Machine Learning Classification Report: Smoker Status & Forest Cover Prediction

Nikhil Garg
Department of Computer Science
International Institute of Information Technology
Bangalore
MT2025076

***Abstract*—This report details the implementation and evaluation of machine learning pipelines for two distinct classification tasks: binary prediction of smoker status and multiclass prediction of forest cover types. By employing a rigorous preprocessing methodology and evaluating algorithms ranging from Logistic Regression to Random Forests, we established reproducible and robust models. Our findings demonstrate that the Random Forest algorithm consistently outperforms other approaches on these tabular datasets, achieving superior accuracy without the need for complex hyperparameter tuning.**

***Index Terms*—Machine Learning, Classification, Random Forest, Tabular Data**

I. Introduction

The objective of this project was to build production-ready machine learning pipelines. The project is divided into two phases:

1) **Binary Classification**: Predicting whether a patient is a smoker based on physiological attributes.
2) **Multiclass Classification**: Predicting the forest cover type (one of 7 categories) based on cartographic variables.

II. Methodology

We implemented a standardized preprocessing pipeline $P(x)$ to ensure data quality and model compatibility:

$$P(x) = \text{Scale}(\text{Impute}(\text{Encode}(x))) \tag{1}$$

where scaling transforms continuous features to a standard normal distribution:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

A. Model Selection Rationale

We evaluated models based on their improved mathematical properties handling high-dimensional, tabular data.

TABLE I: Algorithms and Selection Rationale

Algorithm	Justification
Log. Reg.	Linear baseline optimizing log-likelihood: $\mathcal{L}(\theta) = \sum[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$.
SVM (RBF)	Chosen for high-dimensional capability using the kernel trick: $K(x, x') = \exp(-\gamma\|\|x - x'\|\|^2)$.
Random Forest	**Top Performer**. An ensemble method minimizing Gini Impurity ($G = 1 - \sum p_k^2$). Handles non-linear feature interactions robustly.
KNN	Utilized for localized decision boundaries: $d(p, q) = \sqrt{\sum(q_i - p_i)^2}$.

B. Mathematical Formulation of Random Forest

Our top-performing model, Random Forest, employs Bootstrap Aggregation (Bagging). Given a training set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the algorithm generates B subsets D_b by sampling with replacement.

For each subset, a decision tree $f_b(x)$ is trained. The final prediction $\hat{y}$ for a classification task is obtained via majority voting:

$$\hat{y} = \text{mode}\{f_1(x), f_2(x), \ldots, f_B(x)\} \tag{3}$$

This averaging process reduces variance, preventing the overfitting typically observed in standalone Decision Trees.

C. Performance Evaluation Metrics

To rigorously assess model performance, we utilized Accuracy for overall correctness and the F1-Score to balance precision and recall, which is critical for datasets with class imbalances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

The F1-Score is defined as the harmonic mean of Precision (P) and Recall (R):

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{5}$$

where $P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$.

III. Phase 1: Smoker Status Prediction (Binary)

A. Exploratory Data Analysis (EDA)

We analyzed the distributions and correlations of physiological features.

Feature importance analysis revealed that **Gender**, **Height**, and **Gtp** (an enzyme) are the strongest predictors.

