International Institute of Information Technology
Bangalore

Machine Learning
AIM 511

Project Report

Shikhar Bhadreshkumar Mutta (MT2025114)

`ML-Project Github Link`

December 12, 2025

Contents

1 Dataset - 1 Smoking

Smoking remains one of the most significant public health concerns worldwide, contributing to cardiovascular disease, respiratory illness, increased cancer risk, and measurable biochemical stress on vital organs. Understanding the physiological and clinical markers associated with smoking is essential for developing predictive systems that can identify smoking behavior from health indicators.

The primary objective of this project is to analyze health examination data and build a machine learning model capable of predicting whether an individual is a smoker or a non-smoker. This report presents a structured exploratory data analysis (EDA) to uncover meaningful relationships, variable trends, and physiological signatures that distinguish smokers from non-smokers.

The analysis follows a sequential methodology:

- Conducting EDA to understand distributions, correlations, and feature significance.
- Identifying health markers strongly impacted by smoking.
- Preparing the dataset for downstream machine learning modeling.

2 Dataset Description

The dataset consists of several physiological, biochemical, and anthropometric measurements collected from individuals undergoing routine health examinations. The target variable, `smoking`, is binary:

- **0**: Non-Smoker
- **1**: Smoker

The dataset includes the following variables:

- Age
- Height (cm)
- Weight (kg)
- Waist circumference (cm)
- Eyesight (left/right)
- Hearing (left/right)
- Systolic and diastolic blood pressure
- Total cholesterol, triglycerides, HDL, LDL

- Hemoglobin
- Urine protein level
- Serum creatinine
- Liver enzymes (AST, ALT, GTP)
- Dental caries indicator

The dataset contains no missing values, enabling direct modeling without imputation. Most features are numerical, and the biochemical markers exhibit natural right-skewness, which is common in medical datasets. Liver enzymes and lipid levels play a crucial role in capturing lifestyle-related physiological stress.

3 Exploratory Data Analysis (EDA)

This section presents a comprehensive exploratory data analysis of the smoking dataset. The goal is to uncover trends, distributions, correlations, and health-related patterns that differentiate smokers and non-smokers, thereby informing downstream feature engineering and model development.

3.1 Class Distribution in the Dataset

We begin by analyzing the distribution of the target variable, `smoking`. The dataset presents a binary classification problem, with two groups:

- 0 → Non-Smoker
- 1 → Smoker

Preliminary analysis shows that the dataset exhibits a slightly imbalanced distribution, with non-smokers forming the majority. This class imbalance is minor and manageable through stratified sampling during train–test splitting.

Interpretation:

- The class ratio is acceptable and does not require heavy resampling techniques.
- Metrics such as F1-score and AUC will be more informative than accuracy alone.

Table 1: Smoking Statistics

Smoking	**Count**	**Percentage (%)**
0	24666	63.27
1	14318	36.73

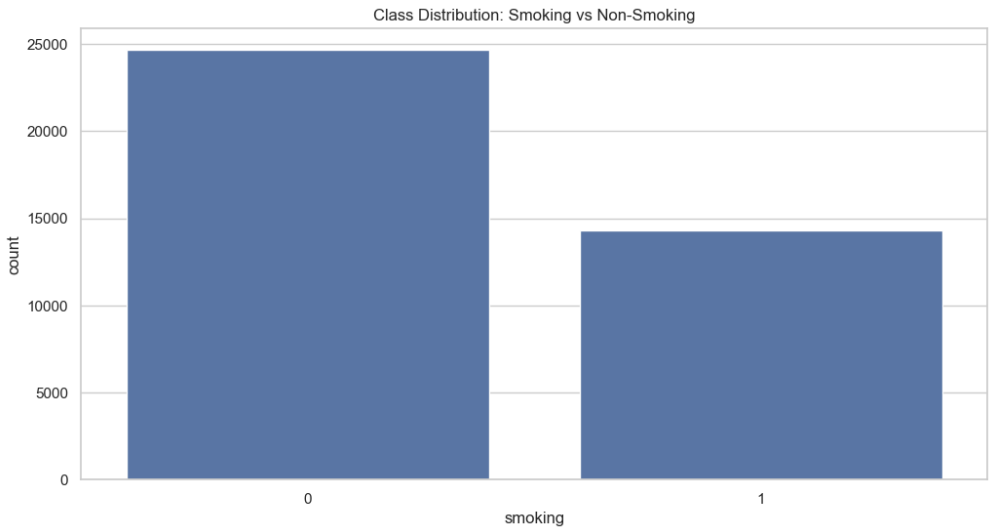


Figure 1: Class Distribution of Smokers and Non-Smokers

3.2 Feature Distributions

To understand the underlying health characteristics, the distributions of key physiological and biochemical markers were examined. Several patterns emerge.

Blood Pressure (Systolic & Diastolic)

Both systolic and diastolic pressure show broad but stable distributions. Smokers tend to fall slightly toward the higher end of the systolic range.

Cholesterol Profile (HDL, LDL, Total Cholesterol)

- HDL (good cholesterol) is generally lower in smokers.
- Triglycerides and LDL show a heavier right tail in smokers.

These patterns align with known clinical effects of smoking.

Liver Enzymes (AST, ALT, GTP)

- All three liver enzymes exhibit significant right-skewness.
- Smokers show distinctly elevated peaks, especially in GTP, associated with liver stress.

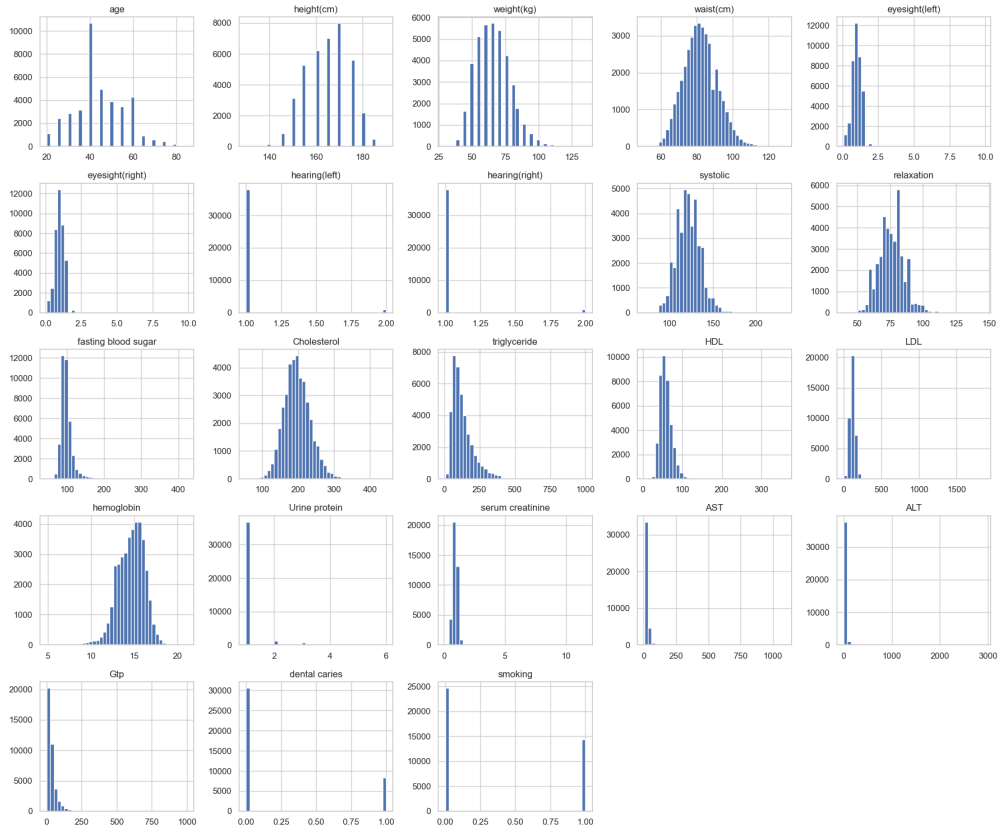
Vision & Hearing Metrics

Eyesight and hearing values show no strong deviations between smokers and non-smokers. These variables behave more like neutral or weak predictors.

Anthropometric Features (Waist, Height, Weight)

- Waist circumference shows a noticeable shift toward higher values in smokers.
- Height and weight distributions are largely similar across classes.

Interpretation: Smoking has measurable physiological impacts reflected in biochemical markers. Features such as GTP, ALT, AST, HDL, triglycerides, and waist measurements show clear separations and will likely be high-importance predictors.



3.3 Feature Distributions by Smoking Status

To better visualize how each feature varies between classes, kernel density plots (KDE) were generated.

Liver Function Markers

Smokers show consistently higher densities in the upper ranges of AST, ALT, and especially GTP. These markers form a distinct "liver-stress signature" strongly correlated with smoking.

