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

He Miao T00732176

Zijian Wang T00729623

DASC 5420

Theoretical Machine Learning

Thompson Rivers University

April 18th, 2024

Abstract

This study presents a comprehensive examination of forest cover type classification using cartographic variables from the Roosevelt National Forest in Northern Colorado. By employing machine learning techniques, we aimed to predict forest cover types exclusively from ecological data without relying on remotely sensed imagery. Our dataset comprised 54 features, including elevation, soil type, and wilderness area information, without any prior scaling or transformation. We compared the performance of logistic regression, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) using a rigorous cross-validation process to ensure robustness and generalizability. KNN outperformed both logistic regression and ANN, demonstrating higher accuracy and a more refined understanding of the relationships between the features and the forest cover types. This research underscores the potential of KNN in environmental modeling and offers insights into effective machine learning strategies for ecological data analysis.

I. Introduction

The classification of forest cover types begins with understanding the distinctive characteristics and ecological roles of seven primary tree species found in the dataset:

1. Spruce/Fir: These evergreen conifers are prevalent in colder, higher elevation areas, often forming dense forests that influence soil acidity through their needle litter. They provide stable habitats for diverse wildlife and are integral to nutrient cycling in these ecosystems.
2. Lodgepole Pine: Known for its resilience to environmental disturbances such as fire, the lodgepole pine is crucial for reforestation efforts, quickly colonizing open areas and improving soil stability.
3. Ponderosa Pine: With its thick, puzzle-like bark and robust form, the ponderosa pine dominates various elevations, offering a unique habitat that supports species adapted to its microclimates. It plays a significant role in fire ecology, promoting biodiversity through its fire-dependent regeneration processes.
4. Cottonwood/Willow: These species thrive in moist, riparian zones, rapidly growing trees that significantly affect local water cycles and provide critical habitats for aquatic and terrestrial wildlife.
5. Aspen: Famous for their striking fall foliage and smooth, white bark, aspen forests support high biodiversity, particularly in transitional zones between forest types. They are vital for maintaining biodiversity, offering a unique ecological niche.
6. Douglas-fir: Not a true fir, Douglas-fir is adaptable across a wide range of altitudes and conditions, known for its significant carbon sequestration capabilities and as a primary habitat provider in many North American forests.
7. Krummholz: This cover type refers to stunted trees at the timberline, shaped by harsh winds and cold, embodying the struggle and adaptation of life at high elevations. Krummholz formations are indicative of ecological boundaries and climatic conditions.

Classifying forest cover types is fundamental to enhancing our ecological understanding. Such classifications illuminate the processes of succession, where forest structures and compositions evolve over time, and nutrient cycling, unique to each type of forest cover. For instance, deciduous forests characterized by species such as aspen support different ecological dynamics compared to coniferous forests like those dominated by spruce/fir, which influence soil acidity through needle fall. This knowledge is crucial for predicting how forests adapt to environmental changes, aiding in the development of strategies to bolster ecosystem resilience and recovery after disturbances.

In terms of forest management and conservation, accurate forest type classification underpins the development of tailored practices that align with the specific needs and resilience of forest types. For

example, management techniques for lodgepole pine, which are adapted to rejuvenate after severe fires, differ significantly from those suitable for ponderosa pines, which thrive under a regimen of frequent, low-intensity fires. Such strategic management is essential for mitigating fire risks, preserving wildlife habitats, and enhancing forest health through appropriate conservation and reforestation efforts.

Furthermore, forests play a crucial role in climate regulation through carbon sequestration, local temperature stabilization, and water cycle management. Different forest types have varied impacts on the climate; evergreens might sequester carbon continuously, whereas deciduous trees offer seasonal benefits such as enhanced ground water recharge and temperature modulation. Accurately classified forest data is vital for constructing reliable climate models, assessing environmental impacts, and planning carbon offset initiatives that contribute effectively to global climate change mitigation.

Preserving biodiversity also hinges on the precise classification of forest cover types. Each forest ecosystem supports a distinct array of flora and fauna, making biodiversity conservation context-dependent. Targeted conservation strategies can then be developed to protect specific ecological communities, such as aquatic species in cottonwood/willow dominated riparian zones or terrestrial fauna in Douglas-fir upland forests. Such nuanced management ensures that diverse biological communities are maintained, supporting the overall ecological balance and sustainability of our natural environments.

Through these multifaceted benefits, the classification of forest cover types not only serves ecological and environmental research but also enhances practical applications in forest management, climate strategy formulation, and biodiversity conservation.