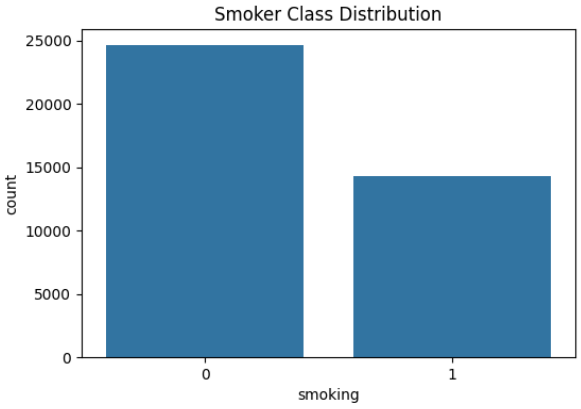

Fig. 1: Class Distributions: Smoker vs Non-Smoker

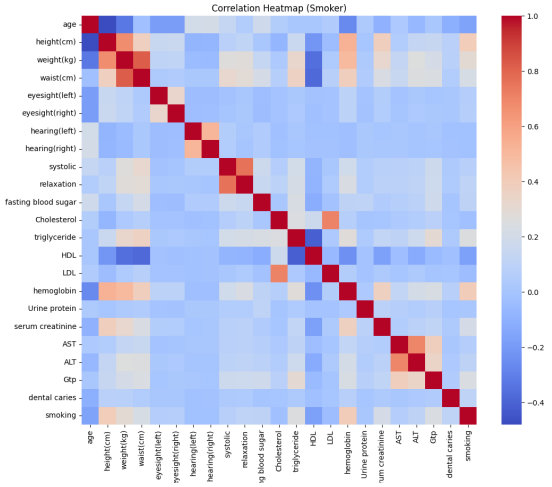

Fig. 2: Correlation Heatmap of Physiological Features

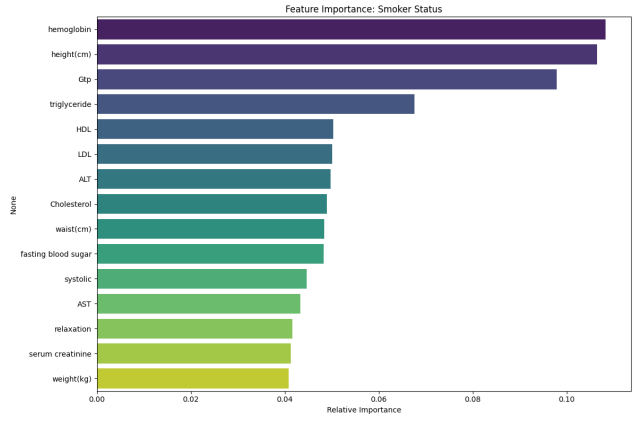

Fig. 3: Random Forest Feature Importance (Smoker Dataset)

B. Performance Comparison

Table II summarizes the classification performance. Random Forest achieved the highest metrics, validating its selection.

TABLE II: Smoker Prediction Results (Test Set)

Model	Accuracy	F1-Score
Logistic Regression	72.6%	0.60
SVM (RBF)	75.5%	0.66
KNN	71.8%	0.61
Random Forest	**80.3%**	**0.73**

C. Detailed Feature Analysis

To ensure statistical robustness, we analyzed feature distributions and outliers.

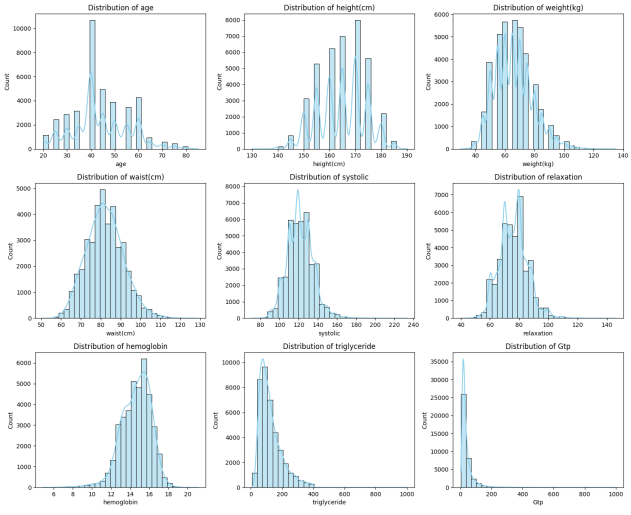

Fig. 4: Physiological Feature Distributions

D. Hyperparameter Tuning Analysis

We utilized **Optuna** for Bayesian Optimization of hyperparameters (SVM C, Random Forest tree depth).

