CSE802 Project Report: Mushroom Classification

Erika Zheng, Junchi Zhu

April 25, 2024

1 Introduction

Mushroom hunting has become a popular activity in the recent years, and it is especially important to distinguish between poisonous and non-poisonous mushrooms. Our goal is to utilize the UCI Mushroom Classification dataset [1] to develop a machine learning model that is able to determine if a mushroom is edible based on its features. The project will also address the imputation of any missing values in the dataset.

2 Dataset

The dataset includes descriptions of 8,124 mushroom samples, each characterized by 23 attributes such as cap shape, cap surface, cap color, presence of bruises, odor, and gill attachment. These details help to identify whether mushrooms are poisonous or edible. Each record pertains to a species from the Agaricus and Lepiota families, labeled as edible, poisonous, or uncertain edibility. For safety, species whose edibility is unknown are classified as poisonous.

3 Methodology

We approach this project by addressing two main steps: handling missing data and classifying the mushrooms based on their attributes.

3.1 Impute Missing Data

To address the challenge of handling missing data when the model is used in the future, we introduced missing values artificially to 20% of the dataset. We then employ two primary strategies for imputation: mode imputation and K-Nearest Neighbors (KNN) imputation.

Mode Imputation The mode imputation method involves replacing missing values with the most frequently occurring value in each column. This approach is particularly effective for categorical data as it preserves the most common category within each feature.

KNN Imputation The KNN imputation technique utilizes the k-nearest neighbors algorithm to estimate and replace missing values. Each missing value is imputed using the mean value from its nearest neighbors, with the distance between samples calculated using present values. This method considers the underlying patterns in the data, potentially leading to more accurate imputations for complex interactions between features.

3.2 Classification

A variety of machine learning models were employed to assess their performance. The models are trained on 80% training and 20% testing data

Machine Learning Models The models used include Logistic Regression, Support Vector Machines (SVM), Gradient Boosting Classifiers, Random Forests, Naive Bayes, Decision Trees, and K-Nearest Neighbors. Each model was trained and evaluated on the original dataset and the imputed datasets Additionally, a combined approach was utilized where a voting classifier integrated predictions from all individual models.

Deep Learning Models Furthermore, a neural network model was designed, consisting of three linear layers with ReLU activation functions.

4 Results

4.1 Impute Missing Data

We compare the feature correlations of the imputed data with those of the original dataset, as illustrated in Figure 1. While there are some discrepancies between the original and imputed datasets, the imputed data from the two imputation methods show no significant differences.

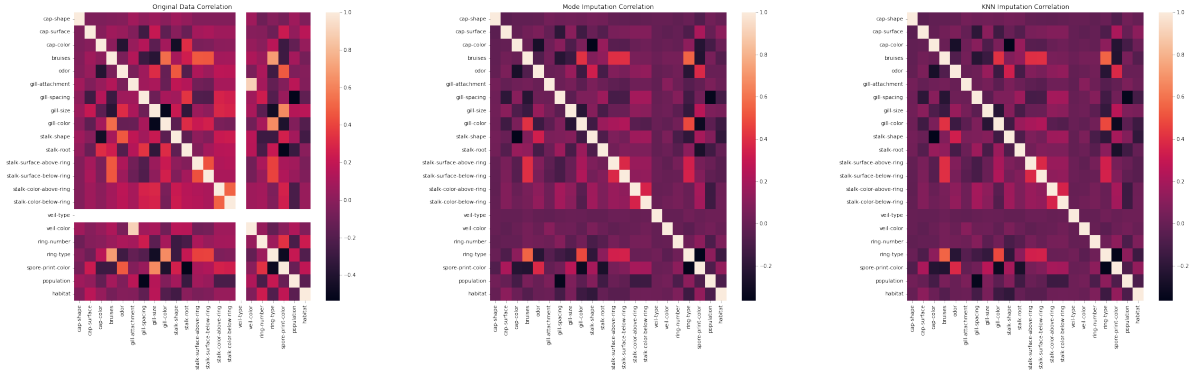

Figure 1: Correlation Graphs

To further validate the data imputation method, we utilized the Random Forest Classifier to assess feature importance across the two methods and the original data.

The feature importance rankings remained consistent across all methods as illustrated in Figure 2.