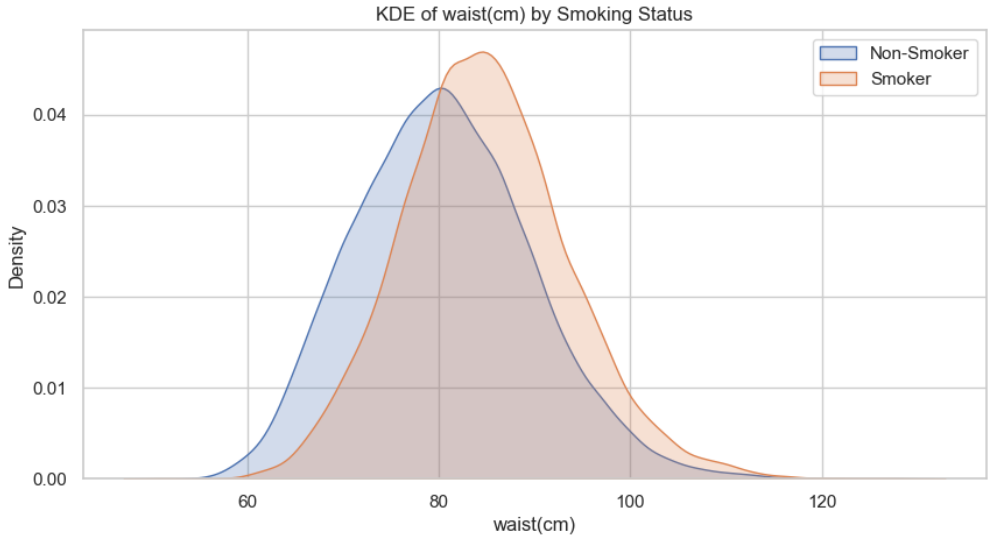
Lipid Profile

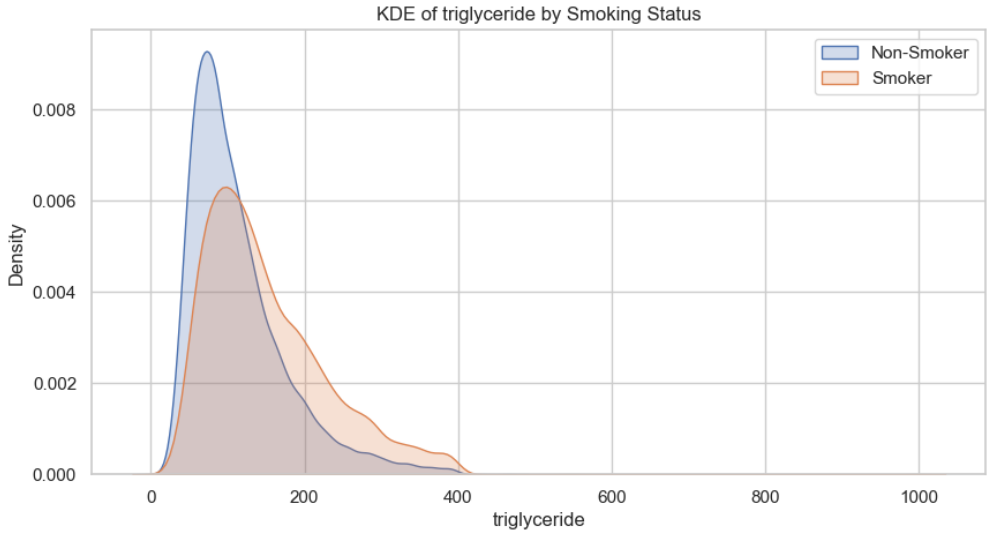
- HDL shifts lower for smokers.
- Triglycerides show strong separation with a visibly heavier tail among smokers.

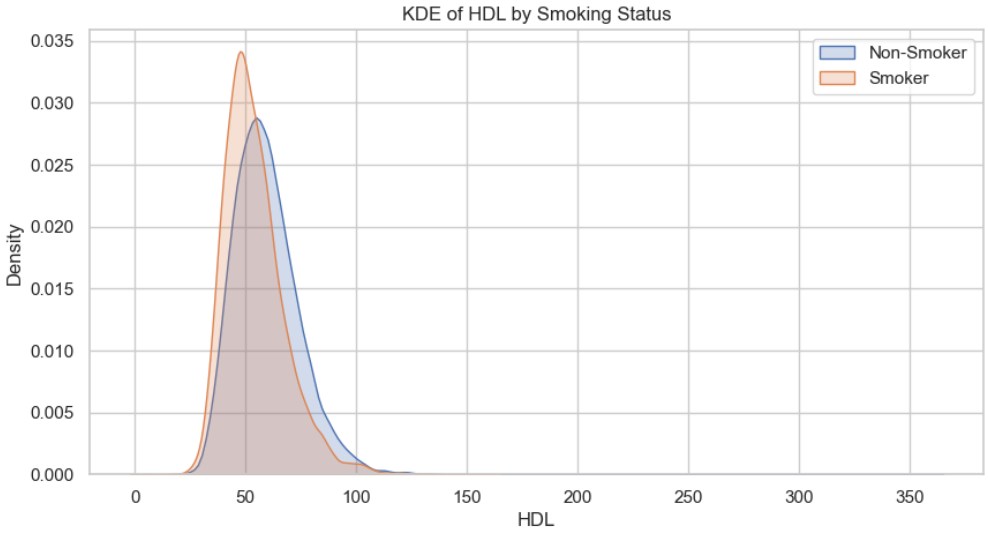
Physical Measures

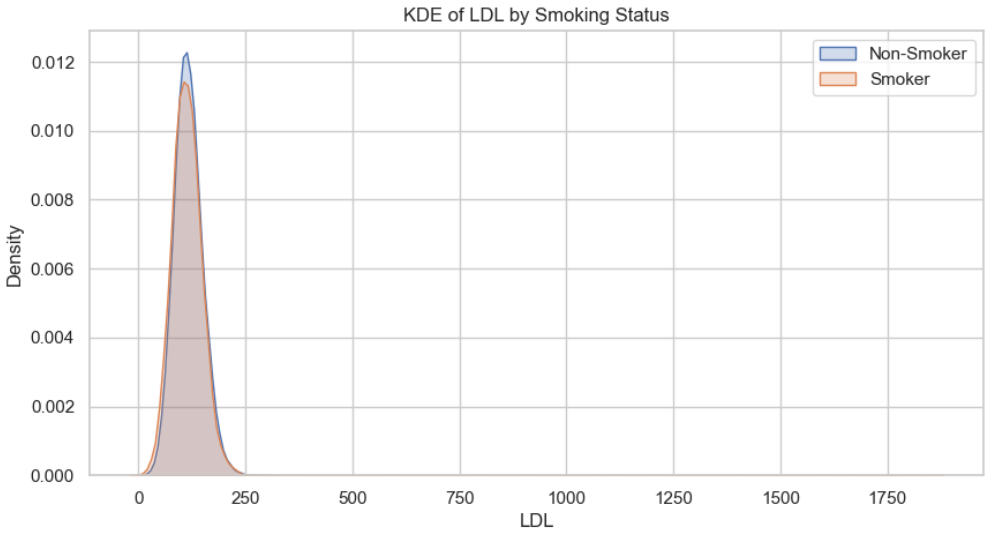
- Waist circumference shows clear upward shifts for smokers.
- Height and weight show no smoking-dependent difference.

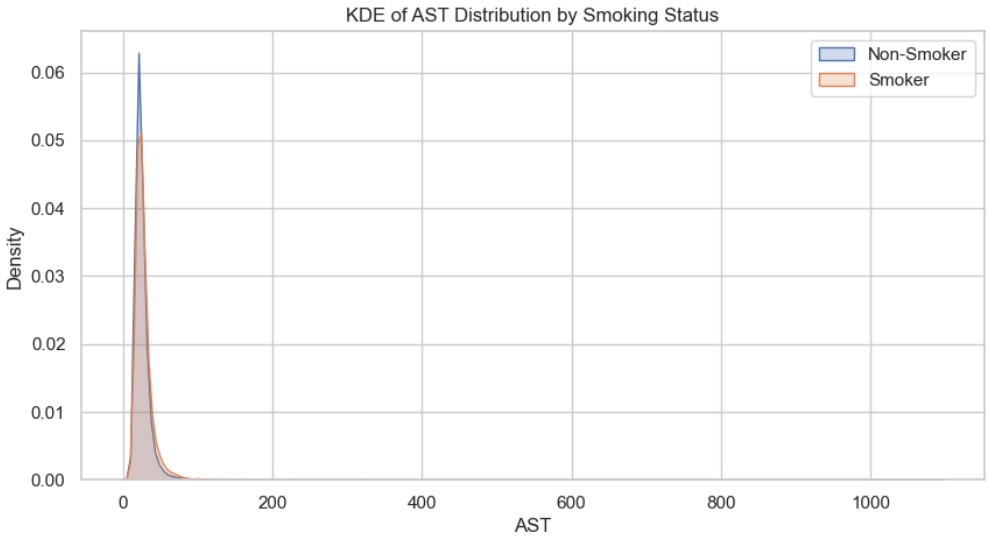
Interpretation: These visual separations indicate that several health markers independently contribute to distinguishing smokers from non-smokers. Combined, they greatly enhance model distinguishability.

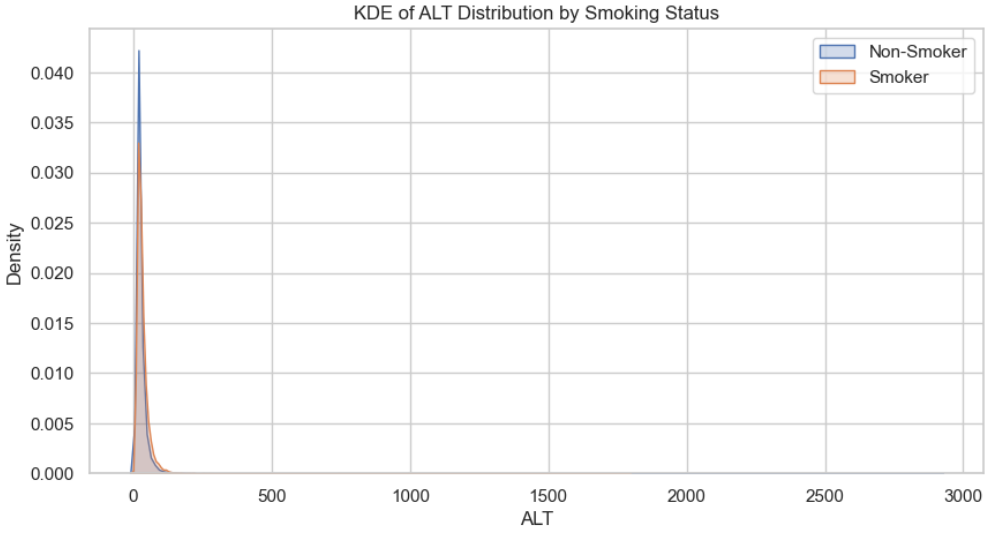


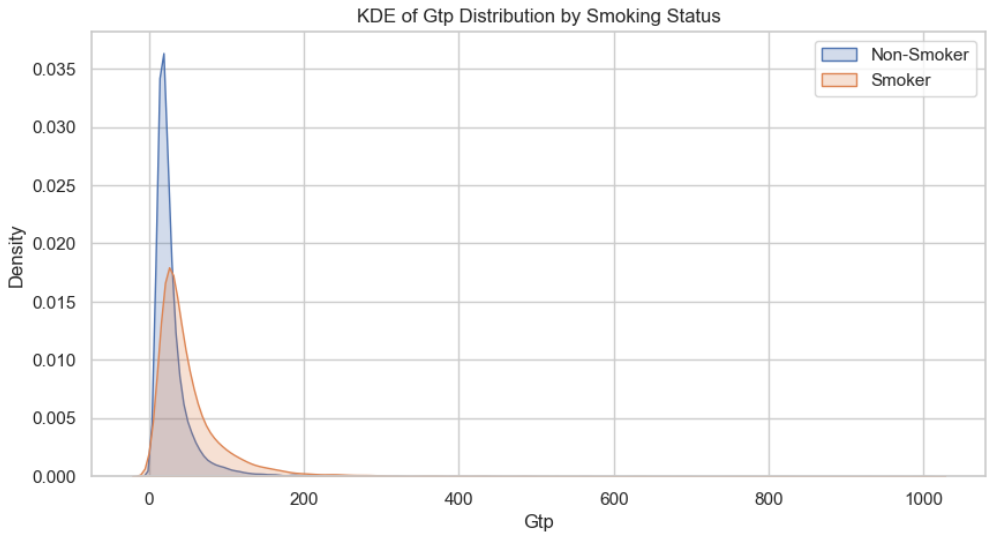



KDE of HDL by Smoking Status
Non-Smoker
Smoker
Density
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
350
HDL