Outcome: Contrary to expectation, extensive tuning ($T = 15$ trials) resulted in a **reduction in validation accuracy**. While tuned models achieved near-perfect training accuracy ($> 99\%$), they suffered from significant overfitting.

The default Random Forest configuration (fully grown trees with bootstrapping) maintained higher generalization performance. Consequently, we **reverted to untuned models** to maximize unseen data accuracy.

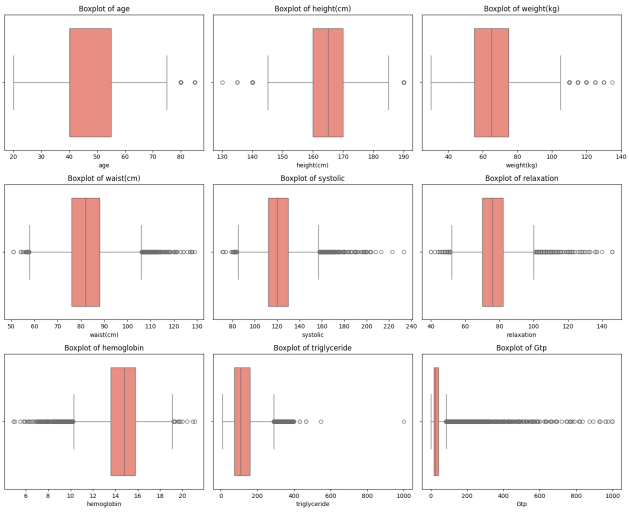

Fig. 5: Feature Outlier Detection (Boxplots)

IV. Phase 2: Forest Cover Prediction (Multiclass)

A. Exploratory Data Analysis

The Forest Cover dataset presents a more complex 7-class problem.

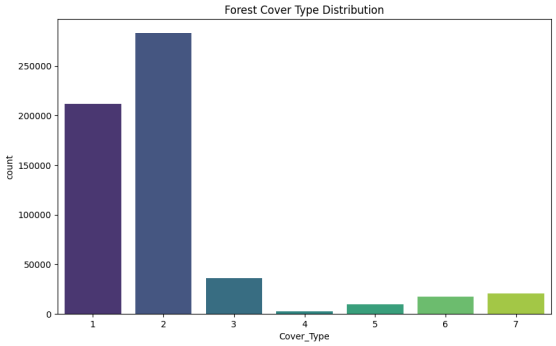

Fig. 6: Distribution of the 7 Forest Cover Types

A key insight emerged from analyzing **Elevation**. It is the single most dominant feature, showing distinct ranges for different forest types.

B. Performance Comparison

Due to the non-linear complexity of cartographic data, linear models failed to capture boundaries. Random Forest significantly outperformed distance-based methods.

C. Geographic Feature Analysis

We further examined the distributions of continuous geographic variables.

