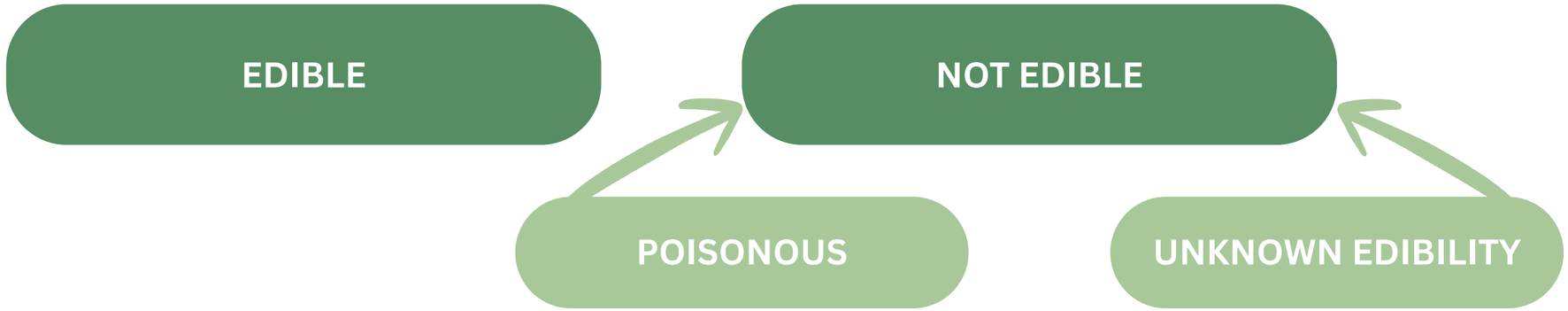
IS THIS EDIBLE?

a Machine Learning project

Kireeva Polina - Morello Gabriele Maria

INTRODUCTION

Goal: to correctly classify the edibility of mushrooms.



A dataset of 61069 hypothetical mushrooms was used for model training.

The assignment aimed to follow the CRISP-DM cycle from Data Understanding to Evaluation Phase.

RELATED WORK

Three papers were chosen to compare and confirm the work done:

Data Curation

1. The first paper is focused on data curation and how data collection influences the algorithm choice.

Algorithms

2. The second paper analyzed different types of algorithms, putting emphasis on the use of metrics to compare effectiveness.

Metrics

3. The third paper evaluated the use of different metrics with a focus on FP and FN for the effects of wrongly classified mushrooms.

PROPOSED METHOD

Data understanding

Data Profiling

```
Data columns (total 21 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   class                  61069 non-null  object
 1   cap-diameter           61069 non-null  float64
 2   cap-shape              61069 non-null  object
 3   cap-surface            46949 non-null  object
 4   cap-color              61069 non-null  object
 5   does-bruise-or-bleed   61069 non-null  object
 6   gill-attachment        51185 non-null  object
 7   gill-spacing           36006 non-null  object
 8   gill-color             61069 non-null  object
 9   stem-height            61069 non-null  float64
 10  stem-width             61069 non-null  float64
 11  stem-root              9531 non-null   object
 12  stem-surface           22945 non-null  object
 13  stem-color             61069 non-null  object
 14  veil-type              3177 non-null   object
 15  veil-color             7413 non-null   object
 16  has-ring               61069 non-null  object
 17  ring-type              58598 non-null  object
 18  spore-print-color      6354 non-null   object
 19  habitat                61069 non-null  object
 20  season                 61069 non-null  object
dtypes: float64(3), object(18)
memory usage: 9.8+ MB
```

Solved problems regarding **data extraction** from a .csv file.