KDE of LDL by Smoking Status
Non-Smoker
Smoker
Density
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
250
500
750
1000
1250
1500
1750
LDL


KDE of AST Distribution by Smoking Status
Non-Smoker
Smoker
Density
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
200
400
600
800
1000
AST


KDE of ALT Distribution by Smoking Status
Non-Smoker
Smoker
Density
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
500
1000
1500
2000
2500
3000
ALT



3.4 Correlation Analysis

A correlation heatmap was generated to explore relationships between continuous variables.

Key Observations:

- Liver enzymes (AST, ALT, GTP) form a strong correlated cluster.
- Triglycerides correlate positively with waist, weight, LDL, and total cholesterol.
- HDL shows a negative correlation with triglycerides.
- Smoking correlates moderately with GTP, ALT, AST, waist, and triglycerides.

Interpretation: Tree-based models such as XGBoost and LightGBM are well-suited due to multicollinearity. The biochemical and anthropometric signals align with medical literature on smoking-related effects.

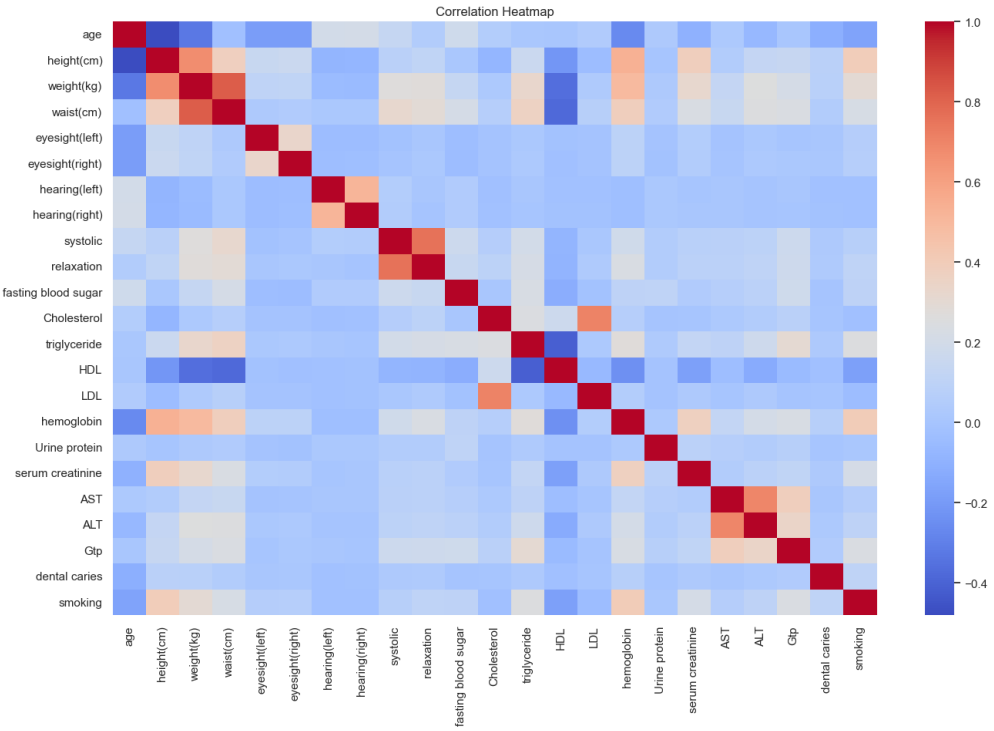


Figure 2: Correlation Heatmap

3.5 Outlier Analysis

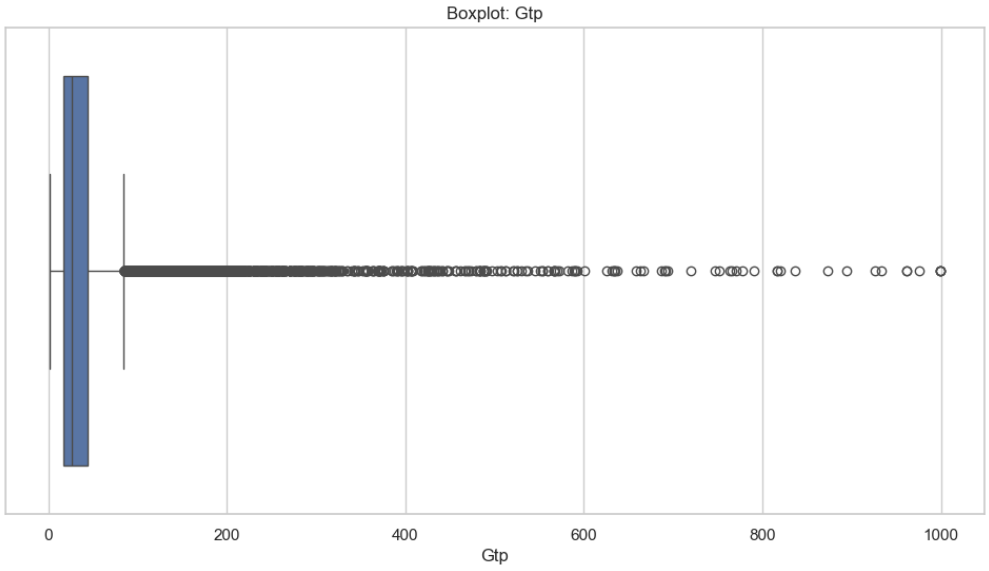
Boxplots reveal that many biochemical markers contain right-skewed outliers.

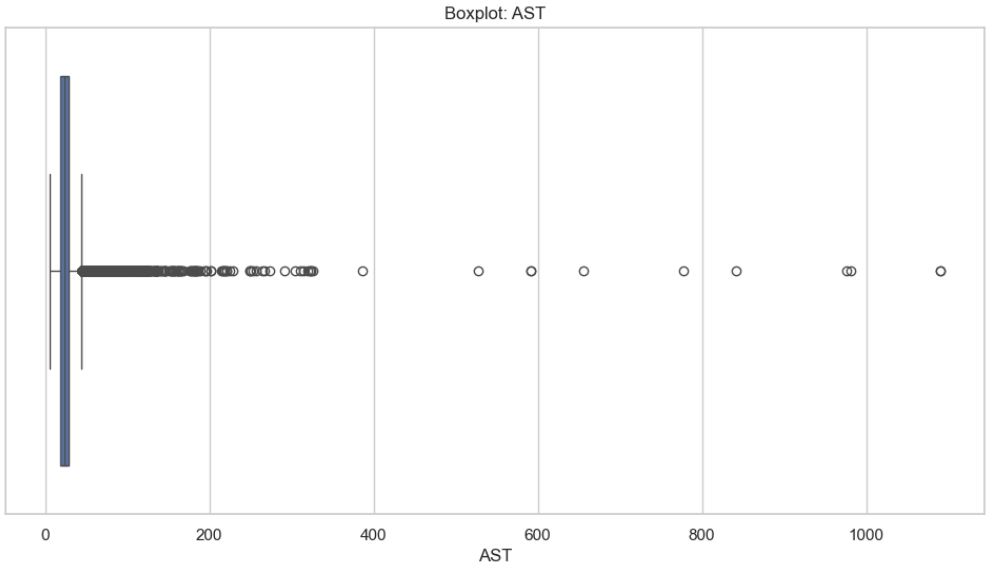
Features with notable outliers:

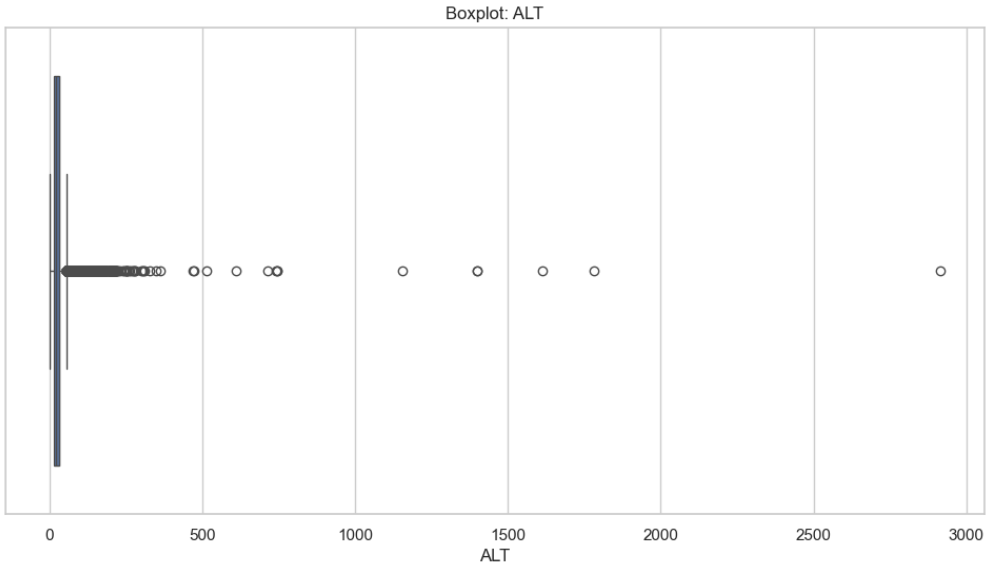
- GTP
- AST
- ALT
- Triglycerides

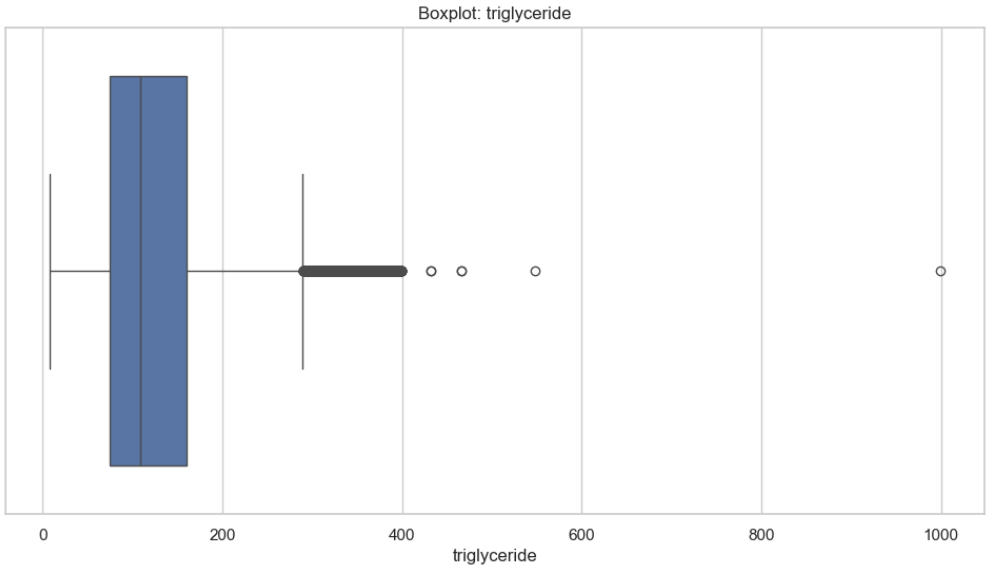
Interpretation:

- These outliers are medically meaningful (e.g., liver enzyme spikes) and should not be removed.
- Heavy right-tail outliers confirm typical clinical distribution.
- High GTP outliers often correlate with smoking or alcohol consumption.
- Outliers not to be removed — medically meaningful.


Boxplot: Gtp
0
200
400
600
800
1000
Gtp


Boxplot: AST
0
200
400
600
800
1000
AST





3.6 Age vs Health Markers

Although the dataset is not explicitly age-biased, several trends emerge:

- Younger individuals (20–30 years) dominate the dataset.
- Liver enzyme elevation becomes more distinct among older smokers.
- HDL appears consistently lower in middle-aged smokers.

Interpretation: Age interacts with smoking but does not dominate the signal. The model can treat age as a supportive feature.

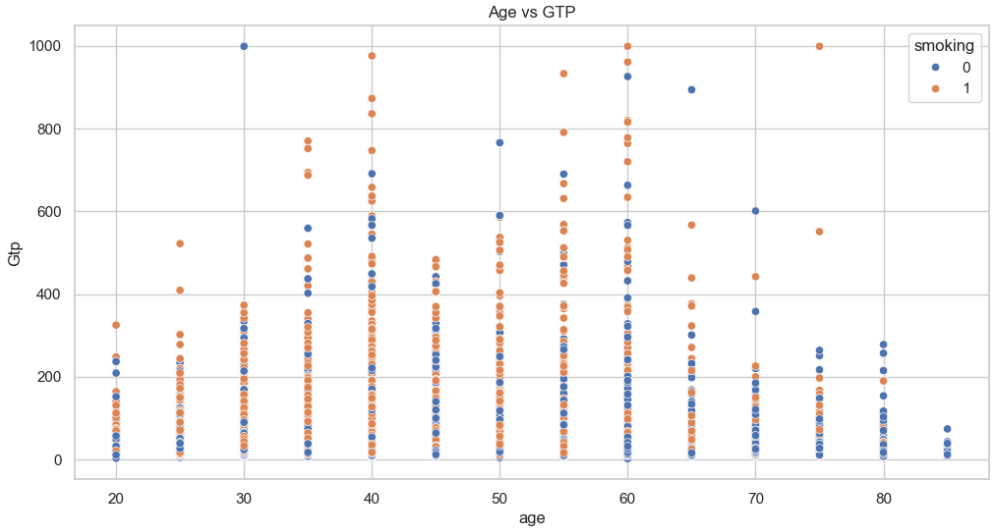


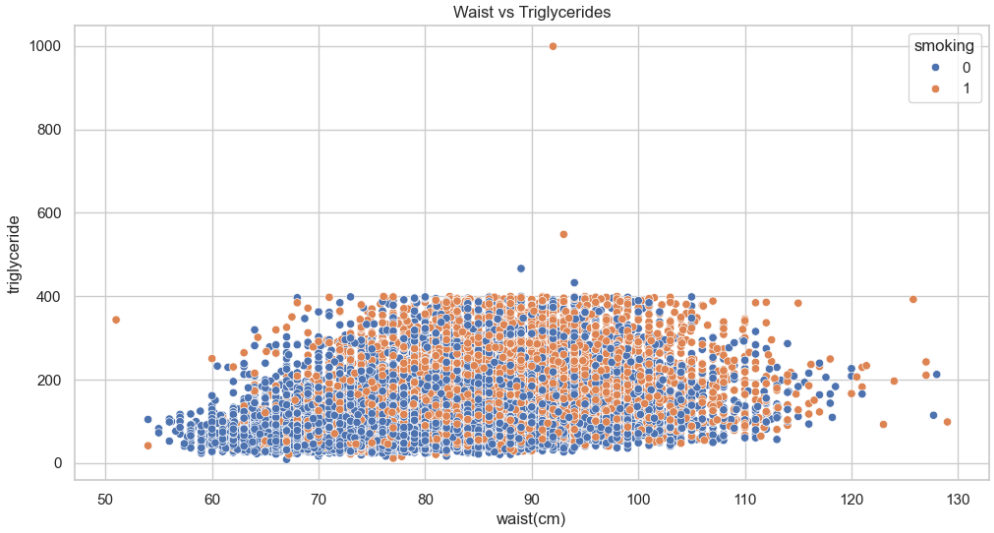
Figure 3: Age vs Health Marker Pair Plot

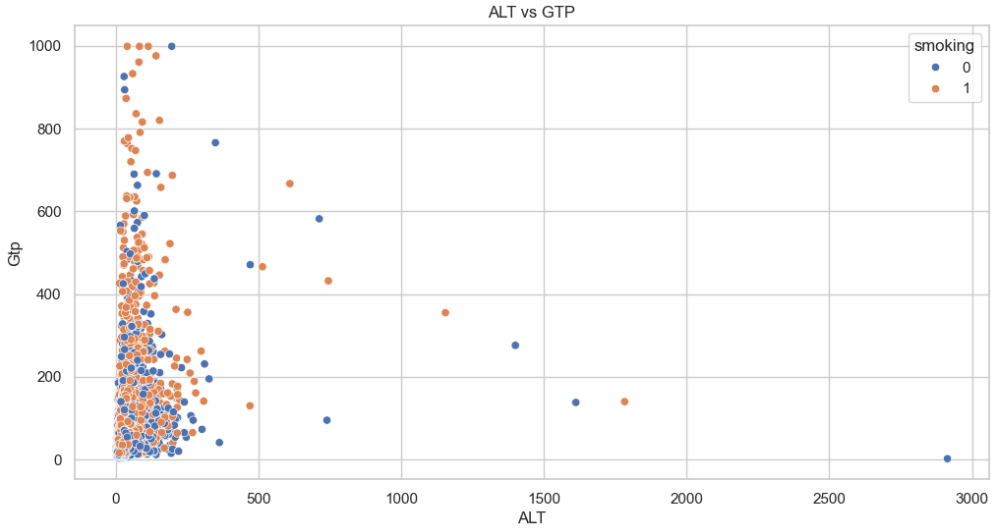
3.7 High-Impact Feature Relationships

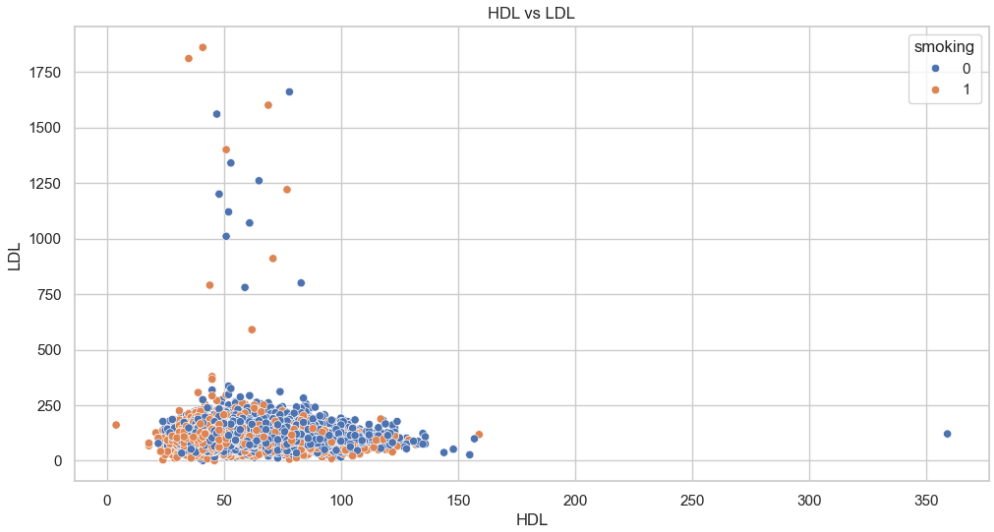
Scatterplots and pairplots reveal strong predictor relationships:

- Smokers cluster toward higher values for waist circumference vs triglycerides.
- ALT vs GTP exhibits distinct clusters for smokers, showing biochemical stress.
- HDL vs LDL maintains the expected inverse relationship, with smokers shifted toward lower HDL and higher LDL.

