DATA MINING REPORT

DECISION TREE, RANDOM FOREST & ARTIFICIAL NEURAL NETWORKS

SHARON ZACHARIA | 22074407

Link to Dataset | GitHub.

INTRODUCTION

In this report, Artificial Neural Networks (ANNs) and machine learning techniques such as Random Forest and Decision Trees are compared using the mushroom dataset. ANNs utilize input and output nodes to model complex decision boundaries, with perceptron serving as basic models and multi-layer networks extending their capabilities. Training involves adjusting weights and biases to minimize loss functions. Decision Trees act as flowcharts, testing attributes at internal nodes and assigning class labels at leaf nodes, offering a straightforward representation of decision-making processes. Random Forests, comprising multiple decision trees, enhance generalization by introducing randomness in feature selection and training instances, thus improving performance and robustness (Tan, et al., 2019).

Mushroom Dataset: The dataset contains descriptions of samples representing 23 species of gilled mushrooms from the Agaricus and Lepiota family. Each species is categorized as either edible, poisonous, or of unknown edibility and not recommended. The category indicating unknown edibility was merged with the poisonous category.

DATA PRE-PROCESSING TASKS APPLIED ON THE DATASET

Class Label Binarization : Class labels were transformed into binary values edible was encoded to 1 and poisonous to 0

Feature Encoding : Categorical features, including 'CapShape' and 'Population', underwent label encoding, replacing categorical values with numerical equivalents. One-hot encoding was performed on all categorical variables using pandas' **get_dummies** function.

Removing Missing Values : Missing values were removed since the missing fields are small percentage of the entire dataset.

Correlation Analysis : Corelation matrix was constructed to assess the relationships between features to gain insights for features , feature selection and model building processes.

Feature Importance Analysis : Significance of each feature was assessed using Random Forest classifier. Bar plot was used to display the feature importance, for gaining information on the characteristics that are most important in classifying mushrooms.

Principal Component Analysis : Principal Component Analysis (PCA) was performed on dataset to investigate dimensionality reduction methods.

COMPARITIVE ANALYSIS OF MODELS

Classification : Artificial Neural Networks , Decision Tree Classifiers & Random Forest are capable of classification tasks , where input data is assigned to predefined classes.

Scalability : Decision Trees and Random Forests are more scalable and efficient compared to ANNs when dealing with large datasets.

Model Interpretability : Due to inherent nature as sets of if-then-else decision rules, decision trees are naturally interpretable and random forests present a challenge in interpretation due to their aggregation of multiple decision trees. The complex and nonlinear nature of ANNs make them considerably more challenging to interpret.

Model Structure : ANNs learn complex relationships through repeated optimization of weights , Decision tress consists of rules based on feature values and random forests is an ensemble method where each tree is trained on sample of data.

Training Methodology : ANNs are trained using gradient descent algorithm whereas Decision Trees use algorithms like CART or ID3 to maximize information gain or Gini impurity reduction and Random Forest utilizes bagging and feature randomization.

Non-Linear Relationships : ANNs are good at capturing non-linear relationships whereas Decision Trees need deeper trees to capture complex patterns. Random trees also excel at capturing non-linear relationships from multiple decision trees.

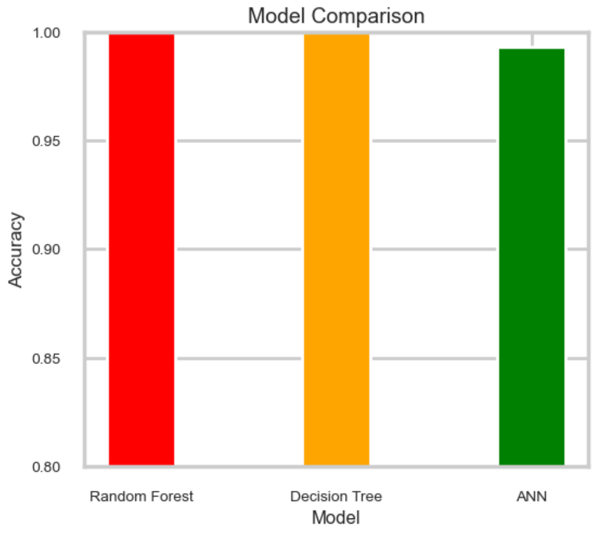

Fig1. Model performance comparison

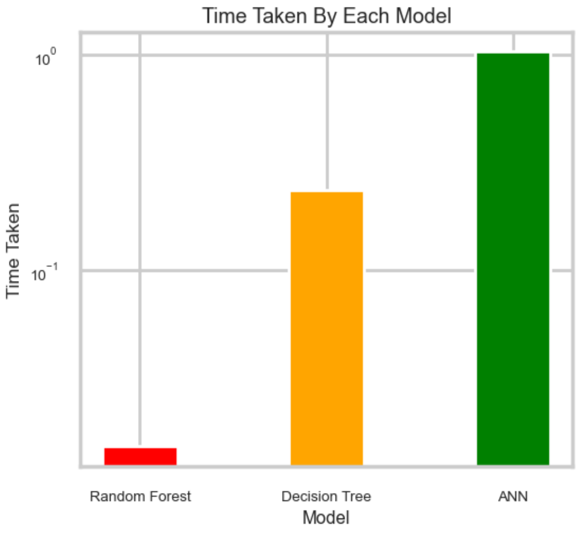

Fig2 . Time Taken by each model

- Random Forest is the fastest among the three algorithms, taking only 0.015 seconds to execute.
- Decision Tree takes 0.235 seconds, which is significantly slower than Random Forest but faster than ANN
- ANN is the slowest, requiring 1.037 seconds to finish. ANN is a potent model that can identify intricate patterns in data. Because of all the calculations and iterations involved in the training process, it takes longer to compute.
- Both Random Forest and Decision Tree achieved high accuracy as compared to ANNs
- The ANN model achieved a slightly lower accuracy of 0.9934, suggesting that both Decision Trees and Random Forest were good at classifying the dataset than ANNs.

REFERENCES

- Tan, P, Steinbach, M, Kumar, V, & Karpatne, A 2019, Introduction to Data Mining EBook: Global Edition, Pearson Education, Limited, Harlow. Available from: ProQuest Ebook Central. [24 April 2024].

- Han, J, Kamber, M, & Pei, J 2011, Data Mining: Concepts and Techniques, Elsevier Science & Technology, San Diego. Available from: ProQuest Ebook Central. [24 April 2024].

- Wes McKinney, Python for Data Analysis, 2e [21 Oct 2017]