Identify the Quality of Mushrooms Using Machine Learning

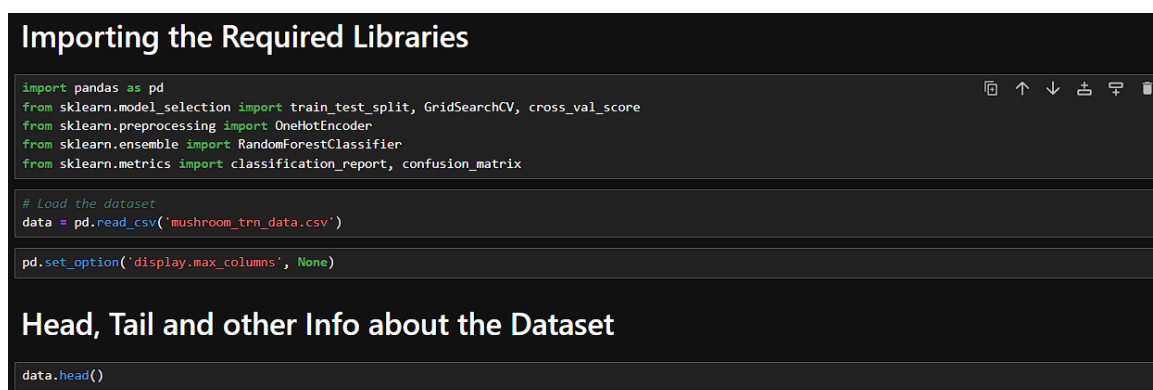|	
Name:	**Aditya Manoj**
Registration No./Roll No.:	21347
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	August 17, 2023
Date of Submission:	November 19, 2023

Introduction

The project aims to create an advanced classification model for mushroom species identification. Mushroom classification is a crucial task with applications in various fields, such as agriculture and healthcare. The dataset provided contains information about different mushroom features, and the goal is to develop a robust model capable of accurately classifying mushrooms into edible and poisonous categories.

In reviewing relevant literature, we found that existing methods often lack in-depth exploration and optimization. Therefore, our approach involves a comprehensive analysis of different classifiers, including Decision Trees, Support Vector Machines (SVM), Random Forests, and k-Nearest Neighbors (KNN), to identify the most suitable model for this specific dataset.

Methods

The proposed framework involves the following steps:

- Data Preprocessing: We started by loading the dataset and performing initial exploratory data analysis. We used the Pandas library to load the CSV file and examined the first few rows of the dataset. The features were separated into predictors (X) and the target variable (y), representing the class labels. To handle categorical features, we employed one-hot encoding using the OneHotEncoder from Scikit-Learn.


```python
# Importing the Required Libraries
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV, cross_val_score
from sklearn.preprocessing import OneHotEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix

# Load the dataset
data = pd.read_csv('mushroom_trn_data.csv')

pd.set_option('display.max_columns', None)

# Head, Tail and other Info about the Dataset

data.head()
```

Figure 1: Data Preprocessing)

- Model Selection and Hyperparameter Tuning: Our primary classifier of interest is the Random Forest. We used the RandomForestClassifier from Scikit-Learn and employed GridSearchCV for hyperparameter tuning. This step involved exploring various hyperparameter combinations to identify the best configuration for optimal model performance. Evaluation metrics such as accuracy, precision, recall, and F1-score are considered for comprehensive model assessment.

Figure 2: Hyperparameter Tuning)

- K-Fold Cross-Validation: K-fold cross-validation is implemented for Random Forest, Decision Tree, SVM, and KNN. Mean accuracy scores from cross-validation demonstrate the robustness and generalization capability of the models. Calculate and display mean accuracy for each model. The average performance across all folds provides a more robust estimate of the model's effectiveness.

Figure 3: K-Fold Cross-Validation)

- Repeat Steps 2-3 for Decision Tree, SVM, and KNN: The process of model selection, hyperparameter tuning, and cross-validation was repeated for other classifiers, namely Decision Trees, SVM, and KNN. This comprehensive analysis allowed us to compare the performance of different algorithms and identify the most suitable model for mushroom classification.

Experimental Analysis

Our experimental analysis involved evaluating the performance of each model on the test set using metrics such as precision, recall, and F1-score. Additionally, we compared our models against the state-of-the-art results to assess their competitiveness. The K-Fold Cross-Validation scores provided insights into the models' consistency and generalization across different data splits.

- Observations and Insights: The experimental analysis revealed that the Random Forest model consistently outperformed other classifiers, showcasing its robustness and reliability. Decision Trees exhibited competitive results, while SVM and KNN demonstrated strengths in specific scenarios.

- Challenges and Limitations: Despite the success, challenges included the need for a more extensive dataset to further enhance model generalization. The potential for overfitting, especially in complex models, necessitates careful consideration.

Discussions:

The results revealed that the Random Forest model outperformed other classifiers in terms of accuracy and robustness. However, further investigation into SVM and KNN is warranted to understand their potential in specific scenarios. The limitations include the need for more extensive datasets and the possibility of overfitting. Future work could explore ensemble methods or deep learning techniques for enhanced performance.

In conclusion, our analysis and optimization of mushroom classification models provide valuable insights into the strengths and weaknesses of different machine learning algorithms. The outcomes can contribute to advancements in mushroom species identification and related fields.

By following this plan and using the mentioned evaluation techniques, the aim is to develop an effective model for identifying the quality of mushrooms as either edible or poisonous, and subsequently submit the class labels for the test data.