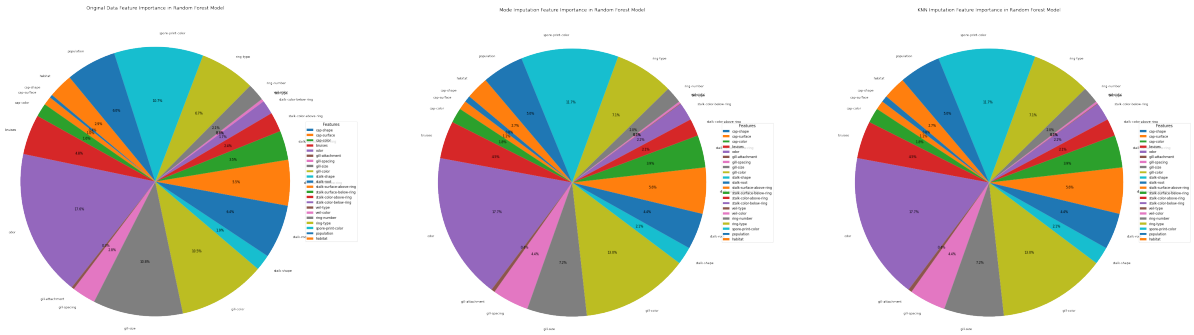

Figure 2: Feature Importance Pie Chart

4.2 Classification

4.2.1 Machine Learning Models

The results in Table 1 shown that the machine learning models generally has high accuracy across all data conditions, with Gradient Boosting, Random Forest, and Decision Tree models consistently achieving near-perfect or perfect scores. The performance slightly decreases in the datasets where missing values were imputed, which suggests that while imputation has low impact on accuracy and the models remain robust.

Model	Original Data (%)	Mode Imputed (%)	KNN Imputed (%)
Logistic Regression	94.77	86.71	86.71
SVM	99.26	96.62	96.62
Gradient Boosting	100.00	99.08	99.08
Random Forest	100.00	99.82	99.82
Naive Bayes	92.18	88.49	88.49
Decision Tree	100.00	98.58	98.58
K-Nearest Neighbors	99.63	97.05	97.05

Table 1: Accuracy of Machine Learning Models Across Different Data Preparations

4.2.2 Deep Learning Models

The neural network has achieve 100.00%, 97.66% and 97.70% testing accuracy on the original dataset, mode imputed and KNN imputed dataset respectively. Detailed insights into the training loss and training accuracy throughout the model's development are shown in Figure 3.

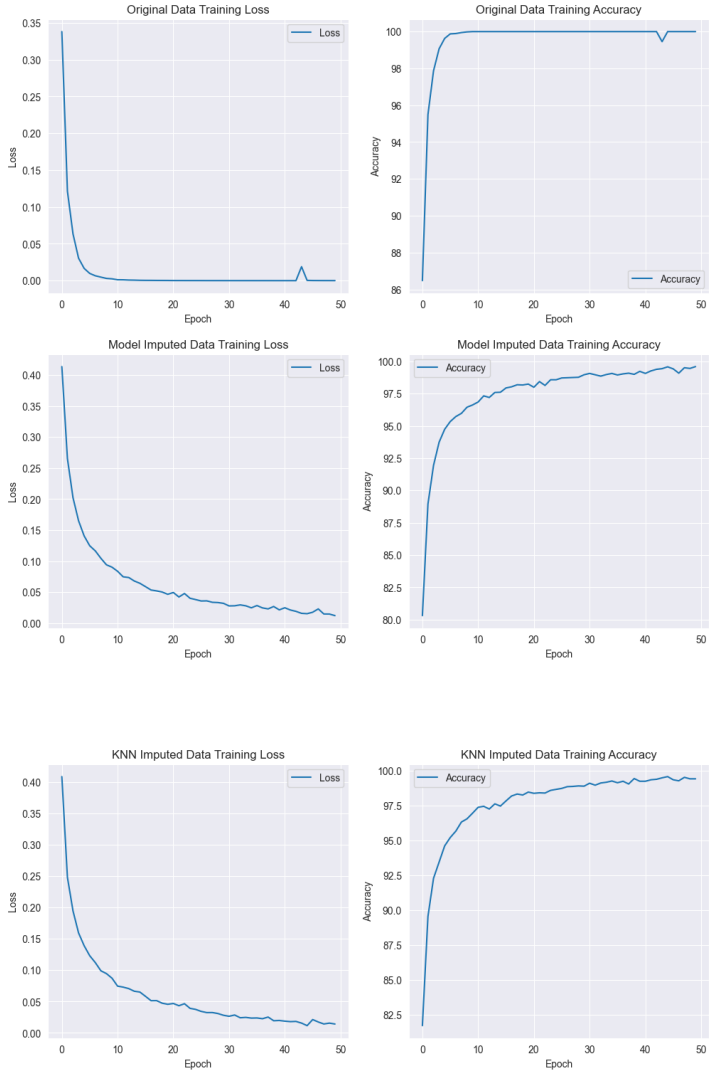

Figure 3: Training Loss and Training Accuracy

5 Summary

This project used various machine learning and deep learning models to accurately classify mushrooms as edible or poisonous based on their physical attributes, achieving high accuracy rates across original and imputed datasets. The imputation methods also shown reliability in the data created. This confirms the project's potential utility in real-world applications.

6 References

1. Mushroom. (1987). UCI Machine Learning Repository. https://doi.org/10.24432/C5959T.