II. Data

In this study, the dataset titled "Forest Cover Type" serves as the primary source for the classification of forest cover types, encompassing a comprehensive range of cartographic variables without the inclusion of remotely sensed data. The data was obtained from the U.S. Forest Service (USFS) and the U.S. Geological Survey (USGS), providing a rich set of attributes that contribute to understanding and predicting forest cover type classifications. (data link: https://archive.ics.uci.edu/dataset/31/covertype)

The dataset is substantial, consisting of 581,012 instances, each representing a 30 x 30-meter cell of land area. Each instance is characterized by 54 independent variables that include both quantitative measurements such as elevation, aspect, slope, and various distances to hydrological and man-made features, as well as qualitative binary classifications representing wilderness areas and soil types. The dataset is structured so that these attributes align with the order of numerals along the rows of the database.

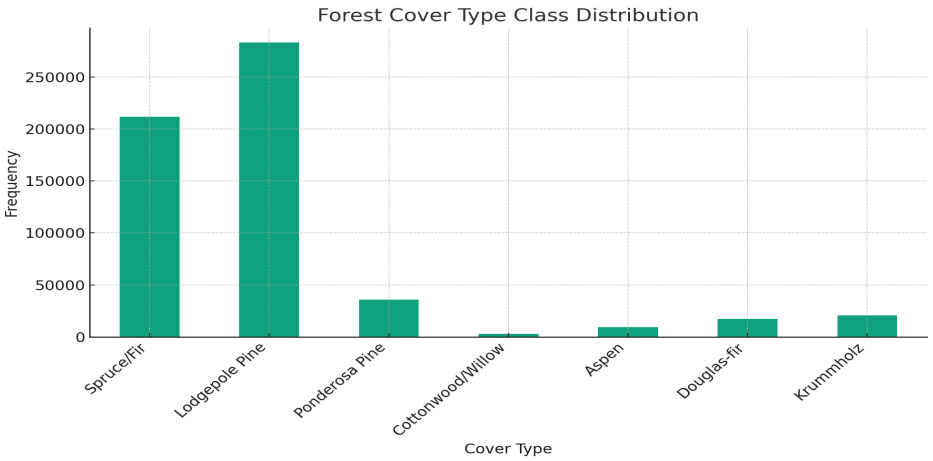

Figure 1: Forest Cover Type Class Distribution

From the available data, the classification goal is to accurately determine one of seven possible forest cover types: Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, or Krummholz. These classes are represented by integers ranging from 1 to 7, which correspond to the order given above. An initial analysis of the dataset reveals an imbalance among the classes, with the Lodgepole Pine and Spruce/Fir types being the most prevalent, whereas Cottonwood/Willow and Krummholz are the least represented.

Statistical analyses of the dataset, such as class distribution, suggest that certain models may need to account for the class imbalance to avoid biased predictions toward the more common classes. To address the imbalance problem, which could predispose machine learning models to exhibit bias towards the more frequent classes, a stratified sampling approach was implemented during the data splitting process. This technique ensures that each class is represented in the training, validation, and test sets in proportions that mirror the overall dataset. These measures ensure that all classes are equally represented during model training and evaluation, which enhances the robustness of the classification models.

So, the "Forest Cover Type" dataset is a comprehensive collection of cartographic variables, rooted in empirical measurements, designed to challenge and refine predictive modeling in ecological contexts. Its utilitarian composition ensures that the derived models are robust, versatile, and reflective of the real-world complexities associated with forest type classification.

III. Method

The study employed the Forest CoverType dataset from the Roosevelt National Forest in Northern Colorado. This dataset includes 54 predictive attributes such as elevation, aspect, slope, and various soil type indicators. The categorical target variable, which delineates the forest cover types, was encoded using one-hot encoding. This transformation converts categorical integer labels into a binary matrix, essential for compatibility with the mathematical operations performed in many machine learning algorithms. To ensure all features contributed equally to the analysis and to mitigate the influence of differing scales, z-score standardization was applied. This method transforms each feature to have a mean of zero and a standard deviation of one, which is crucial for models that assume data is normally distributed or models sensitive to the magnitude of variables, such as K-Nearest Neighbors (KNN).

The dataset was systematically divided into distinct sets; specifically, 10% was reserved for testing, and the remaining 90% for training. This separation is critical to evaluate the model’s performance objectively on unseen data.

The remaining 90% was subdivided, with approximately 22% set aside as a validation set, used exclusively for tuning model parameters. Stratification was employed during these splits to ensure that each class was proportionally represented, mirroring the overall dataset distribution. This is vital to prevent the model from being biased towards the more frequent classes.

