SDS 2019 - Group project

Applied data science and machine learning

Subject: Mushroom classification

Aalborg University - AAU

Cand.IT - IT-Management - 9th semester

Group members:

- Kristian Stavad
- Rasmus Simmelsgaard Hye
- Lasse Hede

AALBORG UNIVERSITY
DENMARK

Data Collected: This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Mushroom Family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended to consume.[1]

- Time period: Donated to UCI ML 27 April 1987

Google Colab link: https://colab.research.google.com/drive/1zilFVvHzCIu5HrQsoeg60jR2o7Kg4FDd

Github link: https://github.com/LassehedeSDS/SDS-Group-Project-KristianStavad-RasmueHye-LasseHede-

We chose to work with this mushroom dataset as we find it an interesting area to investigate further by applying data science and machine learning techniques. We all, from time to time, consume mushrooms, even though we possess limited knowledge about which are edible and which are not. Therefore we found it interesting to investigate on a broader scale, which mushrooms are edible and which are poisonous, and if it would be possible to determine this by looking at the different characteristics such as color, shape, family etc.

Preprocessing

The data analysis was done using Google Colab, which is an online notebook suitable for machine learning objectives. We choose the Mushroom Classification dataset from Kaggle.com and imported it through a url linkage. Initially all the necessary packages for the data analysis, data visualization, machine learning and plotting for this assignment was imported, which includes; Pandas, Matplotlib, NumPy, Seaborn, SciKit Learn etc.

After the initial import of the data and packages, the first initial assessment of the dataset was performed, where the aim was to look for features or variables to drop because of missing or insignificant values or observations. Here we observed that all the variables from the different observations were categorized as single letter strings instead of complete words. This deemed a problem as we could not do calculations with single letter strings. The first task was thus to

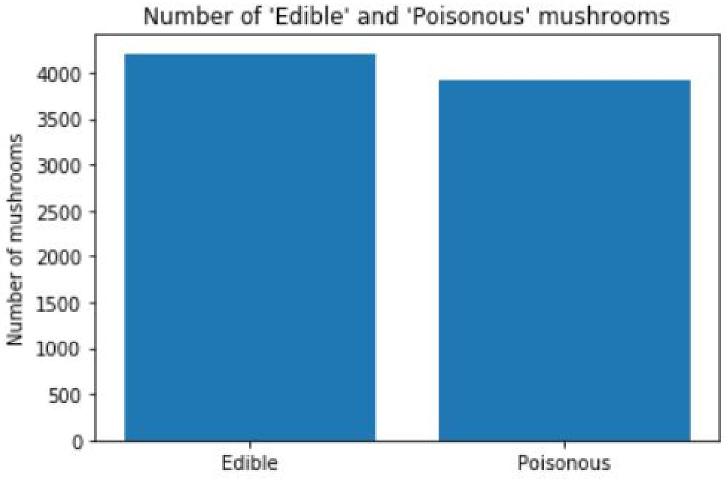
[1] https://www.kaggle.com/uciml/mushroom-classification

transform these strings into integer values with single numbers representing the beforehand stated letter or observed variable. Below you will find an example of this data transformation with the example of the variable “cap-shape”.

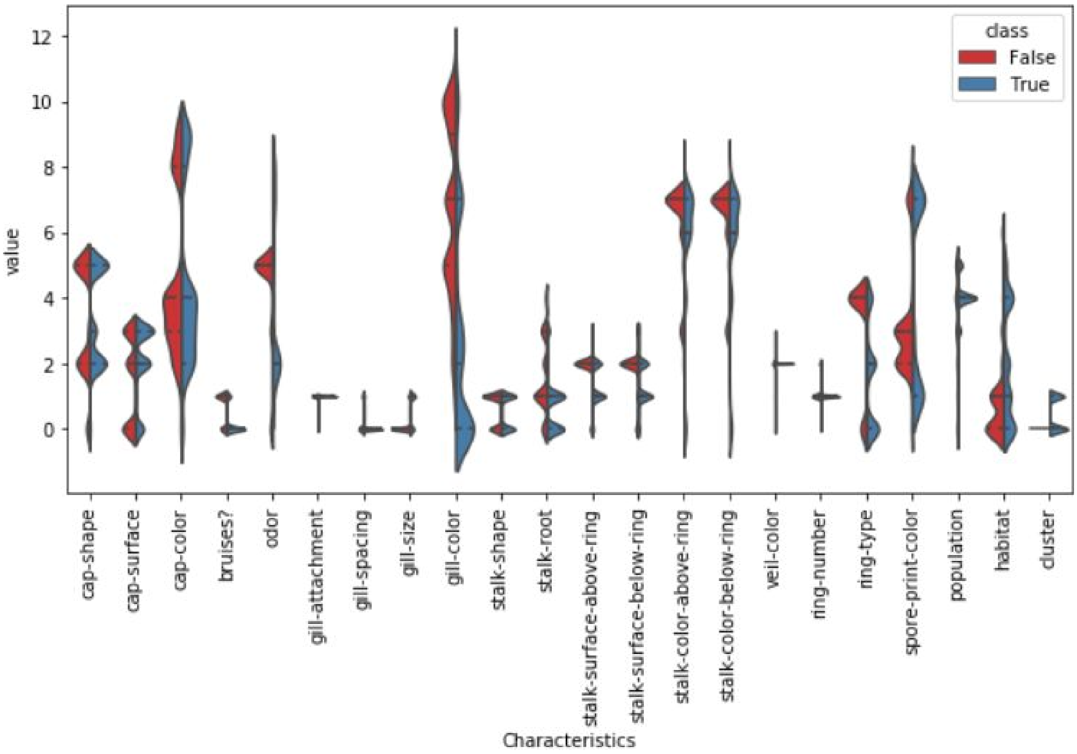
Cap-shape (old strings)	Bell = b	Conical = c	Flat = f	Knobbed = k	Sunken = s	Convex = x
Cap-shape (new integers)	0	1	2	3	4	5

After the transformation from single letter string values to integer values, we decided to change the value of our target value “class” to a boolean value with “True” for poisonous and “False” for edible. We did this to make it easier to communicate the two forms of mushroom we were trying to classify.