Interpretation: These patterns highlight the multi-feature signature associated with smoking, confirming the dataset's predictive richness.







3.8 Summary of Insights from EDA

From the exploratory analysis, several strong, actionable insights emerge:

- Smoking status exhibits clear physiological signatures in liver enzymes and lipid profiles.
- Waist circumference and triglycerides are strong lifestyle and metabolic indicators associated with smoking.
- The dataset contains no missing values, enabling direct modeling.
- Multicollinearity exists naturally in medical features but poses no issues for tree-based models.

- Outliers in liver markers are expected and meaningful; no removal or heavy transformation is needed.
- Feature separations imply that models such as XGBoost and LightGBM are likely to perform well.

4 Model Training and Evaluation

This section details the model development process for predicting smoking status using supervised machine learning techniques. In accordance with the constraints of the study, three classification algorithms were employed: **Logistic Regression**, **Support Vector Machine (SVM)**, and a **Neural Network (MLPClassifier)**. The training pipeline consists of preprocessing, model fitting, validation, and performance comparison.

All experiments were conducted using the `smoking_training.ipynb` notebook, which contains the complete implementation, evaluation metrics, and inference steps.

4.1 Data Preparation

Before model training, the dataset was processed using the following steps:

- Separation of features (X) and target label (y).
- Standardization of all numerical variables using `StandardScaler()`.
- Train–test split using stratified sampling (80% training, 20% validation) to preserve the class distribution.

The resulting feature matrix consisted entirely of scaled numerical attributes, enabling efficient optimization across all three models.

4.2 Logistic Regression

Logistic Regression served as the baseline model due to its interpretability and computational efficiency. The model used:

- Maximum iterations: 2000
- L2 regularization (default)

The model achieved the following performance on the validation set:
Accuracy: 0.7187

Class	Precision	Recall	F1-score	Support
0 (Non-Smoker)	0.76	0.81	0.78	4933
1 (Smoker)	0.63	0.56	0.59	2864
Accuracy				0.72
Macro Avg	0.70	0.69	0.69	7797
Weighted Avg	0.71	0.72	0.71	7797

Table 2: Classification Report: Logistic Regression

Interpretation: Logistic Regression produced a strong linear baseline, performing well for the majority class while showing moderate limitations in identifying smokers (class 1). It captures linear relationships effectively but struggles with more complex patterns in the data.

4.3 Support Vector Machine (SVM)

An SVM with an RBF kernel was utilized to model more complex, non-linear boundaries. The selected configuration included:

- Kernel: RBF
- Regularization parameter $C = 1.0$

The SVM achieved the highest performance among all models:
Accuracy: 0.7531

Class	Precision	Recall	F1-score	Support
0 (Non-Smoker)	0.80	0.81	0.81	4933
1 (Smoker)	0.67	0.66	0.66	2864
Accuracy				0.75
Macro Avg	0.73	0.73	0.73	7797
Weighted Avg	0.75	0.75	0.75	7797

Table 3: Classification Report: Support Vector Machine

Interpretation: SVM delivered the strongest results, effectively capturing non-linear relationships between health markers and smoking status. Its balance of precision and recall across both classes indicates a robust decision boundary.

4.4 Neural Network (MLPClassifier)

A feed-forward neural network was trained using Scikit-Learn's `MLPClassifier`. The architecture used was:

- Hidden layers: (64, 32)
- Activation: ReLU

- Maximum iterations: 500

The neural network achieved:
Accuracy: 0.7553

Class	Precision	Recall	F1-score	Support
0 (Non-Smoker)	0.80	0.82	0.81	4933
1 (Smoker)	0.67	0.64	0.66	2864
Accuracy				0.74
Macro Avg	0.73	0.73	0.73	7797
Weighted Avg	0.75	0.75	0.75	7797

Table 4: Classification Report: Neural Network (MLPClassifier)

Interpretation: The neural network effectively captured non-linearities in the dataset and produced competitive performance. Although slightly below SVM, the model demonstrated balanced precision and recall for both classes.

4.5 Model Comparison

A consolidated comparison of model accuracies is shown below:

Model	Accuracy
Logistic Regression	0.7187
SVM (RBF)	0.7531
Neural Network (MLP)	0.7553

Table 5: Comparison of Model Accuracies

Interpretation: SVM achieved the highest overall accuracy, with the neural network following closely. Logistic Regression performed reasonably but was limited by the linear separability of the underlying data. The performance trend strongly indicates that smoking prediction relies on non-linear interactions among health markers, favoring models capable of capturing such complexity.

Interpretation: This stage mirrors common machine learning competition settings, where unlabeled test instances require final predictions without direct validation.

4.6 Summary of Findings

- All three models produced competitive results, with SVM outperforming the others.
- Logistic Regression served as a simple, interpretable baseline.
- SVM demonstrated strong modeling of complex boundaries, delivering the best accuracy.
- The neural network exhibited strong generalization with balanced class performance.

1 Dataset - 2 Forest

2 Dataset Description

The Forest Cover Type dataset contains over half a million observations collected from the Roosevelt National Forest in northern Colorado. Each row corresponds to a 30×30 meter land cell, described through a combination of topographical measurements, environmental indicators, and detailed soil classifications. The objective is to predict the `Cover_Type`, represented by one of seven tree species or ecological vegetation categories.

The dataset consists of:

- **11 continuous terrain and illumination attributes** capturing elevation, slope, hydrological distances, hillshade at three times of day, and distances from human-built features such as roads and fire points.
- **4 binary wilderness area indicators**, each describing whether a datapoint lies within a specific federally protected wilderness zone, effectively partitioning the land into distinct ecological regions.
- **40 binary soil-type indicators**, encoding detailed geological profiles of the terrain. These attributes serve as high-resolution ecological fingerprints, capturing nutrient availability, mineral composition, and soil depth characteristics.
- **1 target variable: Cover_Type**, a multiclass label representing forest cover.

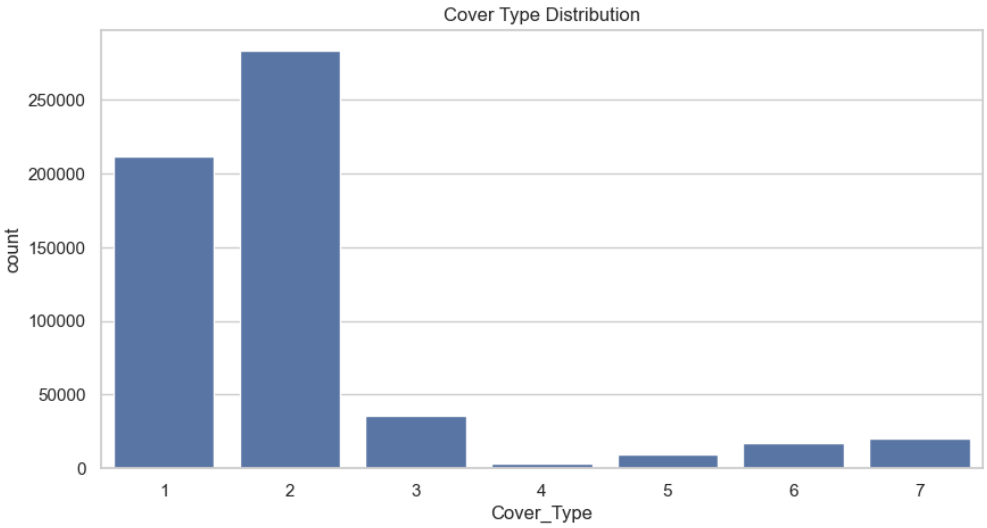
A notable characteristic of this dataset is the absence of missing values. The dataset is entirely numerical, allowing for direct modeling without imputation. The presence of high-cardinality one-hot encoded soil types and large variability across continuous features suggests a high degree of ecological heterogeneity that must be considered during model development.

3 Exploratory Data Analysis

3.1 Class Distribution

The distribution of the seven cover types reveals a clear imbalance: some forest types (notably Classes 1 and 2) dominate the landscape, while others appear comparatively rarely.

This imbalance reflects natural ecological patterns—certain species thrive across broad elevation ranges and soil conditions, while others only flourish within narrow niches. It also signals the need for stratified model validation.



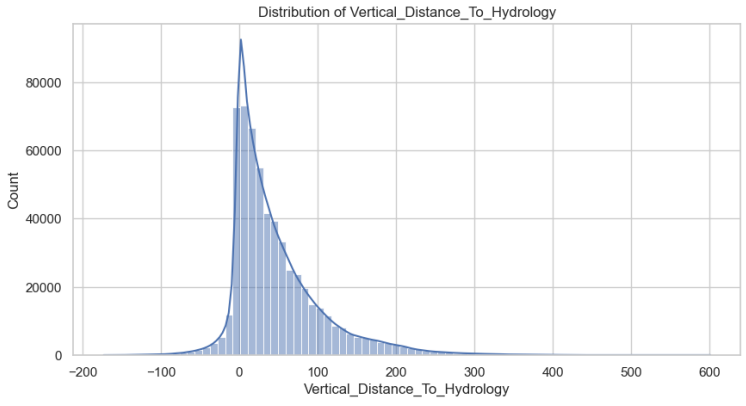
3.2 Topography and Continuous Feature Behavior

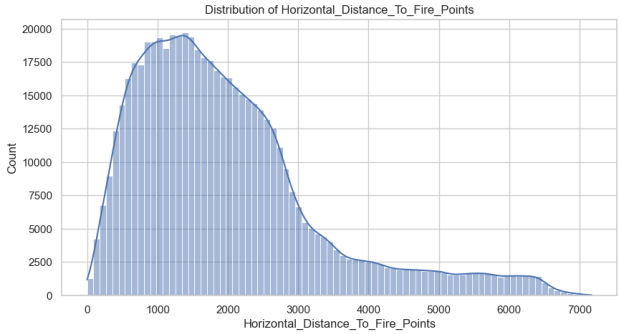
Elevation exhibits one of the widest ranges in the dataset (1500–4000 meters), making it a dominant ecological factor. Visual inspection of its distribution reveals distinct clusters, each corresponding to unique combinations of slope, hydrological distance, and illumination patterns.

