Multiclass Classification of Forest Cover Type

Antonio Squicciarini

August 1, 2025

Problem Introduction

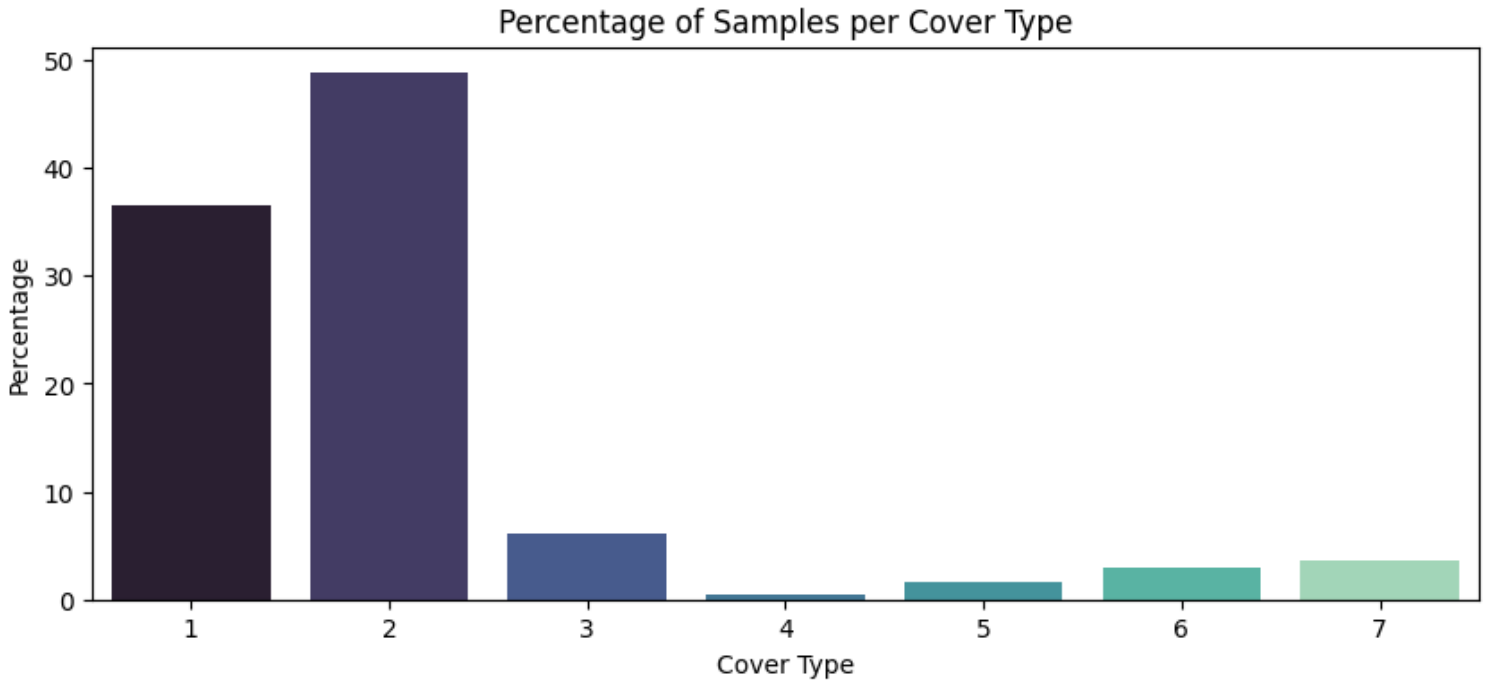
- Predict forest cover type using cartographic features.
- 581,012 observations from Roosevelt National Forest (Colorado).
- 54 features: 10 quantitative + 44 binary (soil type, wilderness areas).
- 7 forest cover types to classify (e.g., Spruce/Fir, Lodgepole Pine).

Dataset Description

- No missing values or duplicates.
- Terrain, hydrology, hillshade, and encoded location data.
- Target distribution is imbalanced.
- Originally published by Blackard and Dean (1998).
- Source: `https://archive.ics.uci.edu/dataset/31/covertype`

Class Distribution

- Classes 1 and 2 dominate.
- Classes 4 and 5 underrepresented.
- Consider resampling or class weights.

