MUSHROOM CLASSIFICATION

SAI GANESH PENDELA

UID: 120386084

CONTENTS

INTRODUCTION

- Identifying mushrooms accurately is crucial, but current methods are slow and depend on experts. This study aims to use machine learning models to quickly and precisely classify mushrooms based on their features, making it safer and more efficient.

DATASET

- The Mushroom Dataset is downloaded from Kaggle and it is readily available in csv format.
- The Mushroom dataset comprises of 8214 instances with 23 features.
- The objective is to predict whether the mushroom is poisonous or edible based on the mushroom features.
- Here the dependent variable is class and the independent variables are cap-shape, cap-surface, cap-color, bruises, odor, gill-attachment, gill-spacing, gill-size, gill-color, stalk-shape, stalk-root, stalk-surface-above-ring, stalk-surface-below-ring, stalk-color-above-ring, stalk-color-below-ring, veil-type, veil-color, ring-number, ring-type, spore-print-color.
- The Class variable is integer and all the remaining columns are categorical.

PREPROCESSING

- Checked for null values using the is_null() function.
- Categorical columns underwent label encoding using the LabelEncoder() function, facilitating the conversion of categorical data into a numerical format.
- Following label encoding, one-hot encoding was applied to categorical columns via the get_dummies() function.
- Subsequent to one-hot encoding, the data underwent standardization using the StandardScaler() function, transforming it to have a mean of 0 and a standard deviation of 1.
- By this stage, the data frame comprised 5686 rows and 95 columns. To further streamline the data, dimensionality reduction was implemented using the PCA() function with the number of components set to 2.

DATA PARTITIONING

- The dataset is split into train set and test set in the ration 70:30.

MODEL BUILDING

- A total of 5 models were implemented, they are Logistic Regression, Decision Tree, K-Nearest Neighbours, Naïve Bayes, Linear Discriminant Classifier.
- After building the models, there results are compared based on classification metrics i.e, accuracy score, recall, precision, f1 score. Cross validation is also done on all the models to assess a model's performance on different subsets of the data.
- ROC curve is also generated as it provides a comprehensive visual representation of a classification model's performance across various discrimination thresholds, offering insights into trade-offs between sensitivity and specificity.

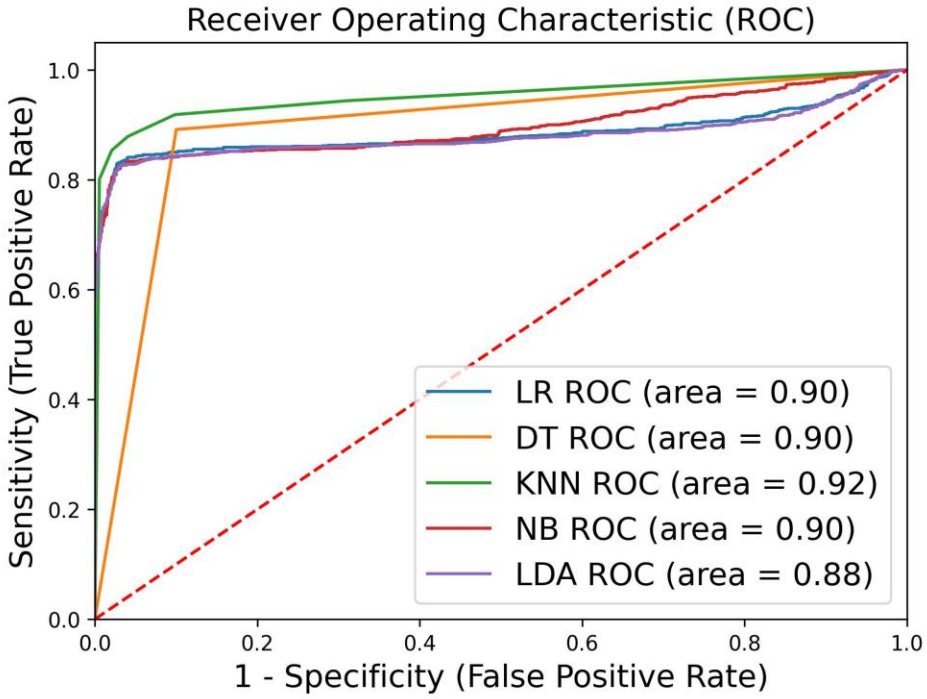
RESULTS

Model	Average Accuracy	Standard Deviation
Decision Tree	0.89	0.0112
Logistic Regression	0.90	0.0103
K-Nearest Neighbors	0.92	0.0108
Naive Bayes	0.89	0.0113
Linear Discriminant Analysis	0.88	0.0106