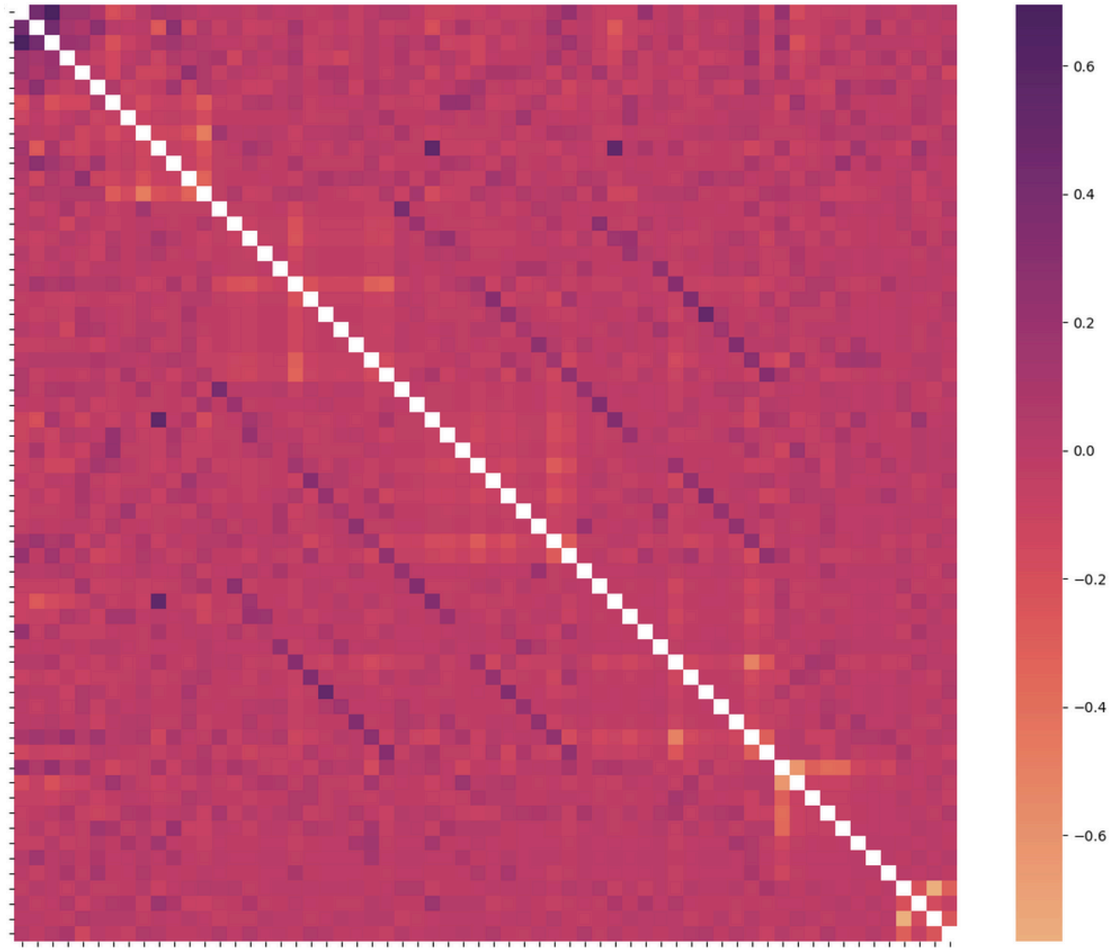
All the features with predominantly **null values** were removed.

Features having letters representing values may cause problems for the modeling phase, so **one-hot encoding** was used.