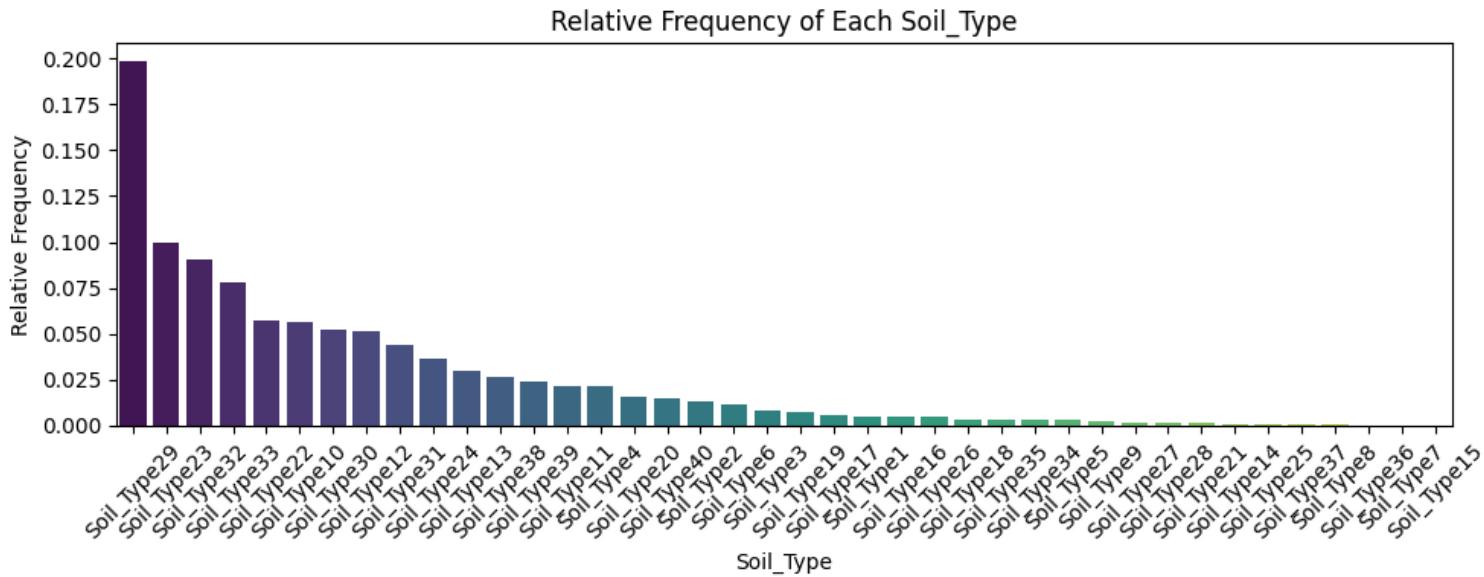


Descriptive Statistics (Quantitative)

- Elevation: 1859–3858 m, mean $\sim$2960 m.
- Slope: median 13, max 66.
- Horizontal Distance to Hydrology: mean $\sim$269 m.
- Hillshade Noon: peaks near 226.
- Data spans varied topography.

