Radhika Puri
CPSC 393
Prof. Ali
22 September 2025

Assignment 02 Technical Report

The objective of this project was to classify mushroom samples as edible or poisonous using a deep Feed-Forward Neural Network (FFNN) and compare its performance with a simpler baseline model, Logistic Regression. The purpose was to evaluate whether the complexity of a neural network provides a meaningful improvement over a traditional linear model.
Out of the three options for this assignment, the mushroom dataset was chosen. I really hate mushrooms as a food, but still like mycology, so this was the natural choice. The mushroom dataset comprises of twenty categorical and continuous features that describe characteristics such as cap shape, color, surface texture, stem properties, and habitat. The target variable ("class") consists of two categories: e (edible) and p (poisonous), where the latter includes mushrooms of unknown edibility.

Analysis

Before modeling, the dataset was cleaned by removing columns with more than 15,000 missing values and dropping rows that still contained missing data. The class column was encoded numerically (e = 1, p = 0) to support binary classification.

A review of category proportions revealed that 54.3% of mushrooms were labeled poisonous and 45.7% edible, indicating a fairly balanced dataset. Among other variables, several categories showed clear dominance.
Cap shape: convex (x) was the most common (48%), followed by flat (f, 23%).
Cap color: brown (n) dominated at 41%, followed by yellow and white (~12% each).
Does-bruise-or-bleed: 79% of samples did not bruise or bleed.
Season: the majority (49%) were collected in autumn, followed by summer (38%).
These proportions suggest certain biological traits (like convex brown caps and woodland habitats) are strongly represented. This slight imbalance in categorical frequency, however, does not appear to bias the target classes significantly.

All categorical predictors were one-hot encoded to convert them into numerical features suitable for model training. Continuous variables, such as cap diameter and stem height, were standardized using Z-score scaling to normalize their ranges. The dataset was then split 80/20 into training and testing sets.

Methodology

A Feed-Forward Neural Network was developed in TensorFlow/Keras to classify mushrooms. The architecture contained three hidden layers, providing enough capacity to capture complex nonlinear relationships in the dataset.

Architecture:
Input Layer: corresponds to all encoded and scaled features.
Hidden Layer 1: 128 neurons, ReLU activation, with Dropout(0.3) for regularization.
Hidden Layer 2: 64 neurons, ReLU activation, Dropout(0.3).
Hidden Layer 3: 32 neurons, ReLU activation.
Output Layer: 1 neuron with sigmoid activation to output class probabilities (edible vs. poisonous).

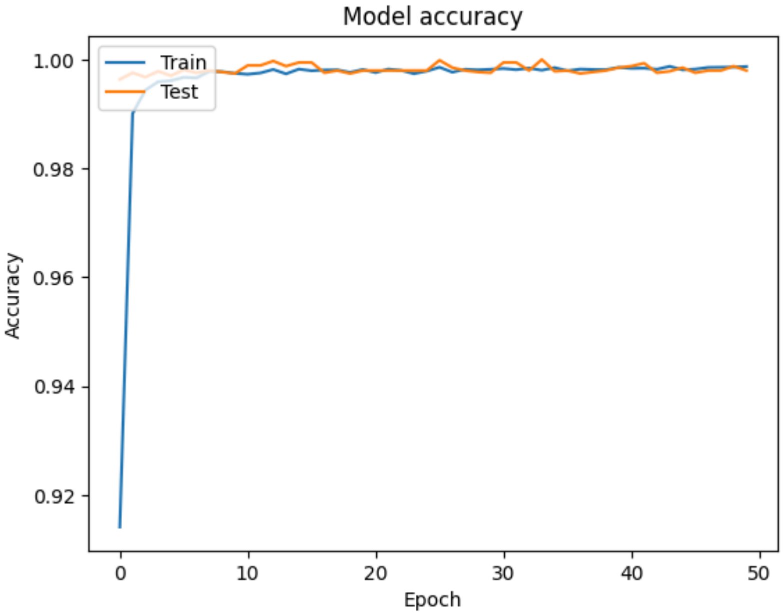
Training was performed over 50 epochs with a batch size of 32, using an 80/20 train-test split.

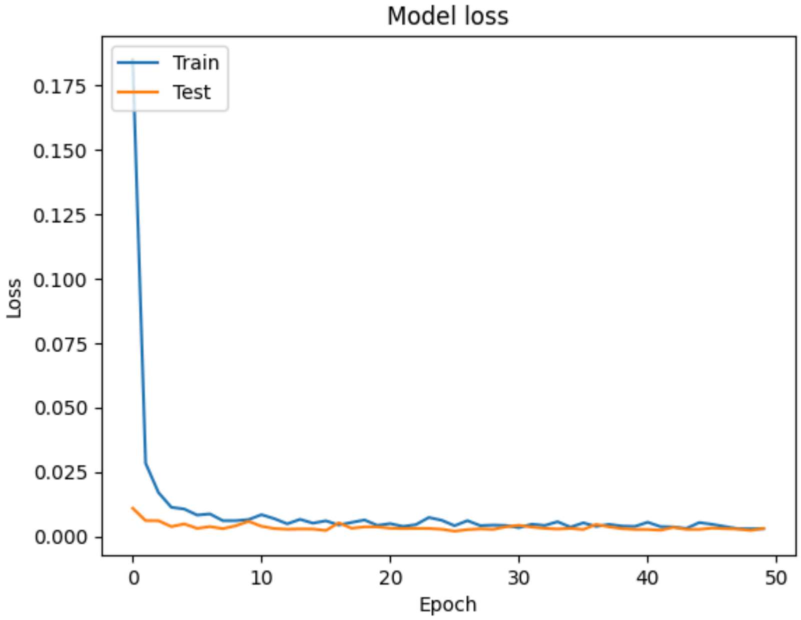
To provide a simpler baseline, a Logistic Regression model was trained on the same preprocessed data. This model assumes linear separability and outputs class probabilities using the sigmoid function. Because it lacks hidden layers, it serves as a useful comparison to determine whether the neural network's depth is necessary for this task.

Results

The neural network achieved near-perfect classification accuracy on both training (0.9987) and test data (0.9980). The small gap between the two values suggests good generalization with minimal overfitting.

Training Curves:





The accuracy plot shows rapid convergence within the first few epochs, stabilizing above 99% accuracy for both training and testing. The loss curve confirms that the error decreased sharply and remained close to zero after convergence.

On the other hand, Logistic Regression achieved about 79% accuracy, considerably lower than the neural network. This result shows that the relationships between features and edibility are not strictly linear, and the FFNN's nonlinear transformations improved classification performance at a significant level. Overall, the Feed-Forward Neural Network vastly outperformed Logistic Regression, achieving over 20% higher accuracy. For real-world use, the FFNN would be the preferred model, as it generalizes extremely well while maintaining high accuracy.

Reflection

This assignment demonstrated the effectiveness of deep learning in handling complex, multi-categorical datasets. The neural network's hidden layers allowed it to learn feature interactions that traditional models like Logistic Regression could not capture. Regularization through Dropout played an important role in preventing overfitting, keeping both training and testing performance consistent. In future iterations, techniques like early stopping could be introduced to further optimize training time.

The exploration of category proportions also offered insight into the biological characteristics of mushrooms, highlighting how class distribution and feature frequency can influence model learning. Although it was disappointing to see that the data used was fabricated for educational purposes, this same model could be utilized on the original real-life dataset from 1987 (https://archive.ics.uci.edu/dataset/73/mushroom) in the future. Overall, this project confirmed that for datasets with many categorical variables and nonlinear dependencies, a Feed-Forward Neural Network provides an advantage over simpler models.