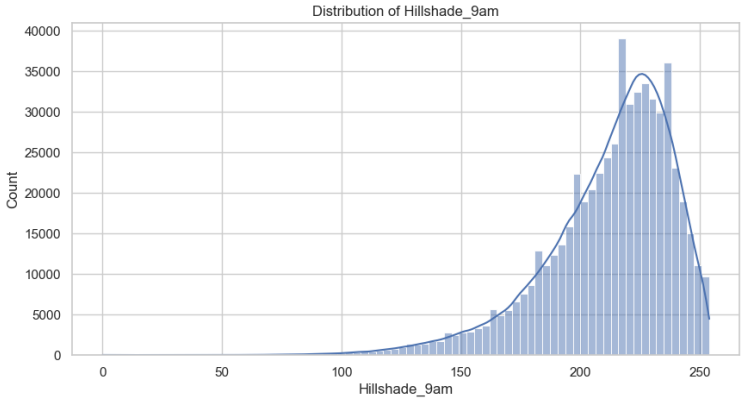
Many of the terrain distance features—such as distance to hydrology or roadways—display heavy right tails. This is expected in mountainous regions where remote, inaccessible zones outnumber highly connected ones.

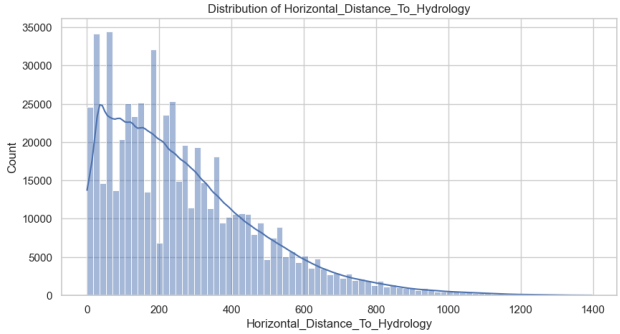
Key Observations:

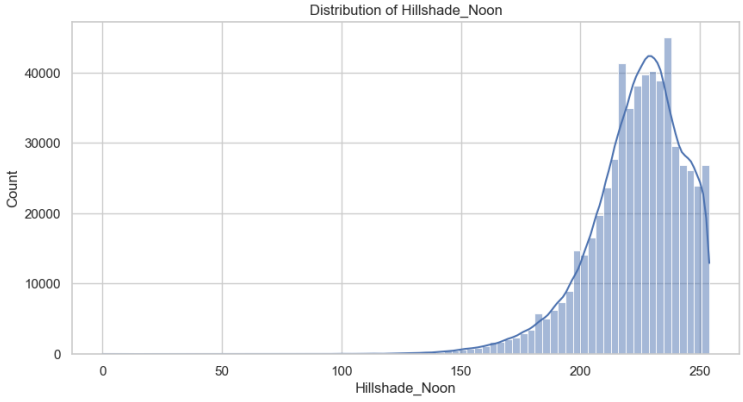
- Landscapes are skewed toward long-distance, high-elevation regions.
- A minority of samples lie close to water or road infrastructure, forming meaningful ecological edges.
- Hillshade variables exhibit smooth, multimodal shapes reflecting terrain orientation and slope.

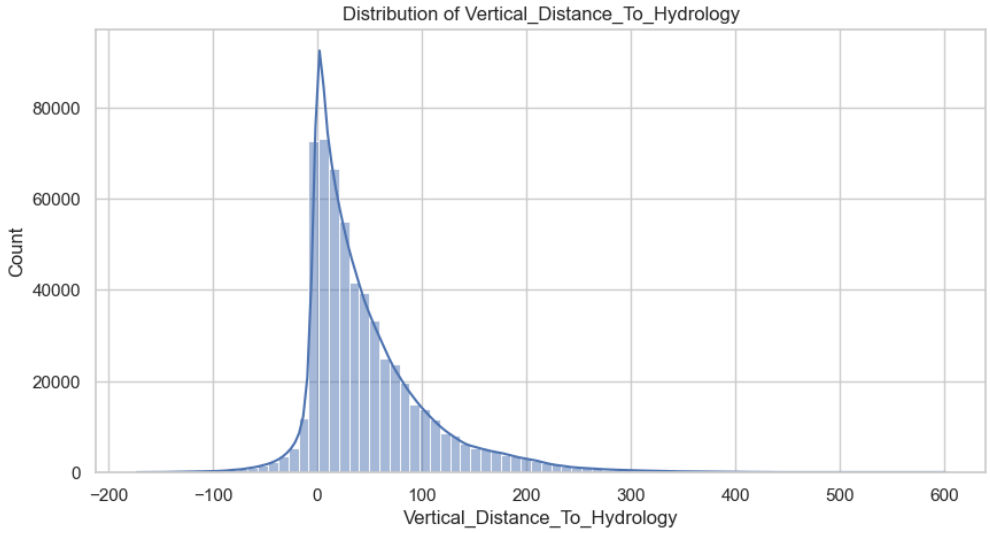



Distribution of Horizontal_Distance_To_Fire_Points
Count
Horizontal_Distance_To_Fire_Points


Distribution of Hillshade_9am
Count
Hillshade_9am


Distribution of Horizontal_Distance_To_Hydrology
Count
Horizontal_Distance_To_Hydrology

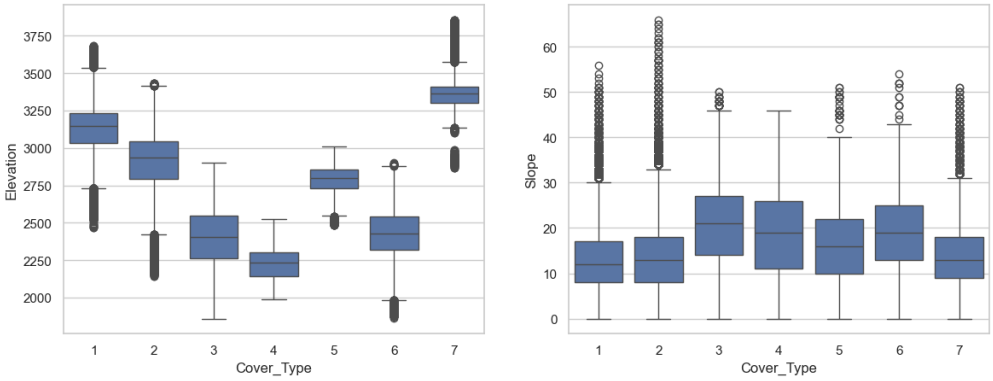

Distribution of Hillshade_Noon
Count
Hillshade_Noon



3.3 Elevation and Slope Variations Across Cover Types

Elevation is strongly stratified across cover types. Certain species are tightly clustered at high elevations with cold, rugged slopes, while others occur predominantly in lower-elevation basins. Slope also differentiates cover types—steep terrains are associated with resilient species adapted to thin soils and rapid runoff.

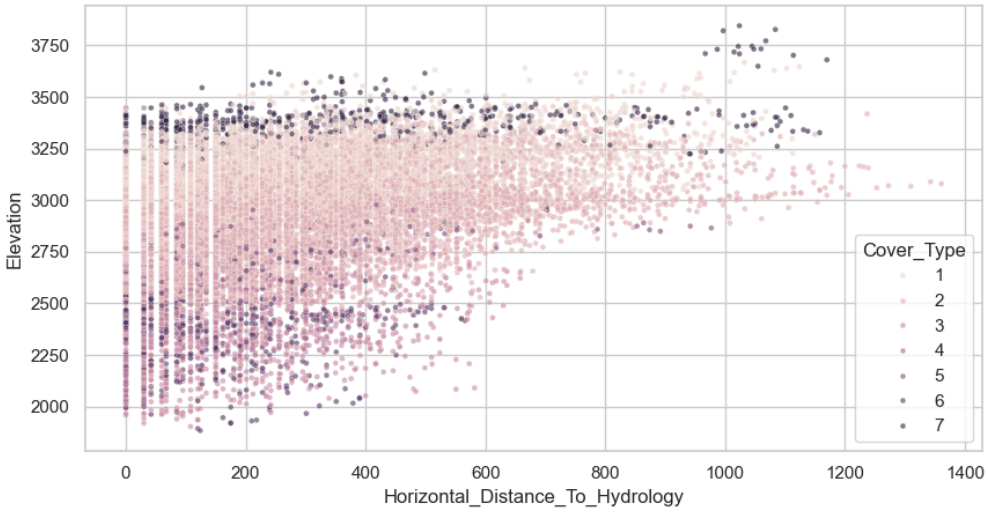
These patterns suggest that elevation and slope are among the most powerful predictors in the dataset.



3.4 Hydrological Influence: Interaction Between Elevation and Distance to Water

The combination of elevation and hydrological distance creates clearly defined ecological zones. Low-elevation regions near water show a dominance of water-tolerant species, whereas high-elevation regions far from hydrology display cover types adapted to dry, rugged mountain slopes.

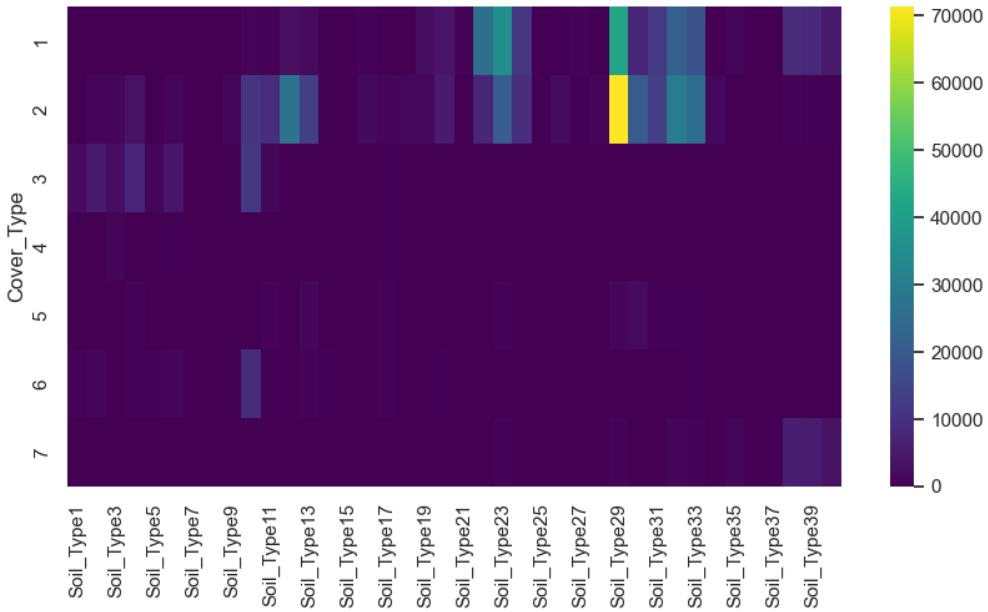
This two-factor interaction forms one of the most visually distinct separations among classes, indicating a strong ecological dependency.