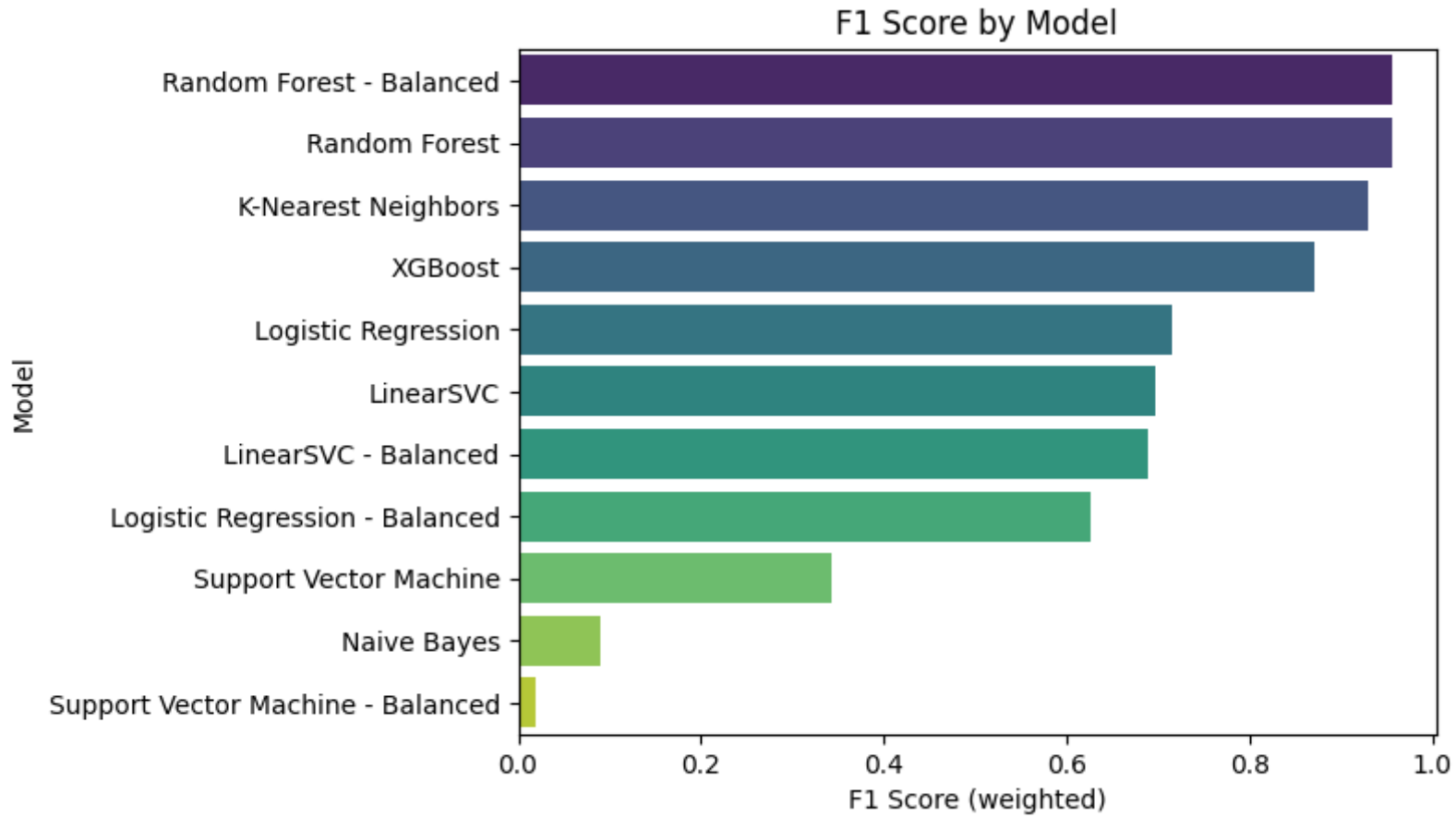
Binary Feature Distributions

- **Soil Types:** 40 categories, mutually exclusive.
- **Wilderness Areas:** 4 classes, 2 dominate (˜85%).
- These features have high predictive value.

