UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI
DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGY

MACHINE LEARNING 2 FINAL PROJECT

Group 37

Convertype Classification

Members:

Trinh Nhat Huy - 23BI14193
Nguyen Minh Tuan - 23BI14441
Nguyen Quang Minh - 23BI14288
Hoang Quang Minh - 23BI14281
Cao Hoang Linh -
Nguyen Le Minh -

Hanoi, February 22, 2026

Contents

List of Figures

List of Tables

1. Data Analysis

2. Results

This section presents the experimental results obtained from different classification models on the Covertype dataset. We evaluate model performance under two scenarios: training with and without SMOTE to address class imbalance.

The comparison focuses on both predictive performance and computational efficiency. Performance is assessed using Balanced Accuracy, Macro F1-score, Weighted F1-score, and Cohen's Kappa, while computational cost is measured in terms of training and prediction time.

The objective is to analyze the trade-offs between accuracy, robustness to class imbalance, and computational complexity across different machine learning models.

2.1. Overall Performance Comparison

This subsection provides a comprehensive comparison of model performance under two experimental settings: training without SMOTE and training with SMOTE. We evaluate classification effectiveness using Balanced Accuracy, Macro F1-score, Weighted F1-score, and Cohen's Kappa. In addition, computational efficiency is assessed through training time and prediction time. The objective is to identify performance trade-offs across different machine learning models and to analyze the impact of SMOTE on imbalanced data handling.

Scenario	Model	Bal. Acc	Macro F1	W-F1	Kappa	Train (s)	Pred (s)	Samples
Without SMOTE	Decision Tree	0.8931	0.8173	0.8819	0.8096	8.84	0.22	406,708
	Random Forest	0.9172	0.8565	0.8973	0.8347	92.39	3.37	406,708
	KNN (k=5)	0.8708	0.8816	0.9307	0.8888	0.29	617.38	406,708
	SVM (RBF)	0.7646	0.6183	0.7243	0.5683	20.38	223.82	20,000
With SMOTE	Decision Tree	0.8980	0.8248	0.8841	0.8137	74.91	0.08	1,388,170
	Random Forest	0.9268	0.8534	0.8968	0.8338	572.91	4.15	1,388,170
	KNN (k=5)	0.9153	0.8672	0.9240	0.8780	0.17	2118.72	1,388,170
	SVM (RBF)	0.7922	0.5884	0.6898	0.5206	9.65	175.44	20,000

Table 1: Overall performance comparison across models with and without SMOTE.

2.2. SMOTE impact analysis

Model	Δ Balance Accuracy	Δ Macro F1	Δ W-F1	Δ Kappa
Decision Tree	+0.0049	+0.0075	+0.0022	+0.0041
Random Forest	+0.0095	-0.0030	-0.0005	-0.0010
KNN (k=5)	+0.0444	-0.0144	-0.0067	-0.0108
SVM (RBF)	+0.0276	-0.0299	-0.0345	-0.0477

Table 2: Detailed performance change after applying SMOTE.

Table 2 shows that SMOTE consistently improves Balanced Accuracy across all models, indicating better recall for minority classes after oversampling. The most significant improvement is observed for KNN, suggesting that distance-based methods benefit substantially from a more balanced class distribution.

However, the gains in Balanced Accuracy are not always accompanied by improvements in Macro F1, Weighted F1, or Cohen's Kappa. In particular, Random Forest and SVM exhibit slight to notable declines in these metrics, implying that while minority class detection improves, overall predictive agreement and precision may decrease. This highlights a trade-off between class-balanced performance and global model stability when applying SMOTE.

3. Discussion

3.1. The Impact of SMOTE on Class Imbalance

The Covertype dataset presents a severe class imbalance challenge, wherein the majority class (Type 2) comprises 48.76% of the instances, whereas the minority class (Type 4) accounts for only 0.47% (a ratio of approximately 103:1). To address this inequality, classification experiments were conducted across two scenarios: the original imbalanced data (Without SMOTE, 406,708 samples) and a synthetically balanced dataset (With SMOTE, 1,388,170 samples).

The application of SMOTE yielded a precise trade-off between performance metrics. Across all evaluated models, SMOTE successfully improved the Balanced Accuracy. For example, the Random Forest classifier exhibited an increase in Balanced Accuracy from 0.9172 to 0.9268. However, this enhancement occurred at the expense of the Macro F1 score, which decreased slightly for models such as Random Forest (from 0.8565 to 0.8534) and substantially for Linear SVM (from 0.5414 to 0.4670). This reduction indicates that the synthetic oversampling induced the models to over-predict minority classes, thereby increasing false positive rates for the majority classes.

3.2. Model Selection and Performance Results

The selection of the optimal model necessitates a balance between predictive performance, inference speed, and scalability.

Random Forest: Random Forest emerged as the strongest performing model overall. It achieved the highest Balanced Accuracy (0.9268 with SMOTE) and the highest Macro F1 score (0.8565 without SMOTE). Its ensemble architecture effectively accommodated the dataset's non-linear decision boundaries and the heterogeneity of continuous and sparse binary features. Furthermore, while SMOTE increased the training time (from 92s to 572s), the prediction time remained under 5 seconds.

K-Nearest Neighbors (KNN): KNN (k=5) demonstrated robust predictive capabilities, achieving a Balanced Accuracy of 0.9153 with SMOTE. However, as a lazy learner, its inference time presented a critical limitation. Predicting the test set required over 600 seconds on the imbalanced dataset and increased to over 2100 seconds with SMOTE. This $O(n \times d)$ query complexity renders KNN impractical for real-time applications.

Support Vector Machines (SVM): The SVM models struggled with this dataset. The Linear SVM performed poorly (Balanced Accuracy of 0.6196 to 0.6736), which shows that the different forest cover types cannot be separated by simple straight lines. Meanwhile, the non-linear SVM (RBF kernel) was too computationally heavy. Training it on the full data would take far too long, so it had to be trained on a tiny subset (only 20,000 samples), resulting in weak performance (Balanced Accuracy ≈ 0.79). Ultimately, SVM proved to be either too rigid or too slow for this problem.

Conclusion on Model Selection vs. SMOTE: The results clearly indicate that choosing a highly capable model (like Random Forest) is fundamentally more effective than applying data

balancing techniques (like SMOTE) to a weaker model. A Random Forest without SMOTE significantly outperformed a Linear SVM with SMOTE across all metrics. While SMOTE provides marginal improvements in detecting minority classes for strong models, the underlying model architecture is the primary driver of success on complex, non-linear data.

3.3. Feature Impacts on Model Choice

Analysis shows that continuous variables like Elevation strongly dictate forest cover types. Random Forest handles this dataset exceptionally well because its tree structure naturally creates hard decision boundaries based on these physical thresholds (e.g., Elevation ranges) alongside the sparse categorical data (44 Soil and Wilderness types). In contrast, models like Linear SVM perform poorly because they attempt to draw straight lines through a heavily non-linear and mixed-type feature space.

3.4. Practical Limitations

1. **SMOTE is Computationally Inefficient:** Balancing the data expanded the training set to nearly 1.4 million samples. This massively increased training time (e.g., Random Forest training jumped from 92s to 572s) while actually reducing the overall Macro F1 score. In a real-world pipeline, the huge hardware cost of SMOTE is not justified.

2. **Unusable Models at Scale:** Theoretical models failed in practical application. RBF SVM ran out of memory and could not even be trained on the full dataset. KNN achieved good accuracy, but its prediction phase took far too long (over 30 minutes), making it entirely impractical for fast, real-time inference.