Machine Learning Approaches for Predicting Forest Cover Types: A Study of Logistic Regression, K-Nearest Neighbors, and Neural Networks

DASC 5420
Theoretical Machine Learning

He Miao T00732176
Zijian Wang T00729623

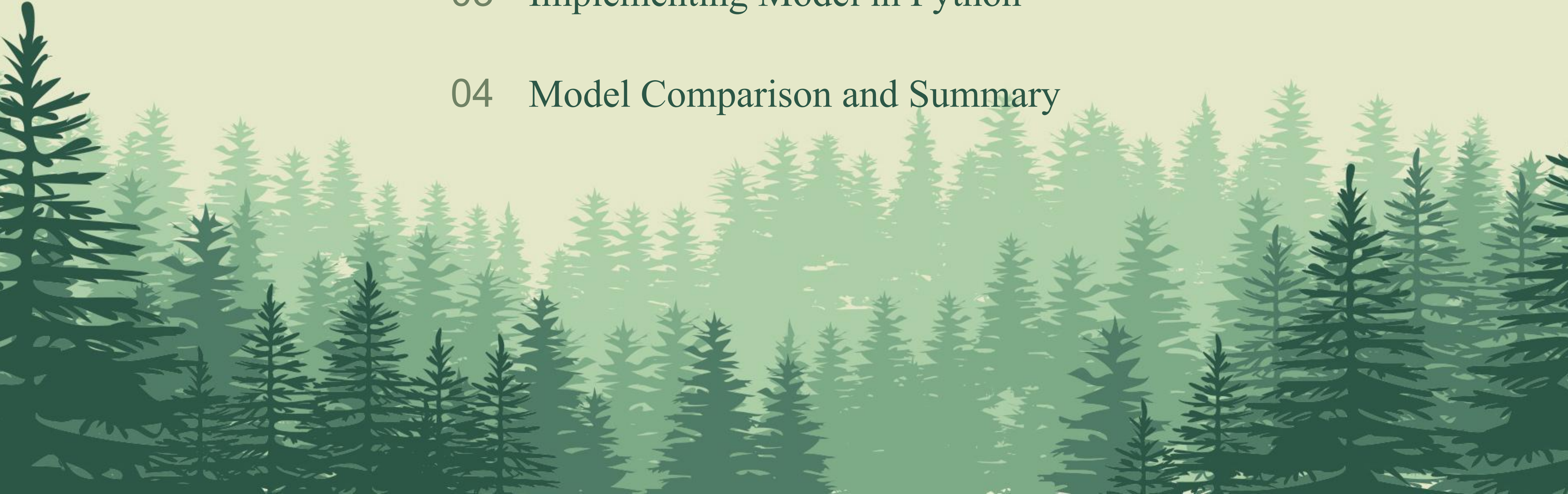
CONTENT

01 Project Introduction

02 Dataset Introduction

03 Implementing Model in Python

04 Model Comparison and Summary



Project Introduction

Project Background

- **Location and Focus:** The study is conducted in the Roosevelt National Forest, Northern Colorado, focusing on forest cover type classification using cartographic variables.
- **Data Utilization:** Instead of relying on remotely sensed imagery, this project employs ecological data, which includes 54 features such as elevation, soil type, and wilderness area information.
- **Importance of Classification:** Accurate classification is essential for effective forest management, conservation efforts, and understanding ecological roles and processes such as succession and nutrient cycling.

Project Objective

- **Primary Aim:** To predict forest cover types using only ecological data, without satellite imagery.
- **Machine Learning Techniques:** Implementation of logistic regression, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) to explore the most effective model.
- **Model Comparison and Validation:** Evaluation of model performance through a rigorous cross-validation process to ensure robustness and generalizability, with KNN showing superior performance.
- **Research Contributions:** The project aims to highlight the potential of KNN in environmental modeling and provide insights into effective machine learning strategies for ecological data analysis, enhancing forest management and conservation strategies.

