Mushroom Classification Using K-Nearest Neighbors (KNN), Support Vector Machine, and Random Forest Algorithms

Atahan Çaldır, Musa Berkay Kocabaşoğlu, Kerem Ersan

1. Problem Description

There are countless types of mushrooms in the world. The biggest problem with mushrooms to date is which mushrooms are edible and which are poisonous.The aim of our project is to create a model that shows which mushrooms are edible and which are poisonous by analyzing some properties of edible and poisonous mushrooms, ultimately, using this model to discover whether an unknown mushroom is edible or not.

2. Solution Method

Using the mushroom dataset with the Python software language and performing operations in the Jupyter notebook, this project will be created using Numpy, Pandas, Sckit-learn, Matplotlib libraries.

In order to get to know the dataset, different characteristics of mushrooms will be examined in this project. 22 different features were used from the dataset provided to examine the relationships between these independent variables and the dependent variable, which is whether the mushroom is edible or poisonous. It will then be determined whether the dataset contains missing data in order to create a usable dataset for the study. In case of finding missing data, we will decide to remove missing instances from our dataset or apply interpolation techniques. Then, the features will be divided into numerical and categorical for reviewing dataset.

After the dataset review, we will split our data 3 parts for training, test and validation. We will use K-nearest neighbor, support vector machine and random forest algorithms for training data. Each algorithm will be tested with different values of its hyper parameters and the most performing version will be used. For example, to set efficient K-nearest neighbor algorithm, different metrics will be used in distance calculation, different k values will be determined. Finally, the best performed version of the 3 algorithms will be compared with their confusion matrices and the most optimized model will be selected.

3. Dataset

Mushroom 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. [1]

4. Algorithms

In this project, K-Nearest Neighbors, Support Vector Machine, and Rondom Forest algorithms will be used for the classification of the mushrooms.

4.1. K-Nearest Neighbors

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. [2]

4.2. Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. [3] In our project, we will use it for classification.

4.3. Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. [4]

5. References

[1] *Mushroom data set.* (1987). UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Mushroom

[2] *K-nearest neighbors algorithm.* (n.d.). IBM. Retrieved from https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20algorithm%2C%20also%20known%20as%20KNN%20or,of%20an%20individual%20data%20point.

[3] *Support vector machine algorithm.* (n.d.). Javatpoint. Retrieved from https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[4] *Random forest classifier.* (n.d.). Scikit Learn. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html