were left with a complete dataset with 15 attributes (including class) and 37065 instances.

Before starting to train the models, we need to transform our data to machine readable inputs, e.g. using OrdinalEncoder or OneHotEncoder to transform nominal features to numeric. We also sometimes apply other methods for better performance, e.g. normalised or standardised numeric features to avoid the problem of different scales between features. We have applied OrdinalEncoder on nominal features and used MinMaxScaler to convert the numeric features to between 0 and 1 for our data.

After transformed the data, we split our data into training set, validation set and testing set, which we will use training set to train the models and validation set to evaluate performance and tune hyperparameters then use testing set to

evaluate the final model performance. Bear in mind that we should not use the testing set result to modify our model, this might lead to overfitting the model with the data we have.

There are 14 different features in our dataset, to reduce the dimensionality and remove less informative features. We applied SelectBest function to choose the best 4 features using Mutual Information as the score function to find out. We got a result that indicated 'gill-color', 'ring-type', 'habitat' and 'season' were the most important features among all. Now we have a much smaller dataset with informative features.

## *B.* Training Modlels

### *1)* Logistic Regression

Logistic Regression will calculate the probability of the target, e.g. whether the mushroom is more likely to be edible or poisonous. The probability will be transform into a binary value (0 or 1) in order to make prediction. Applying Logistic Regression on our data, we received an accuracy score of 57.55

### *2) Random Forest Classifier*

Random Forest classifier is a kind of ensemble learning methods which uses multiple learning algorithms to obtain better performance than alone. Just like its name, consists of a number of individual decision trees, and use majority vote or average to make the final decision based on the trees. We achieved an accuracy score of 94.93 which is much higher than the previous model.

### *3) DecisionTree Classifier*

Decision Tree has a flowchart like structure which each internal node represents a test of a feature and each leaf represents a class label. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. After training the model, we achieved an accuracy score of 93.71 which is also better compare with logistic regression model.

In table I displayed the performance of each classifier. Among all three, Random Forest classifier tend to out perform the rest. Although Decision Tree also performed well and had accuracy scores close to Random Forest, the reason we would prefer the later is that, remember we mentioned before the ultimate goal was to differentiate edible and poisonous mushrooms so people will not consume poison mushroom accidently. Under this circumstance we need to consider recall score as well, it represented the ratio of tp / (tp + fn) where tp is the number of true positives and fn is the number of false negatives. The recall is the ability of the classifier to find all the positive samples. In this case means the ability to correctly identify poisonous mushrooms, so we would prefer the classifier with higher recall score.

TABLE I. PERFORMANCE COMPARISON BETWEEN 3 CLASSIFIERS

| Methods | Logistic Regression | Random Forest | Decision Tree |
|---|---|---|---|
| Accuracy score | 57.55 | 94.93 | 93.71 |
| Precision score | 58.16 | 95.26 | 94.53 |
| Recall score | 77.11 | 95.37 | 93.81 |

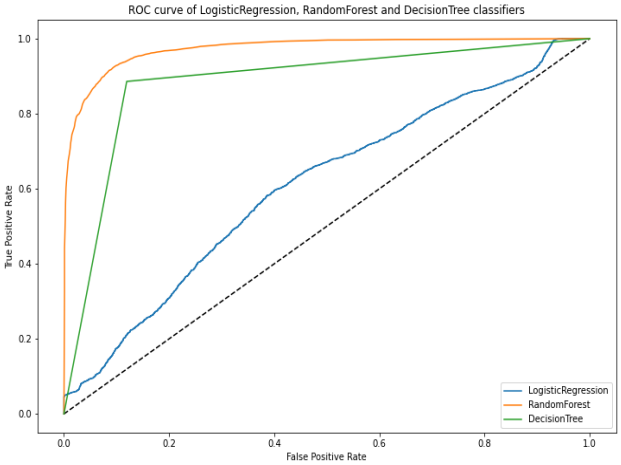


Figure. 5. ROC Curve of LogisticRegression, RandomForest and DecisionTree

## *C. Fine Tuning*

To optimise the performance of RandomForest, we will change the number of estimators by experimenting from the range 1 to 10. As we can see from Fig. 6., the accuracy had improved as the estimator number increased, but stop improving after 3 estimators, the accuracy stop around 0.94. We suspected that might be effected by the feature selection methods.

```
The accuracy of RandomForest with 1 estimators is:0.927158273381295
The accuracy of RandomForest with 2 estimators is:0.9194501541623844
The accuracy of RandomForest with 3 estimators is:0.9441161356628982
The accuracy of RandomForest with 4 estimators is:0.9402620760534429
The accuracy of RandomForest with 5 estimators is:0.9454008221993834
The accuracy of RandomForest with 6 estimators is:0.9438591983556013
The accuracy of RandomForest with 7 estimators is:0.9459146968139774
The accuracy of RandomForest with 8 estimators is:0.9464285714285714
The accuracy of RandomForest with 9 estimators is:0.9486125385405961
The accuracy of RandomForest with 10 estimators is:0.9469424460431655
```

Figure. 6. Result of RandomForest with different estimators.