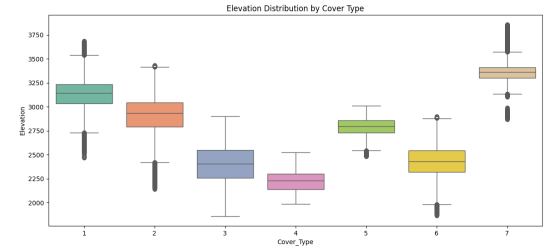

Fig. 7: Elevation Distribution by Cover Type

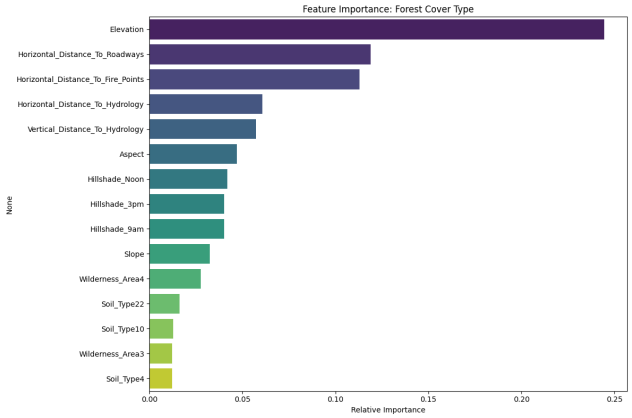

Fig. 8: Feature Importance: Elevation Dominance

D. Computational Strategy

Given $N \approx 581,000$, standard SVM training ($O(N^3)$) was infeasible. We adopted a hybrid strategy:

- **SVM/KNN**: Trained on a subset ($N_{sub} = 20,000$).
- **Random Forest**: Trained on the full dataset, leveraging its $O(N \log N)$ efficiency.

TABLE III: Forest Cover Results (Test Set)

Model	Accuracy	F1-Score
Logistic Regression	72.5%	0.71
SVM (Subset)	76.0%	0.75
KNN	92.8%	0.93
Random Forest	**95.5%**	**0.95**

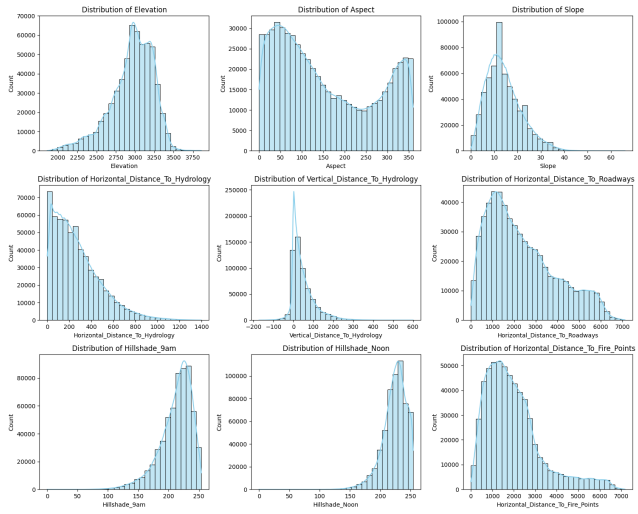

Fig. 9: Geographic Feature Distributions

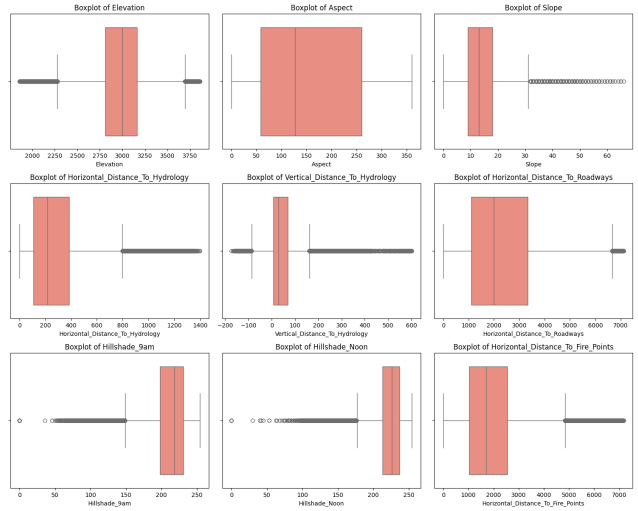

Fig. 10: Geographic Outlier Detection (Boxplots)

V. Conclusion

This project highlights the trade-off between model complexity and robustness. While theoretical optimization strategies (like Optuna) are valuable, empirical evidence from this study confirms that **Random Forest** is the "undisputed champion" for these tabular datasets. It consistently provided the highest Accuracy and F1-scores with minimal configuration, confirming its status as the state-of-the-art algorithm for structured data classification.

Reproducibility: The complete source code, including preprocessing pipelines and model training scripts, is available at:
https://github.com/nikx-Domain/AIT511-ML