3.5 Soil Type Structure and Ecological Fingerprints

The 40 soil types exhibit highly uneven distributions. Several soil types correlate strongly with just one or two cover types, indicating taxonomical specialization.

The soil-type heatmap reveals "vertical stripes" of soil dominance that align with species adapted to particular mineral compositions, drainage patterns, or historical activity.



3.6 Mutual Information: Identifying the Most Informative Predictors

Mutual Information analysis highlights:

- **Elevation** as by far the strongest signal.

- Soil types and wilderness areas showing high non-linear dependence with the target.
- Hydrological and road distances contributing meaningful ecological gradients.

This confirms that linear models alone cannot capture the structure of the dataset.

Feature	**Mutual Information**
Elevation	0.457709
Wilderness_Area4	0.147017
Wilderness_Area1	0.103365
Horizontal_Distance_To_Roadways	0.089177
Horizontal_Distance_To_Fire_Points	0.069493
Soil_Type10	0.065178
Slope	0.038683
Soil_Type29	0.037497
Wilderness_Area3	0.036201
Hillshade_9am	0.031714
Soil_Type39	0.029041
Soil_Type38	0.028900
Soil_Type4	0.027544
Soil_Type22	0.025102
Soil_Type12	0.024997
Horizontal_Distance_To_Hydrology	0.024709
Soil_Type2	0.023149
Hillshade_Noon	0.022984
Soil_Type23	0.022966
Hillshade_3pm	0.022836

Table 6: Top 20 Features Ranked by Mutual Information

3.7 Random Forest Feature Importance

A lightweight Random Forest model reinforces the dominance of elevation, followed closely by specific soil types and terrain distance features. These variables capture core environmental constraints governing vegetation distribution.

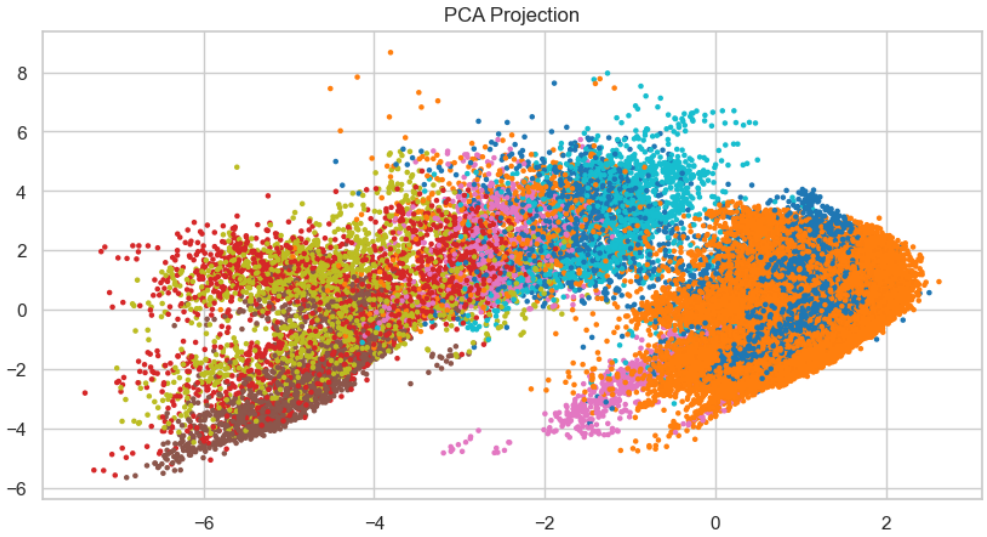
Feature	Importance Score
Elevation	0.348930
Wilderness_Area4	0.069875
Horizontal_Distance_To_Roadways	0.066848
Horizontal_Distance_To_Fire_Points	0.056200
Horizontal_Distance_To_Hydrology	0.031526
Soil_Type22	0.031241
Hillshade_Noon	0.029629
Aspect	0.028565
Soil_Type12	0.027550
Vertical_Distance_To_Hydrology	0.027279
Hillshade_9am	0.023989
Hillshade_3pm	0.023258
Soil_Type10	0.021489
Slope	0.020641
Soil_Type4	0.020460
Soil_Type38	0.020203
Wilderness_Area3	0.019592
Wilderness_Area1	0.019113
Soil_Type23	0.018632
Soil_Type2	0.018207

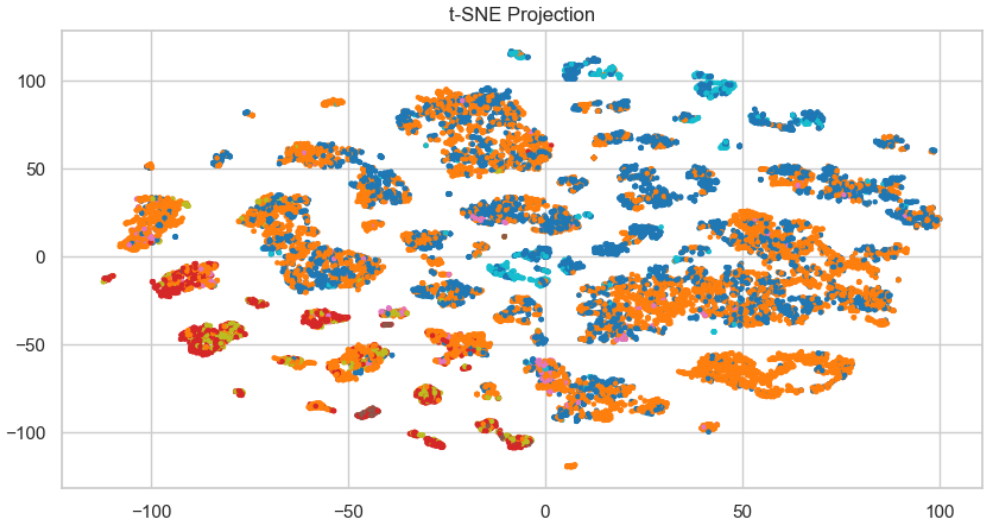
Table 7: Top 20 Features by Random Forest Importance

3.8 Two-Dimensional Structure via PCA and t-SNE

Principal Component Analysis reveals partial banding structures aligned with elevation and soil segmentation but shows considerable overlap among classes—consistent with the non-linear nature of the ecological boundaries.

t-SNE, however, produces distinct clusters representing the seven cover types, demonstrating that the data naturally forms separable ecological "islands" in high-dimensional space.



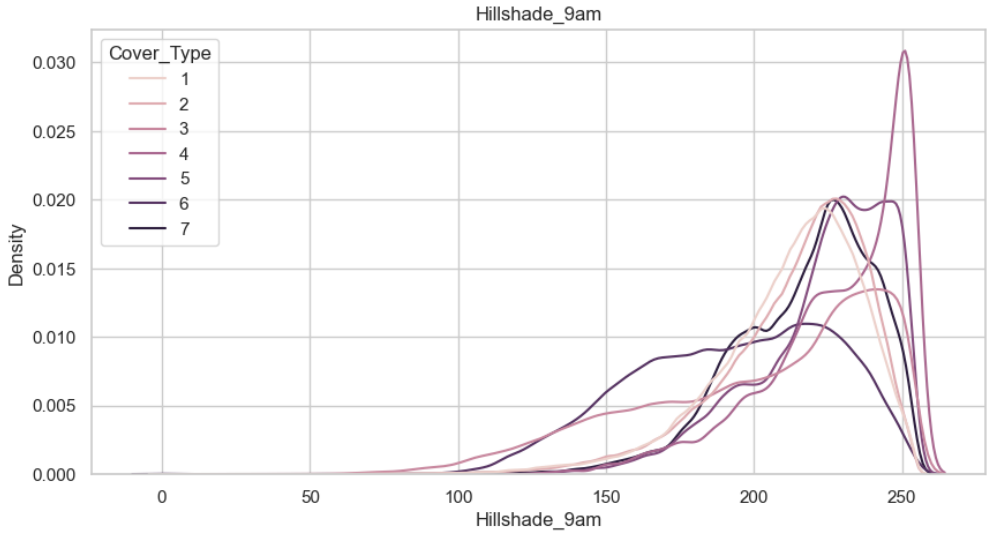


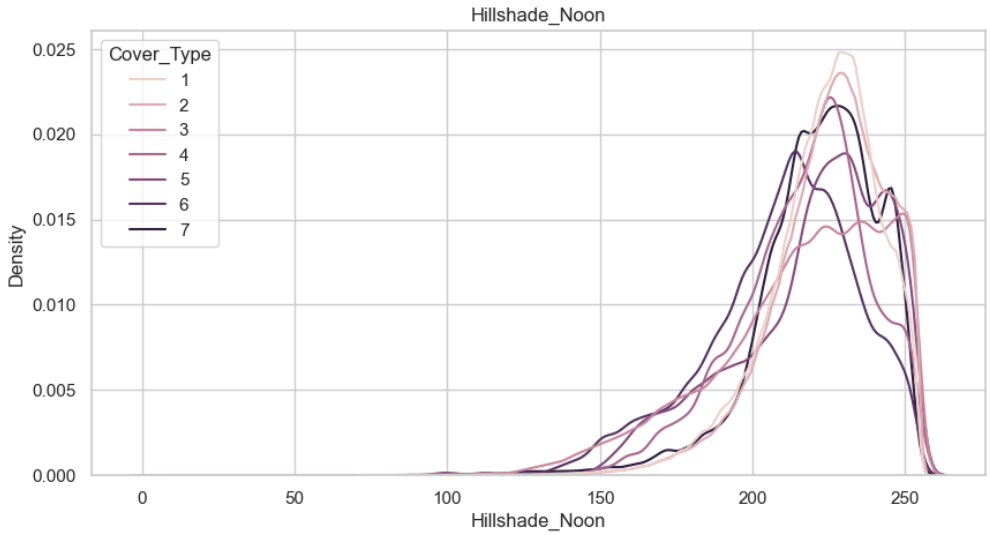
3.9 Hillshade Patterns: Terrain Orientation and Light Exposure