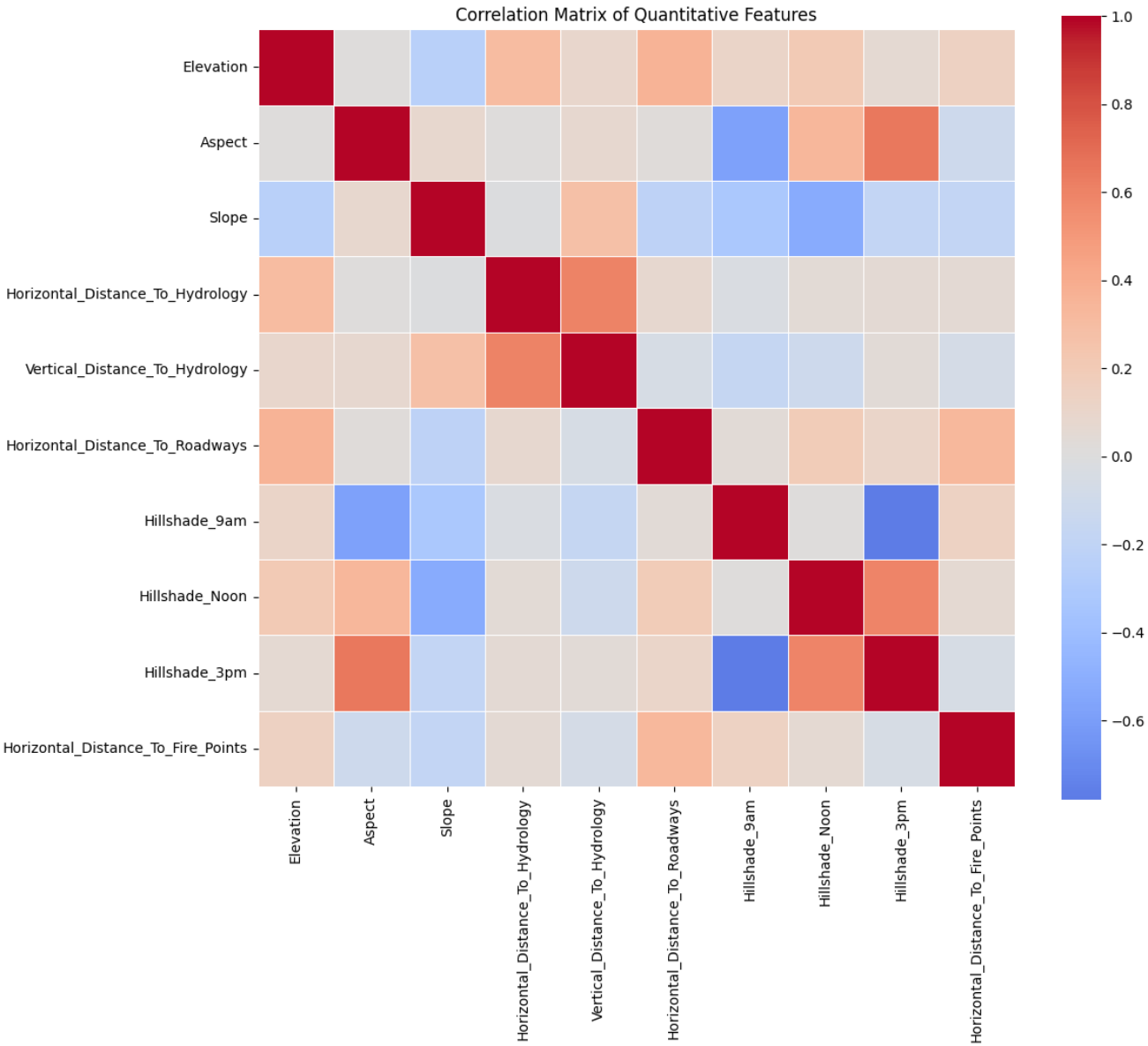


Correlation and Mutual Information

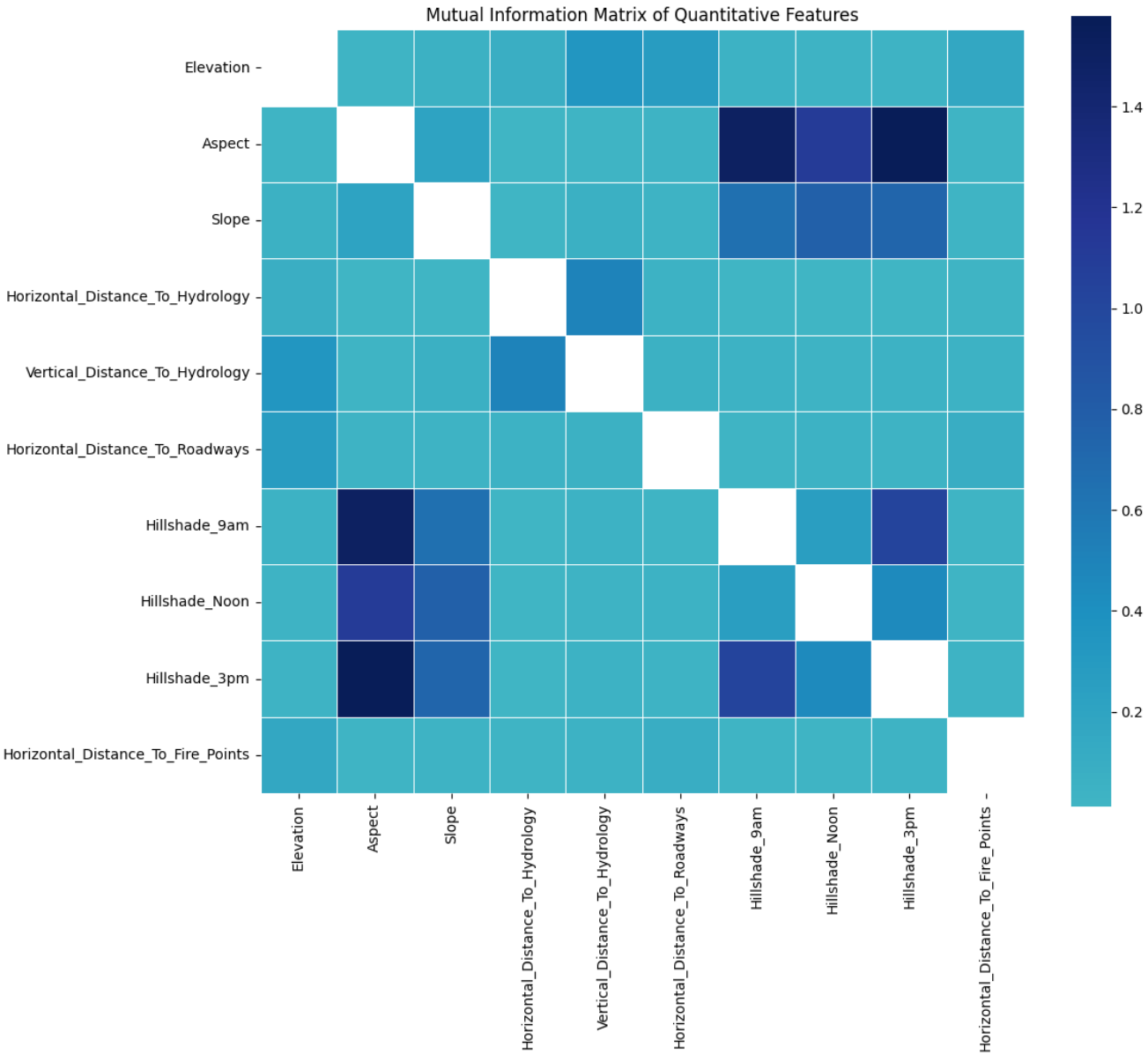
- Weak linear correlations.
- Strongest: Hillshade 9am & 3pm.
- Mutual Info: nonlinear relationships e.g., Aspect, Elevation.
- Useful for tree-based models.