Cross-validation is integrated into this setup to enhance the validation process. Specifically, a 5-fold cross-validation was employed during hyperparameter tuning to ensure comprehensive model validation across multiple data subsets. This approach not only mitigates the risk of model overfitting but also confirms the model’s ability to perform consistently across varied scenarios and provides a stable estimate of model accuracy. This methodical validation strategy, combining both a dedicated validation set and cross-validation, offers robust insights into the generalizability and reliability of the predictive models being evaluated.

To manage the high dimensionality of the data, PCA dimensionality reduction was applied, aiming to reduce the number of variables while retaining 95% of the original data variance. This reduction not only enhances computational efficiency but also mitigates the risk of overfitting by simplifying the model’s complexity.

The comparative analysis encompassed Logistic Regression, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN). Each model was evaluated using a combination of grid search and cross-validation to fine-tune the hyperparameters optimally:

- Logistic Regression: The model was fine-tuned over several hyperparameters, most notably the regularization strength (C) with values tested including 0.1 and 10, and the type of penalty applied (l2). Regularization helps prevent the model from fitting too closely to the training data, thereby reducing overfitting.
- K-Nearest Neighbors (KNN): The primary hyperparameter optimized was the number of neighbors (n_neighbors), with values such as 3, 5, and 7 explored. Selecting the optimal number of neighbors is crucial as it balances the bias-variance tradeoff inherent in KNN.

Both models utilized a 5-fold cross-validation approach within the training data. This method partitions the data into five subsets, iteratively training the model on four subsets while validating on the fifth, ensuring comprehensive evaluation across different data samples and enhancing generalizability.

Artificial Neural Networks (ANN): The ANN was constructed with two hidden layers to capture complex patterns in the data:

- Layer Configuration: The first hidden layer consisted of 128 neurons, and the second contained 64 neurons, both using the rectified linear activation function (ReLU). This function introduces non-linearity into the model, allowing for learning more complex functions.
- Dropout Regularization: A dropout rate of 0.5 was employed to randomly deactivate certain neurons during training, effectively simplifying the model temporarily and preventing overfitting.
- Dynamic Learning Rate Adjustment: The learning rate was initially set with Adam optimizer, and adjustments were made using a ReduceLROnPlateau scheduler. This scheduler reduces the learning rate by a factor of 0.1 if no improvement is seen in the validation loss for 10 consecutive epochs, allowing for finer adjustments in learning as the training progresses.
- Training Regime: The network was trained over 1000 epochs, with early stopping implemented to cease training if the validation loss did not improve for a significant number of epochs. This strategy prevents overtraining and ensures the model does not learn the idiosyncrasies of the training data at the expense of its ability to generalize.

The combined use of dedicated validation data and systematic cross-validation provides a robust mechanism for assessing model efficacy. The validation set allows for the immediate evaluation of model changes (such as learning rate adjustments in ANNs), while cross-validation aggregates performance metrics across multiple data folds to assure model stability and reliability across diverse data scenarios.

Final evaluation was conducted on the independent test set, with metrics including accuracy, precision, recall, and F1-score to provide a comprehensive assessment of model efficacy. This evaluation confirms the model's ability to generalize to new, unseen data, an essential factor for practical deployment.

Github link: https://github.com/flistz/DASC-5420/blob/main/ML_Project.ipynb

IV. Results

This study evaluated the effectiveness of Logistic Regression, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) in classifying forest cover types based on cartographic variables

from the Roosevelt National Forest. Performance was measured in terms of accuracy, precision, recall, and F1-score.

The accuracy reported here is not from a singular evaluation but rather derived from a 5-fold cross-validation process. This cross-validation not only reinforces the reliability of the reported accuracy figures but also provides a holistic measure of each model's performance across the entirety of the dataset. This ensures that the performance is not contingent on a particular subset of data, which could introduce bias, but is representative of the model's generalizability.