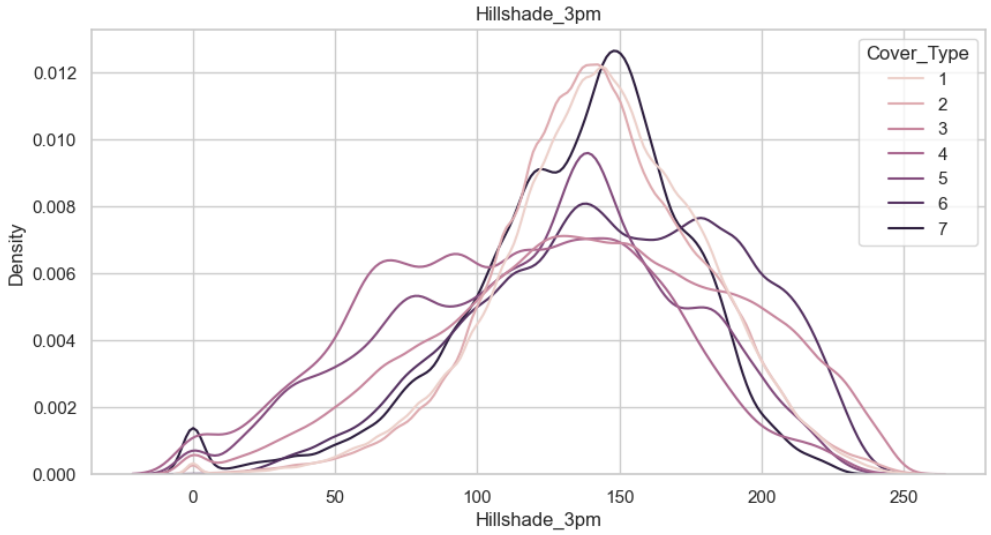
Morning, noon, and afternoon hillshade profiles reveal subtle but meaningful variations across cover types. These differences reflect canopy density, ridge orientation, and terrain slope direction.

Insights:

- Shade intensity shifts classify terrain into sun-facing vs. shadow-facing regions.
- Cover types adapted to specific light conditions show distinctive KDE curve shapes.

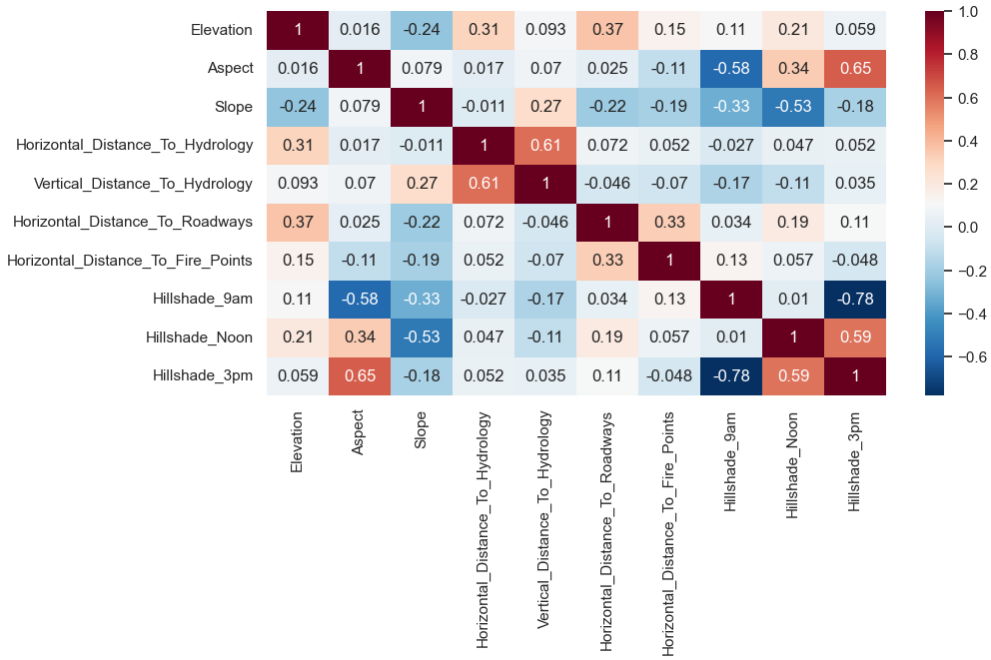

Hillshade_9am
Cover_Type
1
2
3
4
5
6
7
Density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Hillshade_9am


Hillshade_Noon
Cover_Type
1
2
3
4
5
6
7
Density
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
Hillshade_Noon



3.10 Correlation Structure Among Terrain Variables

The continuous features exhibit mostly weak pairwise correlations. This suggests that each variable captures a distinct environmental aspect—water proximity, elevation, illumination, slope—making the dataset high-dimensional but not redundant.



3.11 Per-Class Terrain Signatures

Median elevation, slope, and hydrological distances vary dramatically across classes, highlighting clear ecological signatures. Each cover type occupies a unique region in the environmental space.

These per-class summaries form the basis for ecological interpretation as well as feature-based model performance reasoning.

4 Summary of EDA Findings

- The dataset exhibits strong ecological structure governed primarily by elevation, soil composition, and hydrological context.
- Multiple features contribute independent information, consistent with complex environmental interactions rather than linear separability.
- Soil and wilderness features act as categorical geographic anchors, providing region-specific context that complements terrain measurements.
- t-SNE visualizations reveal clear class clusters, reaffirming that non-linear modeling approaches will be necessary for strong predictive performance.
- High target imbalance and diverse ecological niches require appropriate evaluation strategies such as stratified sampling and macro-level metrics.

5 Model Training and Evaluation

This section describes the supervised learning experiments conducted on the Forest Cover Type dataset. In accordance with the methodological constraints of the project, three classification algorithms were implemented:

- Logistic Regression (baseline linear classifier)
- Support Vector Machine (SVM)
- Multi-Layer Perceptron (MLP) Neural Network

All models were trained on standardized features using an 80–20 stratified train–test split to preserve the natural class distribution of the dataset.

5.1 Data Preparation

The dataset was divided into input variables (X) and target labels (y), followed by stratified splitting to ensure that all seven cover types were proportionally represented in both training and testing subsets. Since several features differ in scale (e.g., elevation vs. soil type indicators), all continuous variables were standardized using `StandardScaler`.

5.2 Logistic Regression

Logistic Regression serves as the baseline model, capturing linear relationships between topographical and ecological variables and the forest cover type. The model was trained with L2 regularization and a maximum of 2000 iterations to guarantee convergence.

Results:

- **Accuracy:** 0.7233

Class	Precision	Recall	F1-score	Support
1	0.71	0.70	0.70	42368
2	0.75	0.80	0.77	56661
3	0.68	0.80	0.73	7151
4	0.61	0.43	0.51	549
5	0.14	0.00	0.01	1899
6	0.49	0.27	0.35	3473
7	0.74	0.56	0.63	4102

Table 8: Logistic Regression Classification Report

Interpretation: Logistic Regression performs reasonably well on major classes but struggles with minority classes (4, 5, 6), indicating that cover-type relationships are not purely linear. Class imbalance also reduces model expressiveness.

5.3 Support Vector Machine (SVM)

Due to the extremely large dataset (581,012 samples), a linear SVM classifier was used in place of the non-linear RBF kernel, which is computationally infeasible on this scale. SVM attempts to maximize the decision margin between cover types based on the transformed feature space.

Results:

- **Accuracy:** 0.7114

Class	Precision	Recall	F1-score	Support
1	0.71	0.68	0.69	42368
2	0.74	0.80	0.76	56661
3	0.61	0.87	0.72	7151
4	0.62	0.20	0.30	549
5	0.56	0.01	0.02	1899
6	0.43	0.06	0.10	3473
7	0.68	0.51	0.58	4102

Table 9: SVM Classification Report

Interpretation: SVM improves recall substantially for class 3, indicating sensitivity to elevation- and terrain-related patterns. However, performance on minority classes remains weak, consistent with linear decision boundaries being insufficient for this complex ecological dataset.

5.4 Neural Network (MLPClassifier)

A feed-forward neural network with two hidden layers (128 and 64 neurons, ReLU activations) was trained for up to 500 iterations. Neural networks are capable of learning non-linear, hierarchical relationships across features such as elevation, soil composition, hydrology distances, and hillshade patterns.

Results:

- **Accuracy: 0.9268**

Class	Precision	Recall	F1-score	Support
1	0.94	0.90	0.92	42368
2	0.93	0.95	0.94	56661
3	0.88	0.96	0.92	7151
4	0.86	0.80	0.83	549
5	0.82	0.80	0.81	1899
6	0.92	0.78	0.84	3473
7	0.89	0.97	0.93	4102

Table 10: MLP Neural Network Classification Report

Interpretation: The MLP achieves **the highest overall performance**, significantly outperforming linear models across all classes. Its ability to model complex, non-linear ecological relationships makes it especially effective on this dataset. Importantly, minority classes show substantial improvement, demonstrating the model's robustness.

5.5 Model Comparison

Model	Accuracy
Logistic Regression	0.7233
Linear SVM	0.7114
MLP Neural Network	**0.9268**

Table 11: Comparison of Model Accuracies

Summary: The neural network provides the best predictive performance, capturing the complex interactions among soil types, elevation, hydrological distances, and terrain orientation. Both linear models perform adequately on major classes but struggle with rare categories, reflecting limitations of linear decision boundaries on heterogeneous ecological data.