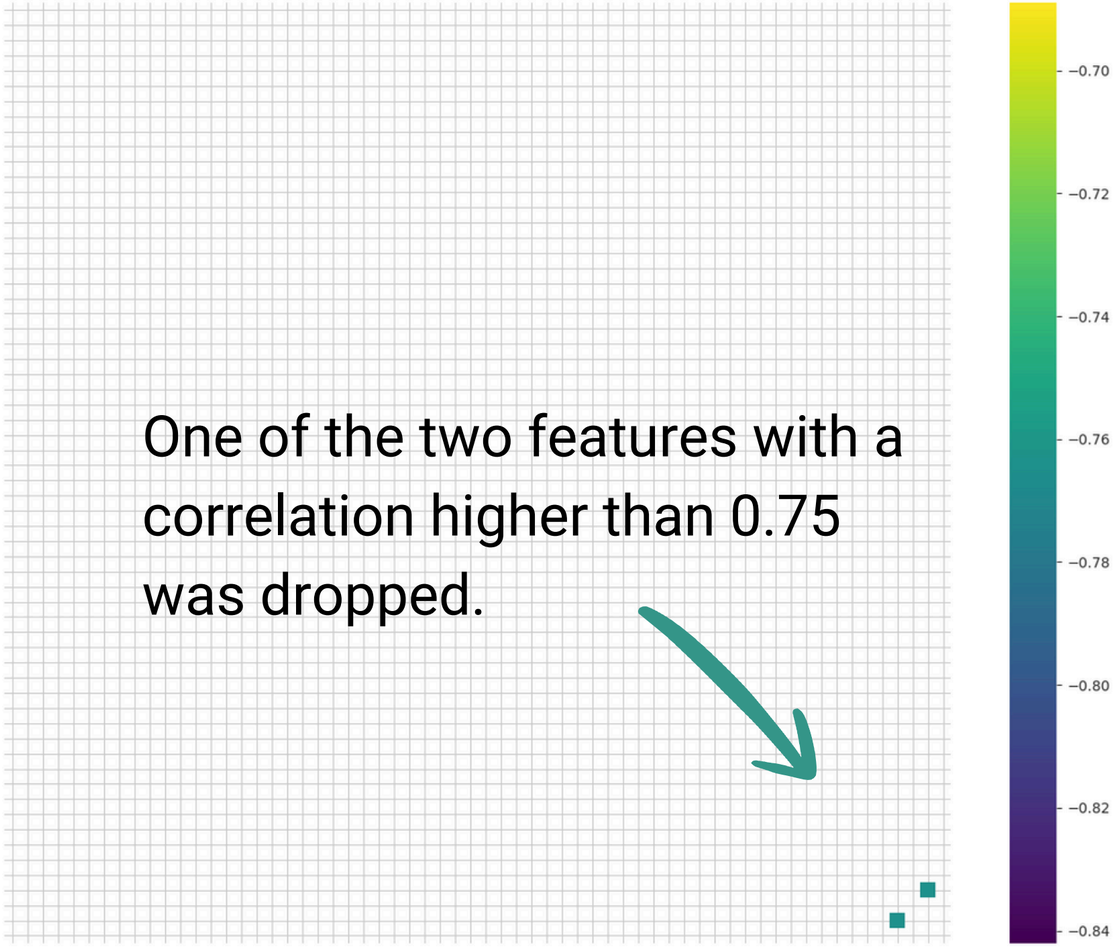
Numerical features had different scales so a **normalization** was needed.

Data understanding

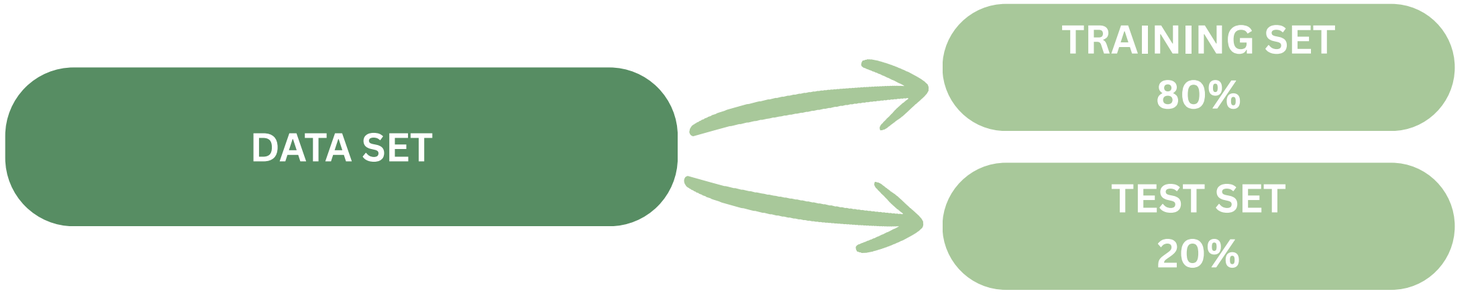
Features may be redundant for the training so a correlation heatmap was applied and a graphical representation helped evaluate them.