The overall goal of this project is to classify edible and poisonous mushrooms using different machine learning techniques, so the first interesting visualization was to visualize the total amount of different mushroom classes. To do this we stored the observations in two variables ‘poisonous’ and ‘edible’ and printed the length of each variable. The result showed that there are 3916 poisonous and 4208 edible mushrooms in the dataset which we visualized in the barchart below.

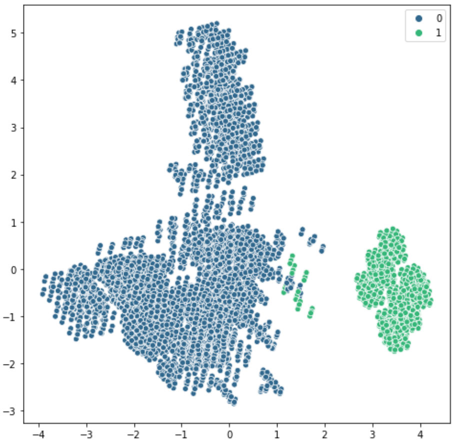


Furthermore we created a violin plot to get a visualisation of the correlation between “class” and the different attributes of the mushrooms. The violin plot suggest that there might be a correlation between “gill-color” and whether a mushroom is poisonous or edible. As you can see below there is a clear distinction between mushrooms with “gill-color” values ranging from 0 to 4, which are mainly poisonous and mushrooms with “gill-color” values from 5-12 which are mainly edible.



Unsupervised machine learning

As part of the unsupervised machine learning data structuration work, we used inertia, KMeans analysis and visualizations of the clusters found in the data to look for similarities. Finding the right amount of clusters helps with getting denser clusters, effectively meaning a more precise visualization. We chose to work with 2 clusters after reading the inertia graph, indicating that most of the observed variety can be explained within the two clusters (see figure to the right).



By investigating the characteristics of the two clusters it became clear that all observations in cluster 1 are poisonous mushrooms while cluster 0 holds observations from all edible mushrooms as well as half of the poisonous mushrooms. Meaning that for half of the poisonous mushrooms, the variance was not a determining factor in terms of classifying if it's an edible or poison mushroom. This cluster investigation was done using itertool. We can therefore conclude that further data exploration and machine learning is needed, to be able to more accurately decide the type of mushroom and we can conclude that the KMeans analysis, in our case, is not accurate enough in terms of classifying a mushroom as poisonous or edible.

Supervised machine learning

In the supervised part of this machine learning analysis we started by splitting the data into two parts, a training set containing 75 % of the data and a test set containing the last 25%. We did this in order to train the algorithms before fitting it to the actual test data. This way we can conclude whether the algorithm actually works first by training it on a part of the data and then showing it some new data to make predictions. After splitting the data we fitted three different algorithms on our test data. The first one we used was Logistic Regression which was able to predict with a 95 % accuracy. We chose Logistic Regression because it is a binary classification algorithm that works by fitting a regression line to the dataset which then works as a predictor of the probability that a new sample belong to one class or another. Next we used K-nearest neighbour (k-NN), which classifies the data based on its nearest neighbours. In this case we chose to classify based on the 5 nearest neighbours, and with this parameter k-NN was able to classify the data within a 99 % accuracy. The last model we chose was Decision Tree algorithm which is a flowchart-like tree structure. The decision tree was able to predict with a 100 % accuracy. Thus we can conclude that Decision Tree is the most accurate classifier of the three we chose with k-NN as a close second.

To evaluate and ensure the accuracy of the applied algorithms we used cross validation. Cross validation is a technique for evaluating machine learning models by splitting the data into subsets e.g. 20 % of the data and performing evaluations on each of the subsets, letting us know if the statistical analysis generalize to an independent dataset and in this case finding a score that refers to how accurate the algorithm can classify the mushrooms as either poisonous or edible. For each of the algorithms we chose a cross validation of 5. Below is a summary of the results of the cross validation scores.

Algorithm	**Mean / Accuracy**
Logistic Regression	**0.94 (94% accurate)**
KNN (K-Nearest Neighbour)	**0.99 (99% accurate)**
Decision Tree	**1.0 (100% accurate)**

Conclusion

Based on our machine learning analysis, we can conclude, that it is possible to use machine learning on mushroom classifications and accurately predict, whether a mushroom is edible or poisonous. Through our process we can also conclude, that in this case supervised machine learning is more precise than unsupervised machine learning to indicate mushroom types. Even though the 'Decision Tree' algorithm was the most precise algorithm in supervised machine learning, we must acknowledge, that there is other less precise supervised machine learning algorithms. Logistic Regression was in this case "only" able to predict 94 % of the mushrooms accurately.

Furthermore we can conclude that unsupervised and supervised machine learning can complement each other well, since they together provide an interesting insight in the dataset with different calculations and visualizations. Finally we acknowledge that since we were working on a relatively clean dataset, which didn't require much data cleaning, data cleaning is still a very important process in machine learning in order to sort out missing or faulty values as the algorithms are more precise when they work with clean data.