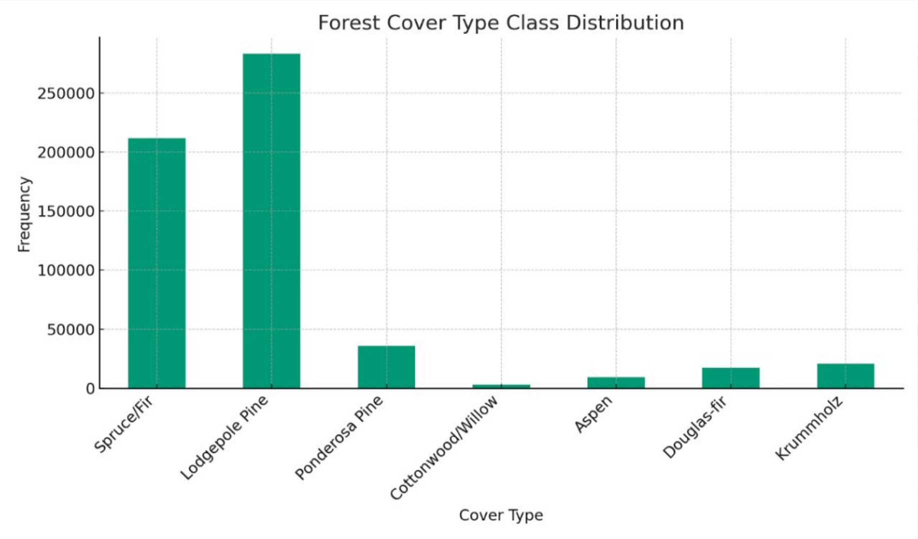
Dataset Introduction

Dataset Overview

- **Source:** Roosevelt National Forest, Northern Colorado
- **Objective:** To predict forest cover types using cartographic variables, without reliance on remotely sensed imagery.
- **Number of Features:** 54, including elevation, soil type, and wilderness area information.
- **Data Handling:** No prior scaling or transformation of data.

Response Variable Descriptions

- **Spruce/Fir**
- **Lodgepole Pine**
- **Ponderosa Pine**
- **Cottonwood/Willow**
- **Aspen**
- **Douglas-fir**
- **Krummholz**



Implementing Model in Python

Data Preprocessing

- **One-Hot Encoding**
  - One-hot encoding is used in multi-class classification to convert categorical labels into a binary matrix, ensuring that machine learning models treat each class as distinct without any ordinal relationship.
- **Data Standardization**
  - Z-score standardization applied to normalize features (mean = 0, standard deviation = 1), crucial for models sensitive to variable scales like KNN.

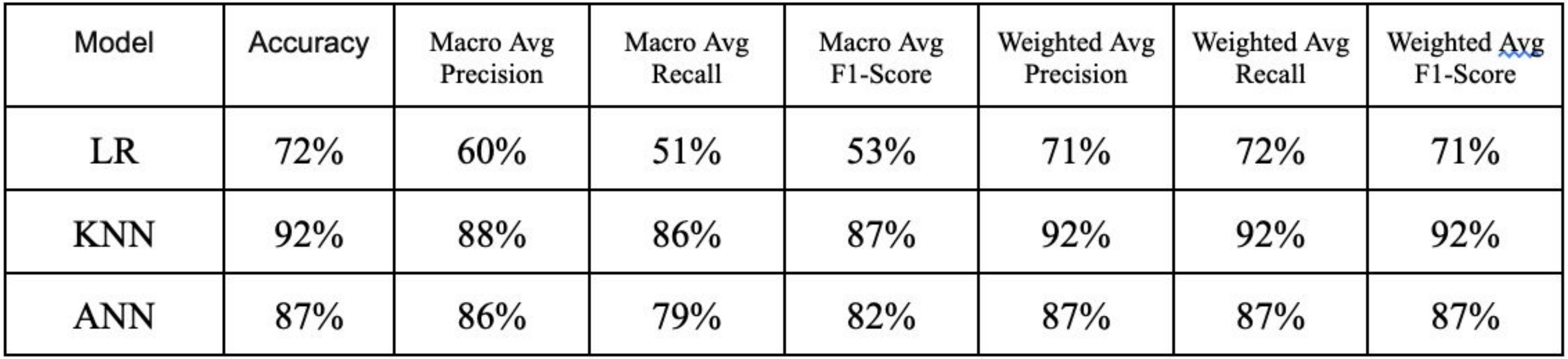
Data Splitting and Validation

- **Data Splitting**
  - Systematic division with 10% reserved for testing and 90% for training.
  - Of the training segment, 22% set aside as a validation set for model tuning, ensuring class representation mirrors overall dataset distribution.
- **Cross-Validation**
  - Integration of 5-fold cross-validation in the hyperparameter tuning phase to enhance model validation and prevent overfitting.

Model Implementation

- **Dimensionality Reduction**
  - PCA applied to reduce dimensions while retaining 95% of original data variance, enhancing computational efficiency and reducing overfitting risk.
- **Model Analysis**
  - **Logistic Regression:** Tuned for regularization strength and penalty type.
  - **K-Nearest Neighbors (KNN):** Optimized for the number of neighbors to balance the bias-variance tradeoff.
  - **Artificial Neural Networks (ANN):** Two hidden layers with ReLU activation and dropout regularization to prevent overfitting.

Model Comparation and Summary



Model	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-Score
LR	72%	60%	51%	53%	71%	72%	71%
KNN	92%	88%	86%	87%	92%	92%	92%
ANN	87%	86%	79%	82%	87%	87%	87%

Table 1: Comparative Classification Metrics

- Macro average calculates metrics for each class independently and then takes the average. This treats all classes equally, regardless of their frequency in the dataset.

- Weighted average calculates metrics for each class like the macro average, but it takes into account the relative size (support) of each class. This means it gives more weight to the metrics of larger classes.