One of the two features with a correlation higher than 0.75 was dropped.



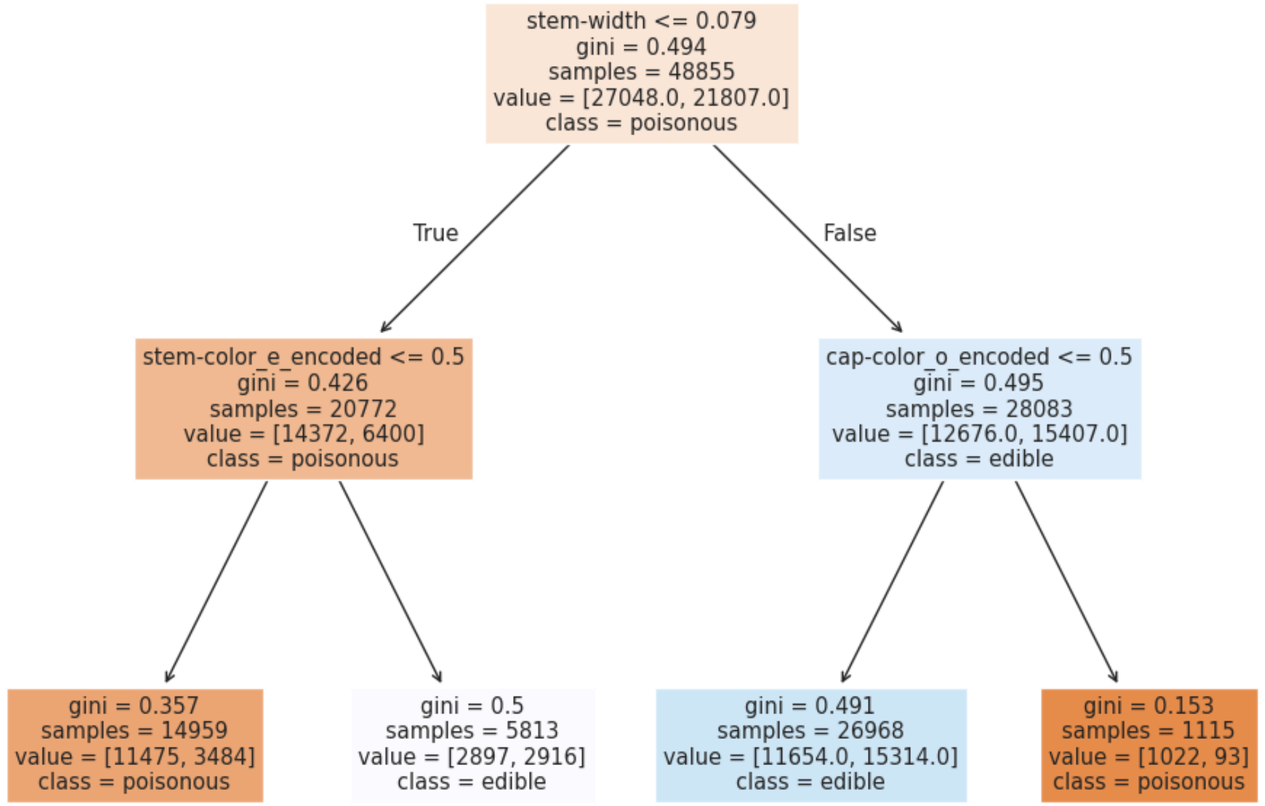
Modeling and **Evaluation**



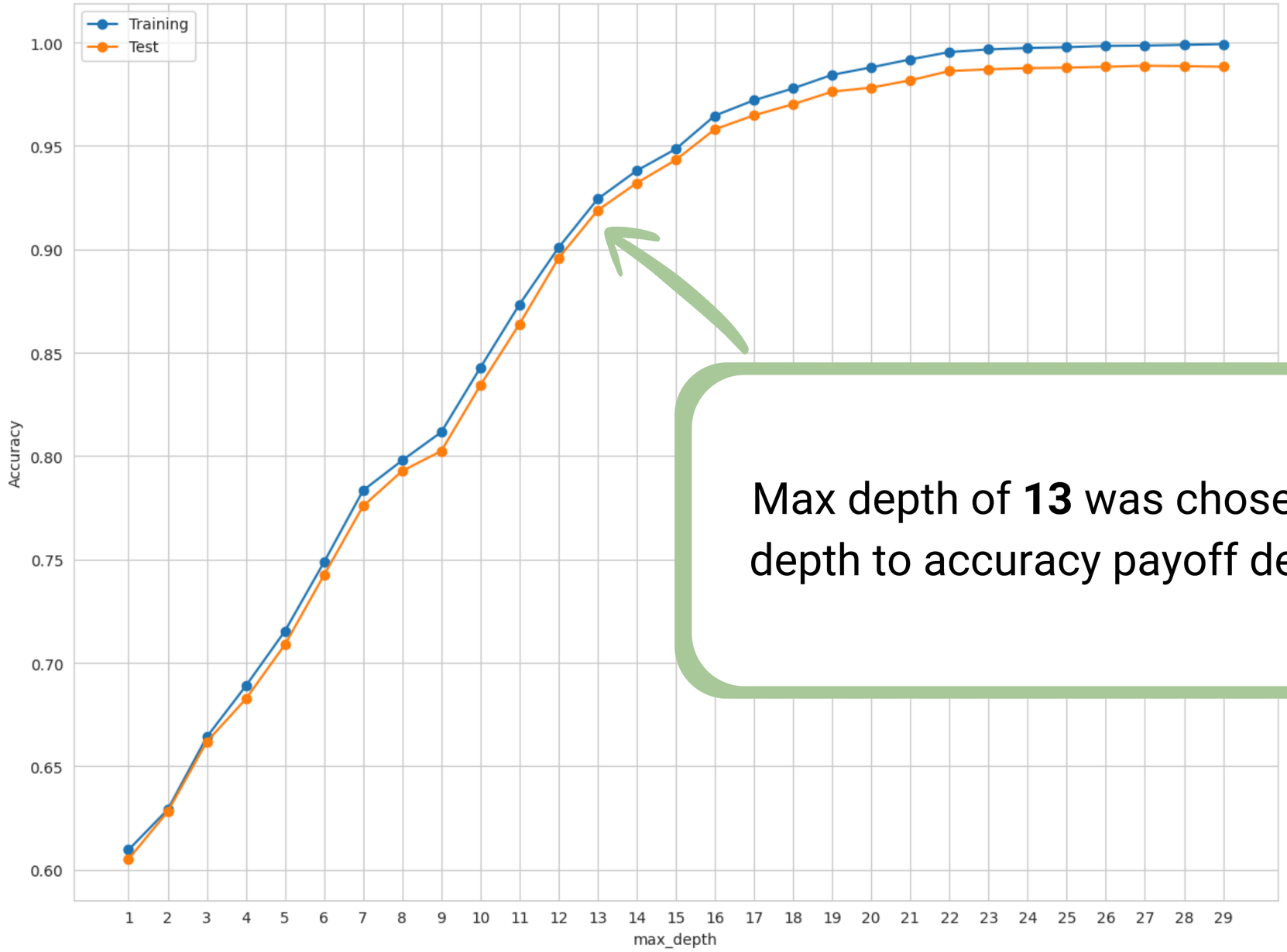
A shallow decision tree as a baseline of a simple algorithm

Accuracy:

62.8%



Modeling and **Evaluation**



Max depth of **13** was chosen as the depth to accuracy payoff decreases

RESULTS

Decision Tree

depth = 13

91.9%

Random Forest

99.6%

AutoML

budget = 60s

XGBClassifier

99.5%

After **feature importance** analysis the features with 0% importance for the Decision Tree were removed from the dataset, and the results of all models did not show a reduction in accuracy.

CONCLUSION

The solution provided for this assignment was the use of tree classifiers and the implementation of AutoML supported the choice of algorithm type to train the model.

Future Improvements

Evaluating other techniques for nominal features.

Providing an analysis on FP and FN.

Further feature selection and hyperparameters analysis.