Correlation Matrix

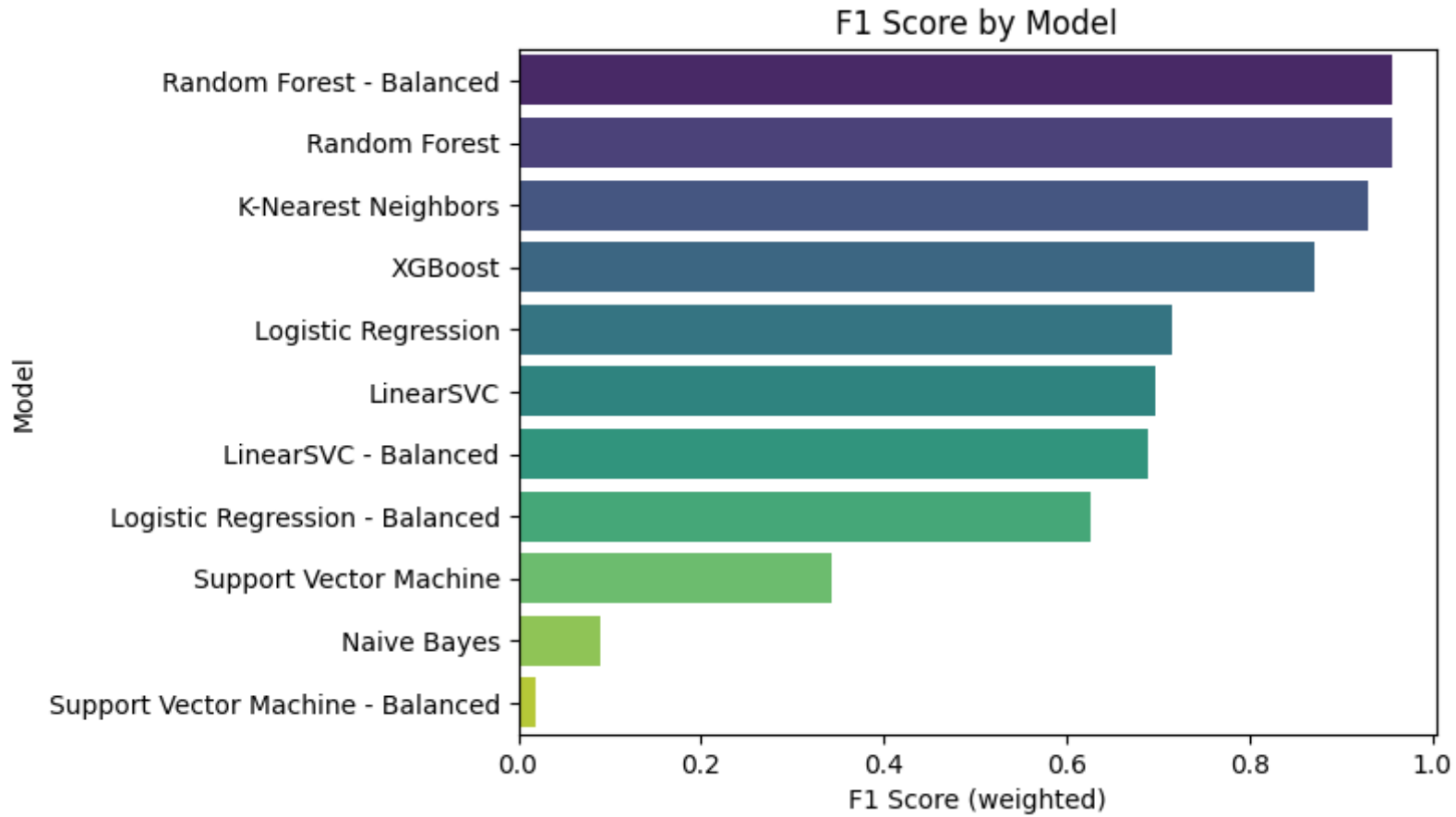


Mutual Information Matrix



Mutual Information Matrix

- Weak linear correlations.
- Strongest: Hillshade 9am & 3pm.
- Mutual Info: nonlinear relationships e.g., Aspect, Elevation.
- Useful for tree-based models.