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

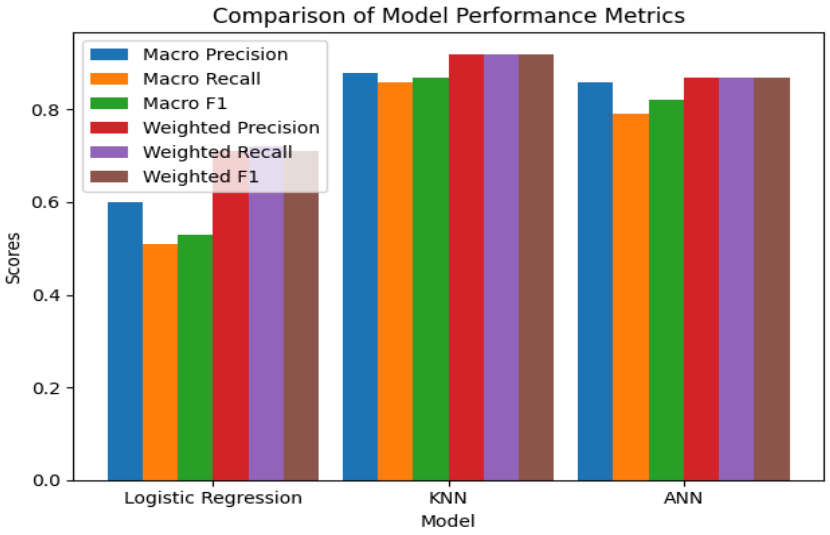

Figure 2: Comparative Model Performance Metrics

As depicted in Figure 2 and summarized in Table 1, KNN achieved the highest accuracy at 92%, with the highest macro and weighted averages across precision, recall, and F1-score. ANN followed with a respectable accuracy of 87%, while LR lagged with an accuracy of 72%. Notably, the KNN model demonstrated consistently high scores across all metrics, indicating a strong balance between identifying relevant data points (precision) and retrieving a comprehensive set of data points (recall).

The superior performance of KNN could be attributed to the nature of the dataset, which contains nuanced patterns and relationships that are better captured by the instance-based learning of KNN. The presence of intricate ecological patterns that do not conform to a linear separability may have reduced the effectiveness of LR, while the ANN, although robust, might require more extensive fine-tuning to match the performance of KNN.

The results obtained from this study do not merely highlight the best-performing model but also reflect the underlying complexity of the ecological data. They reinforce the concept that the choice of algorithm plays a pivotal role in predictive tasks, where each model's intrinsic properties should be aligned with the data's characteristics. In the context of forest cover type classification, KNN proved

to be particularly effective, capturing the inherent complexity and diversity of natural ecosystems represented within the data.

V. Conclusion

The objective of this study was to evaluate and compare the efficacy of various machine learning models in the classification of forest cover types using cartographic variables. The study presents a thorough analysis of Logistic Regression, K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN) through a rigorous cross-validation framework. The findings conclusively demonstrated that KNN, which attained an accuracy of 92.16%, outperformed its counterparts in terms of precision, recall, and F1-score across all classes, as reflected in Figure 2 and Table 1.

KNN's superior performance can be attributed to its ability to navigate the complexity and non-linearity of ecological data, making it a preferable choice for similar classification tasks. While Logistic Regression presented a simpler model with less computational demand, its lower performance metrics indicate a trade-off between simplicity and accuracy. ANN showed promise, yet it requires fine-tuning and potentially more data to achieve the level of performance seen in KNN. And the application of PCA for dimensionality reduction prior to model training proved to be a significant step in enhancing model performance. By focusing on the most informative features, PCA allowed each model to operate within a reduced feature space, leading to improved computation times and potentially better generalization capabilities.

This study underscores the importance of algorithm selection in ecological data analysis, which could influence future research direction in the field. Given the success of KNN in this context, future work could explore the integration of other instance-based learning algorithms or ensemble methods that may further improve classification performance. And the study reinforces the potential of data-driven strategies in environmental management.

And the findings from this research hold substantive implications for the practical application of machine learning in forest type classification—a critical component of ecological management. The ability to accurately distinguish between various forest cover types can significantly bolster efforts in sustainable forest management, particularly in crafting strategies tailored to the needs of specific ecosystems. For instance, accurate classification can enhance fire management practices by identifying forest areas more susceptible to wildfire, allowing for preemptive measures such as controlled burns or the strategic clearing of underbrush. In pest management, distinguishing forest types prone to infestation can guide the allocation of resources for early detection and containment efforts, preserving the health and biodiversity of the forest. In the realm of conservation, detailed knowledge of forest cover distribution is invaluable. It aids in protecting endangered habitats, especially those that may not be visibly distinct but are ecologically unique. For example, certain bird species might only nest in a specific forest type; precise classification enables conservationists to pinpoint and protect these critical nesting grounds.

In conclusion, the study presents a compelling case for the integration of machine learning models into the domain of forest ecology. The advanced classification capabilities demonstrated here offer a multitude of practical applications, from enhancing forest resilience against natural threats to informing policy for climate action. By translating these insights into actionable measures, stakeholders can better safeguard forest ecosystems, ensuring their health and sustainability for generations to come.

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.

Github link: https://github.com/flistz/DASC-5420/blob/main/ML_Project.ipynb