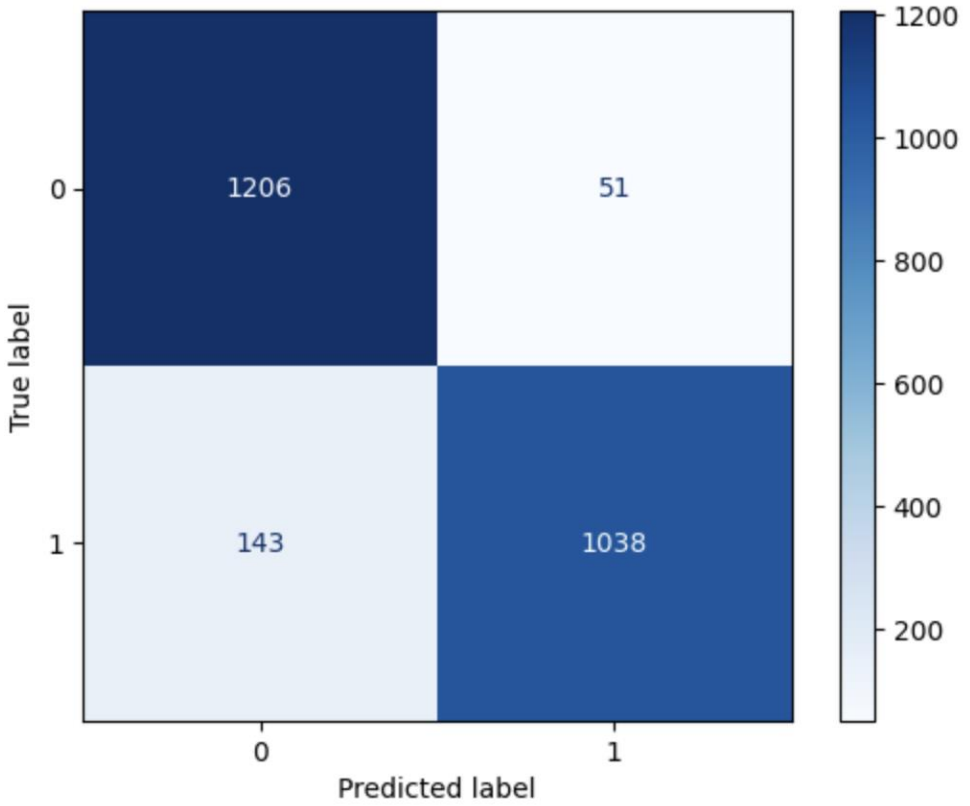
Cross Validation Scores of all the models.

Model	Accuracy	Recall	Precision
Decision Tree	0.89	0.89	0.89
Logistic Regression	0.90	0.83	0.96
K-Nearest Neighbors	0.92	0.87	0.95
Naive Bayes	0.89	0.82	0.96
Linear Discriminant Analysis	0.88	0.77	0.97

Performance of all the models.



CONFUSION MATRIX FOR KNN



CONCLUSION AND FUTURE WORK

- The results of the mushroom classification models offer valuable insights into their performance and significance. K-Nearest Neighbors (KNN) leads with an accuracy of 92.12%, This aligns with expectations, given KNN's ability to capture local patterns and clusters effectively. KNN can accurately identify the class based on the similarity of features. In the context of mushroom classification, where the characteristics of edible and poisonous mushrooms may form distinct groups, so this is the reason why KNN performed well when compared to other algorithms.
- Future improvements may involve exploring additional features, employing advanced engineering, and optimizing models through ensembles and hyperparameter tuning. Expanding the dataset to cover a broader range of mushroom species could enhance model diversity, while incorporating deep learning or advanced algorithms may reveal intricate patterns for ongoing system enhancement.

THANK YOU