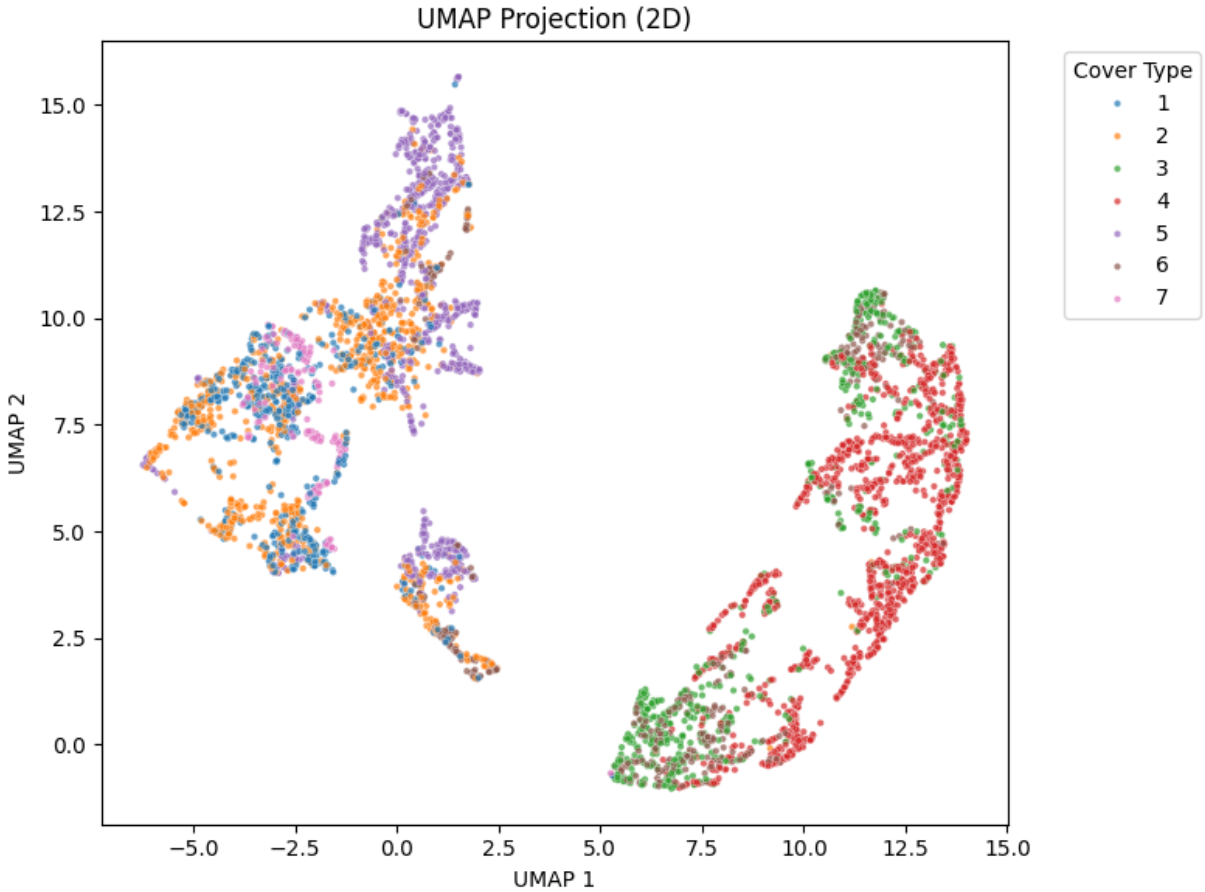


Feature Distributions by Cover Type

- Elevation: strong class separation.
- Aspect, Slope: overlapping, subtle trends.
- Roadway and Fire Point distances: class clusters.

Dimensionality Reduction

- **PCA:** global trend, low separation.
- **t-SNE:** good local clustering.
- **UMAP:** best balance, clear groups.
- Each method reveals different structure.



Model Selection

- **Tree-based models:** Random Forest, XGBoost.
- **KNN:** good performance, not scalable.
- **Linear models:** baseline only.
- **SVM & Naive Bayes:** poor results due to feature structure.

Modeling Pipeline

- Stratified train-test split (80-20).
- Scaled quantitative features.
- Concatenated with binary features.
- Evaluated on Accuracy, F1, Recall, G-Mean.
- **Note:** Given data imbalance, F1 score is preferred to reduce bias toward majority classes.

