Capstone 2: supervised learning

Mushroom classification model

Dataset & research question

This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like "leaflets three, let it be" for Poisonous Oak and Ivy.

From this dataset we aim to determine the following:

- **Can mushroom samples be correctly identified as edible or poisonous based on a set of physical attributes?**

Model specification & results

For this analysis, five classification models were fit to the data. **K-fold cross validation** was used to assess the accuracy of each model. For the random forest and KNN classifiers, **hyperparameter tuning** was done to determine the optimal number of estimators and neighbors, respectively. The models compared for this analysis and their respective accuracies can be found below:

Rank	Model	Accuracy
1	Random forest classifier, n_estimators = 6	95.8%
2	Gradient boosting classifier	95.1%
3	Support vector classifier	94.4%
4	KNN classifier, n_neighbors = 2	93.1%
5	Logistic regression classifier	91.9%

Practical uses & considerations

For our purposes, the best model was the **random forest classifier** with an accuracy of **95.8%**. Ultimately the top factor in selecting the best model was accuracy, since misidentifying a poisonous mushroom could be a fatal error.

- This model could be used by mushroom collectors to determine whether the mushrooms they encounter in the field are edible or poisonous.
- This can mean the difference between survival and death in certain instances, so it is important that the model has the highest possible accuracy.
- Since the model would likely not be usable out in the field, it would be less useful to survivalists and more useful for individuals collecting the mushrooms and later determining whether or not they are edible. That is, unless a portable application of the model could be created to test samples in the field.

Weak points & further analysis

This model does not have 100% accuracy so it is not a foolproof method to determine whether or not a collected mushroom is edible. Further hyperparameter tuning could be done to improve the accuracy closer to 100%, but that is beyond the scope of this analysis. For the purposes of our analysis, an accuracy of 95.8% is acceptable.

Additionally, we included all 22 categorical variables as features in the model. Further analysis could be done to determine which features could be dropped without sacrificing accuracy, in order to improve the performance of the model.

Additional information

- Link to full analysis can be found here.
- Link to dataset: https://www.kaggle.com/uciml/mushroom-classification
- UCI source: https://archive.ics.uci.edu/ml/datasets/Mushroom