Analysis of Bayesian Networks, DT Classifiers, and RF on the Mushroom Dataset

In this study, we explore the use of Bayesian Networks and Decision Tree Classifiers along with Random Forests on the Mushroom dataset. The Mushroom dataset contains 22 categorical features describing different characteristics of various mushroom species to classify whether they are edible or poisonous.to achieve the goal, we design a machine learning pipeline that has Data preprocessing steps, including handling missing values and dimensionality reduction, are implemented to optimize the performance of these models.

Data Preprocessing

The stalk-root feature in the Mushroom dataset contains missing values, as shown in Figure 2. We replace these missing values with the most frequent value (mode) in their respective columns. This approach ensures that the dataset remains consistent and complete for model training. All features in the Mushroom dataset, including the label colom are categorical. We convert these categorical values into numerical values using Label Encoding. This transformation is essential for algorithms like Decision Trees and Random Forests, which require numerical input. The dataset is almost balanced. Both classes, edible represented as e and poisonous represented as p, are a nearly equal number of instances, as shown in Figure 1.

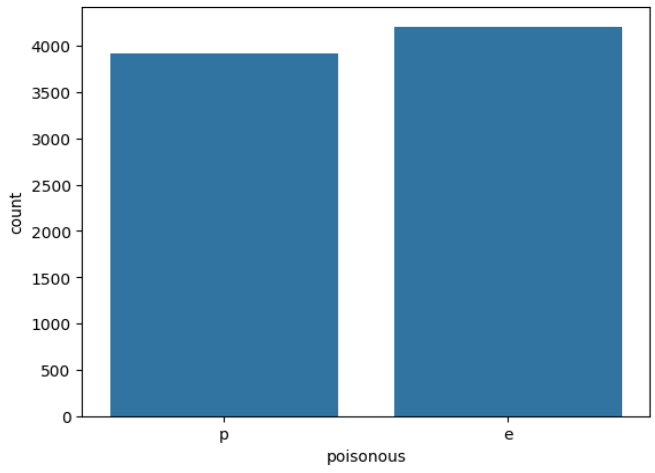

Figure 1 Data Distribution Of Mushroom Data

#	Column	Non-Null Count	Dtype
0	cap-shape	8124 non-null	object
1	cap-surface	8124 non-null	object
2	cap-color	8124 non-null	object
3	bruises	8124 non-null	object
4	odor	8124 non-null	object
5	gill-attachment	8124 non-null	object
6	gill-spacing	8124 non-null	object
7	gill-size	8124 non-null	object
8	gill-color	8124 non-null	object
9	stalk-shape	8124 non-null	object
10	stalk-root	5644 non-null	object
11	stalk-surface-above-ring	8124 non-null	object
12	stalk-surface-below-ring	8124 non-null	object
13	stalk-color-above-ring	8124 non-null	object
14	stalk-color-below-ring	8124 non-null	object
15	veil-type	8124 non-null	object
16	veil-color	8124 non-null	object
17	ring-number	8124 non-null	object
18	ring-type	8124 non-null	object
19	spore-print-color	8124 non-null	object
20	population	8124 non-null	object
21	habitat	8124 non-null	object

Figure 2 Features in the Mushroom dataset

Dimensionality Reduction using PCA

The dataset has the highest dimensionality with 22 features. We reduce the dimensionality of the dataset using Principal Component Analysis (PCA). PCA helps to reduce the number of features while retaining the most important information from the original dataset. We select the top 5 principal components for training the machine learning models. The importance of the features is represented in Figure 3.

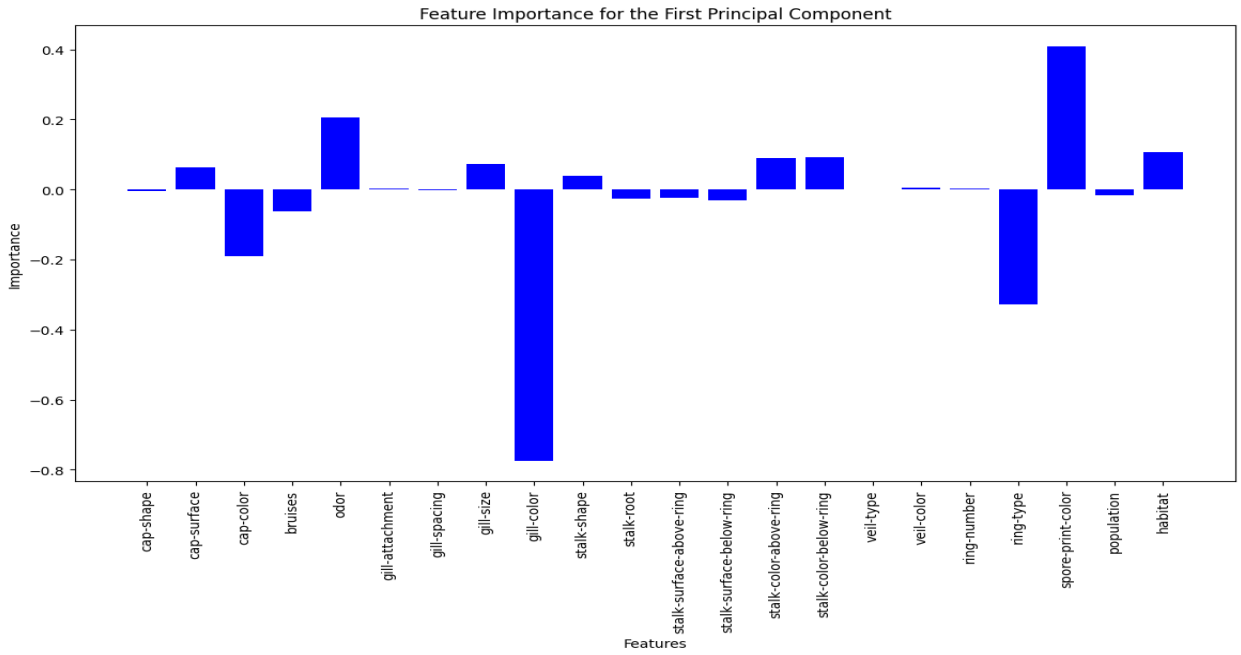

Figure 3 Feature importance extracted using PCA.

We trained a Decision Tree classifier on the reduced dataset, taking advantage of its simplicity and interpretability. The Decision Tree model splits the data based on feature values, creating a tree-like structure that is easy to visualize and understand. Additionally, we employed a Random Forest classifier, an ensemble method that builds multiple Decision Trees and merges their results to improve accuracy and robustness. By averaging the results of multiple trees, Random Forest reduces overfitting and enhances generalization. For Bayesian Networks, we utilized HillClimbSearch to automatically select the network structure, optimizing it to fit the data best. This probabilistic graphical model captures complex feature dependencies, providing strong performance metrics.

We evaluated the performance of the Decision Tree(DT), Random Forest(RF), and Bayesian Network(BN) classifiers on the Mushroom dataset. Each model demonstrated excellent performance, achieving an accuracy of 99% as shown in Table 1. The precision, recall, and F1 scores for all models were also perfect, at 1.0 (or 100%) for both classes, indicating that the models were highly effective in distinguishing between edible and poisonous mushrooms.

Table 1Result of the Machine Learning Model

Model	Accuracy	Precision	Recall	F1 Score
DT	0.996	1.00	1.00	1.00
RF	0.996	1.00	1.00	1.00
BN	0.996	1.00	1.00	1.00