Therefore, we did another experiment by using a model without feature selection to observe performance. The results are incredibly good as shown in Fig. 7. This might be an indication of overfitting that we need to be aware.

```
The accuracy of RandomForest with 1 estimators is:0.9921634121274409
The accuracy of RandomForest with 2 estimators is:0.9903648509763617
The accuracy of RandomForest with 3 estimators is:0.9970452209660843
The accuracy of RandomForest with 4 estimators is:0.9978160328879754
The accuracy of RandomForest with 5 estimators is:0.9985868448098664
The accuracy of RandomForest with 6 estimators is:0.9984583761562179
The accuracy of RandomForest with 7 estimators is:0.9988437821171634
The accuracy of RandomForest with 8 estimators is:0.998972250770812
The accuracy of RandomForest with 9 estimators is:0.9992291880781089
The accuracy of RandomForest with 10 estimators is:0.998972250770812
```

Figure. 7. Result of RandomForest with different estimators

## *D. Evaluation*

By predicting the testing dataset with the new model which set the estimators to 9 and evaluate final performance. In comparison of table I and table II, the accuracy score had drop from 94.93 to 90.54 and recall score drop from 95.37 to 91.64 which is acceptable. If the scores dropped dramatically then we might consider reviewing our training methods as this could indicate overfitting.

TABLE II. PERFORMANCE OF RANDOMFOREST ON TESTING DATA

| Methods | Accuracy score | Precision score | Recall score |
|---|---|---|---|
| Random Forest | 90.54 | 91.01 | 91.64 |

## V. REFLECTION

The final performance of our model achieved accuracy above 90% with some space to improve. We might consider to include more features during feature selection so the model can learn more information for prediction. Other methods e.g. Multi-Layer Perceptron Neural Network might allow us to include more features to perform differently. One major obstacles to apply the model in real world is that the dataset we acquired is a fairly small dataset with only 173 species, the current recorded species of mushroom is around 14,000, not to mention the species that have not been discovered yet. The appearance of the mushroom is also an issue. They do not always appear in typical shape with caps. This makes it hard to keep them in standard records. Perhaps by applying image recognition with CNN will have a more accurate result and can be apply to wider range of species.

## REFERENCES


[1] M. J. Carlile, S. C. Watkinson, G. W. Gooday, "The Fungi as a Major Group of Organisms," in *The Fungi*, 2nd Ed. Cambridge: Academic Press, 2001, pp. 1-9.

[2] M. J. Carlile, S. C. Watkinson, G. W. Gooday, "The Fungi as a Major Group of Organisms," in *The Fungi*, 2nd Ed. Cambridge: Academic Press, 2001, pp. 185-243.

[3] F. W. E. Diggins, "The true history of the discovery of penicillin, with refutation of the misinformation in the literature," *British Journal of Biomedical Science*, vol. 56, Iss. 2, pp. 83–93, Jan. 1999.

[4] Britannica, The Editors of Encyclopaedia, "Penicillin." *Encyclopedia Britannica*, https://www.britannica.com/science/penicillin (accessed Mar. 16, 2022)

[5] Accessed: Mar 18, 2022, [Online]. Available: https://discuss.boardinfinity.com/t/what-do-you-mean-by-convolutional-neural-network/8533

[6] K. Alex, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems, vol. 25, 2012.

[7] Accessed: Mar 18, 2022, [Online]. Available: https://medium.com/dataseries/basic-overview-of-convolutional-neural-network-cnn-4fcc7dbb4f17#:~:text=The%20activation%20function%20is%20a,of%20neurons%20as%20input.%E2%80%9D%20%E2%80%94

[8] Accessed: Mar 18, 2022, [Online]. Available: https://www.dataversity.net/brief-history-deep-learning/#

[9] Accessed: Mar 18, 2022, [Online]. Available: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/#:~:text=Adam%20is%20a%20replacement%20optimization,sparse%20gradients%20on%20noisy%20problems.

[10] V. Sabaratnam, K.H. Wong, M. Naidu, P. R. David " Neuronal health - can culinary and medicinal mushrooms help?" J Tradit Complement Med. pp. 62–68, Jan-March. 2013.

[11] Accessed: Mar 18, 2022, [Online]. Available: https://grocycle.com/parts-of-a-mushroom/

[12] W.E. Brandenburg, K.J. Ward. "Mushroom poisoning epidemiology in the United States." Mycologia. p.p. 637-641. Jul-Aug. 2018.

# APPENDIX

Task 1 Code
https://colab.research.google.com/drive/1YYHcEX7-oN_ilc4ctLok1DTwKIeKaCVb?usp=sharing

Task2 Code
https://colab.research.google.com/drive/17QTM_s4nuxE7Q-Bqcln5fYqLN9FawVij?usp=sharing