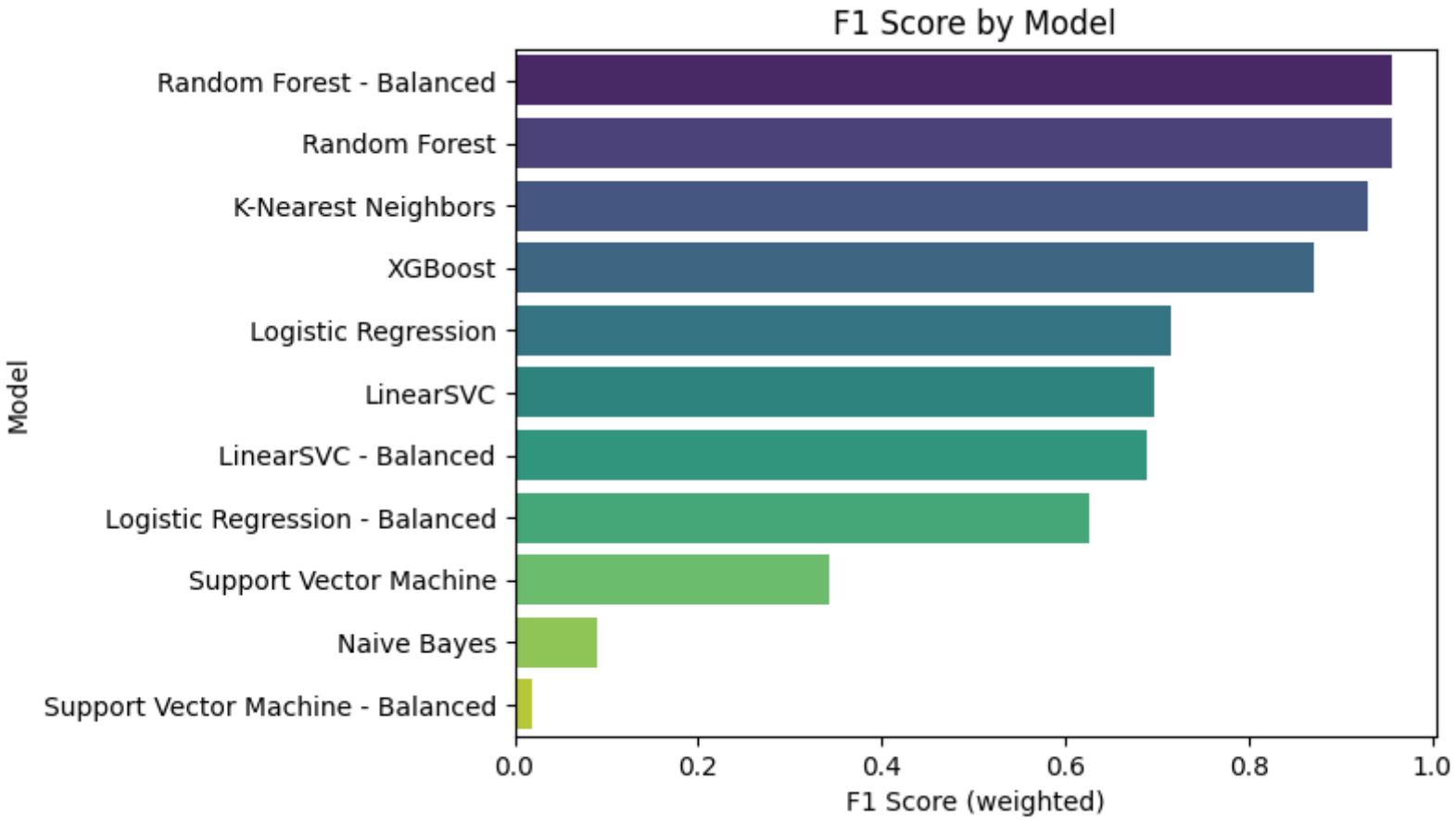
Top Model Results

- **Random Forest Balanced:**
  - Accuracy: 95.6%, F1: 95.6%, G-Mean: 94.8%
  - Recall: 90.8%, Specificity: 98.9%
- **XGBoost:**
  - Accuracy: 87.2%, F1: 87.1%, G-Mean: 89.9%
  - Recall: 83.3%, Specificity: 96.9%
- **KNN:**
  - Accuracy: 92.9%, F1: 92.9%, G-Mean: 92.2%
  - Recall: 86.3%, Specificity: 98.4%

Performance Comparison Table

Model	Accuracy	F1 Score	G-Mean	Recall	Specificity
Random Forest - Balanced	0.96	0.96	0.95	0.91	0.99
Random Forest	0.96	0.96	0.95	0.91	0.99
K-Nearest Neighbors	0.93	0.93	0.92	0.86	0.98
XGBoost	0.87	0.87	0.90	0.83	0.97
Logistic Regression	0.73	0.72	0.69	0.51	0.93
LinearSVC	0.71	0.70	0.65	0.45	0.93
LinearSVC - Balanced	0.68	0.69	0.76	0.63	0.93
LogReg - Balanced	0.60	0.63	0.81	0.71	0.92
SVM	0.43	0.34	0.49	0.27	0.88
Naive Bayes	0.11	0.09	0.63	0.46	0.87
SVM - Balanced	0.07	0.02	0.55	0.36	0.86

Performance Comparison



- Random Forest and KNN lead.
- Linear models fall short.
- Balanced versions help recall.

Future Work

- **Hyperparameter Tuning:** GridSearchCV, cross-validation.
- **Interpretability:** SHAP, permutation importance.
- **Class Balance:** SMOTE, class-weighting.
- **Pipeline Integration:** Feature selection, dimensionality reduction.