Model Comparison and Summary

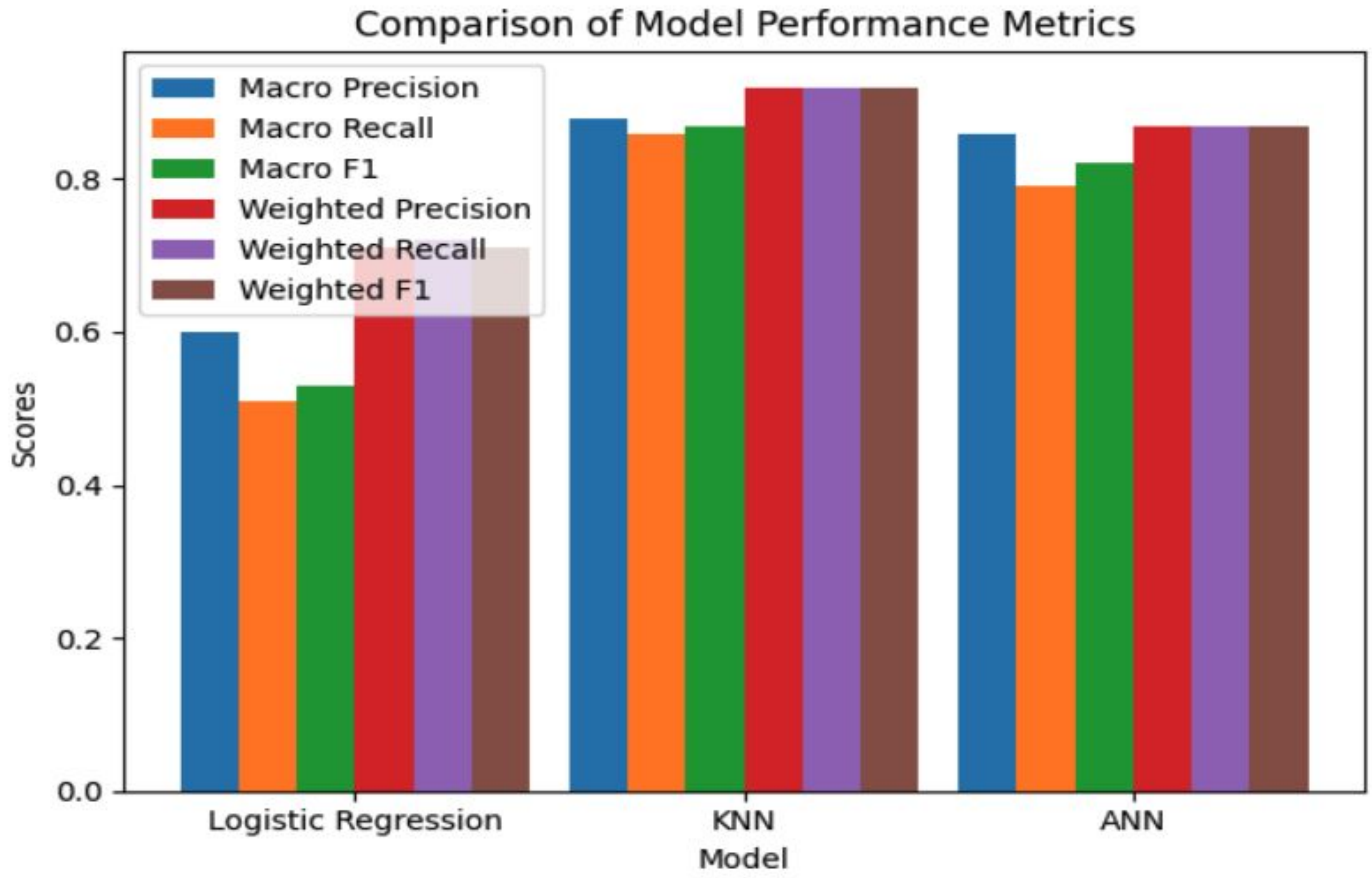


Figure 2: Comparative Model Performance Metrics

Key Findings

- KNN emerged as the top performer with 92.16% accuracy, surpassing other models in precision, recall, and F1-score.
- Logistic Regression offers simplicity and lower computational demands but at the cost of accuracy.
- ANN requires more data and fine-tuning to achieve optimal performance.

Implications for Ecological Management

- Accurate forest type classification enhances sustainable forest management, aids in fire and pest management, and supports conservation efforts.

- Identification of specific forest types can help in strategic interventions like controlled burns and targeted conservation practices.

- Machine learning facilitates detailed understanding of forest cover distribution, crucial for protecting unique ecological habitats.

Thank you !

DASC 5420
Theoretical Machine Learning

He Miao T00732176
Zijian Wang T00729623

Reference

1. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in Python. 2nd ed. New York: Springer; 2023.