Forest Cover Type Prediction Project

Adrian Lariani, Aidan Mayes Poduslo

Problem Statement

Goal:

- Predict the forest cover type from cartographic variables using machine learning models

Dataset:

- Source: UCI (UC Irvine) Forest Cover Type - https://archive.ics.uci.edu/dataset/31/covertype
- Samples: 581,012 rows
- Features: 54 (10 continuous, 44 binary/categorical)
- Classes: 7 forest cover types

Objective:

- Build accurate ML models
- Interpret predictions to understand key features driving cover type classifications

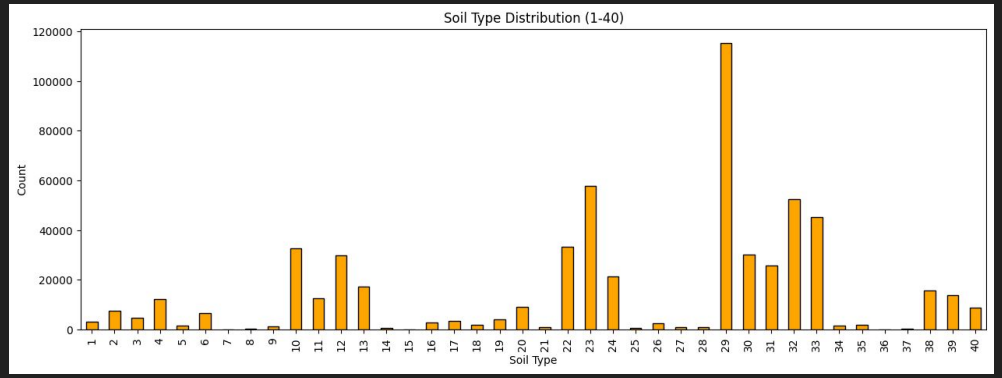
Dataset

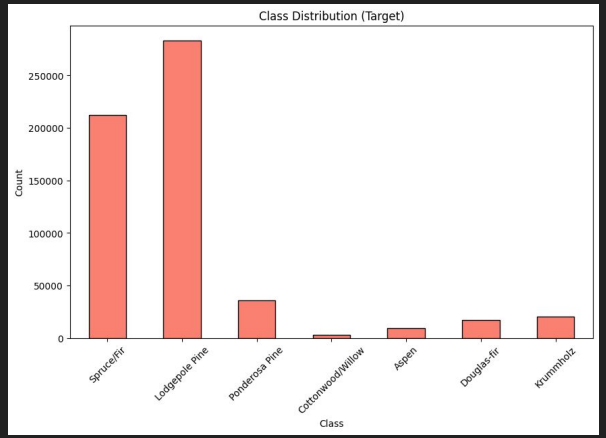
Features:

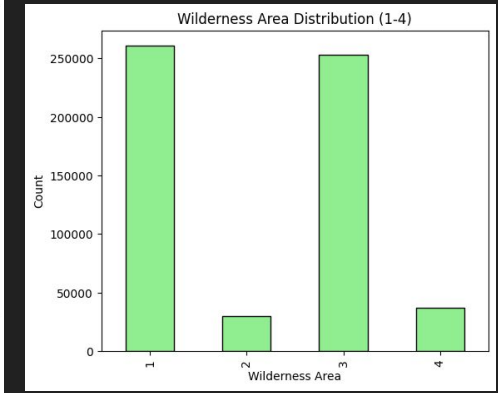
- Continuous: Elevation, Slope, Horizontal/Vertical Distance to Hydrology, Hillshade, etc.
- Categorical/Binary: Soil type (40 categories), Wilderness area (4 categories)

Data Distribution:

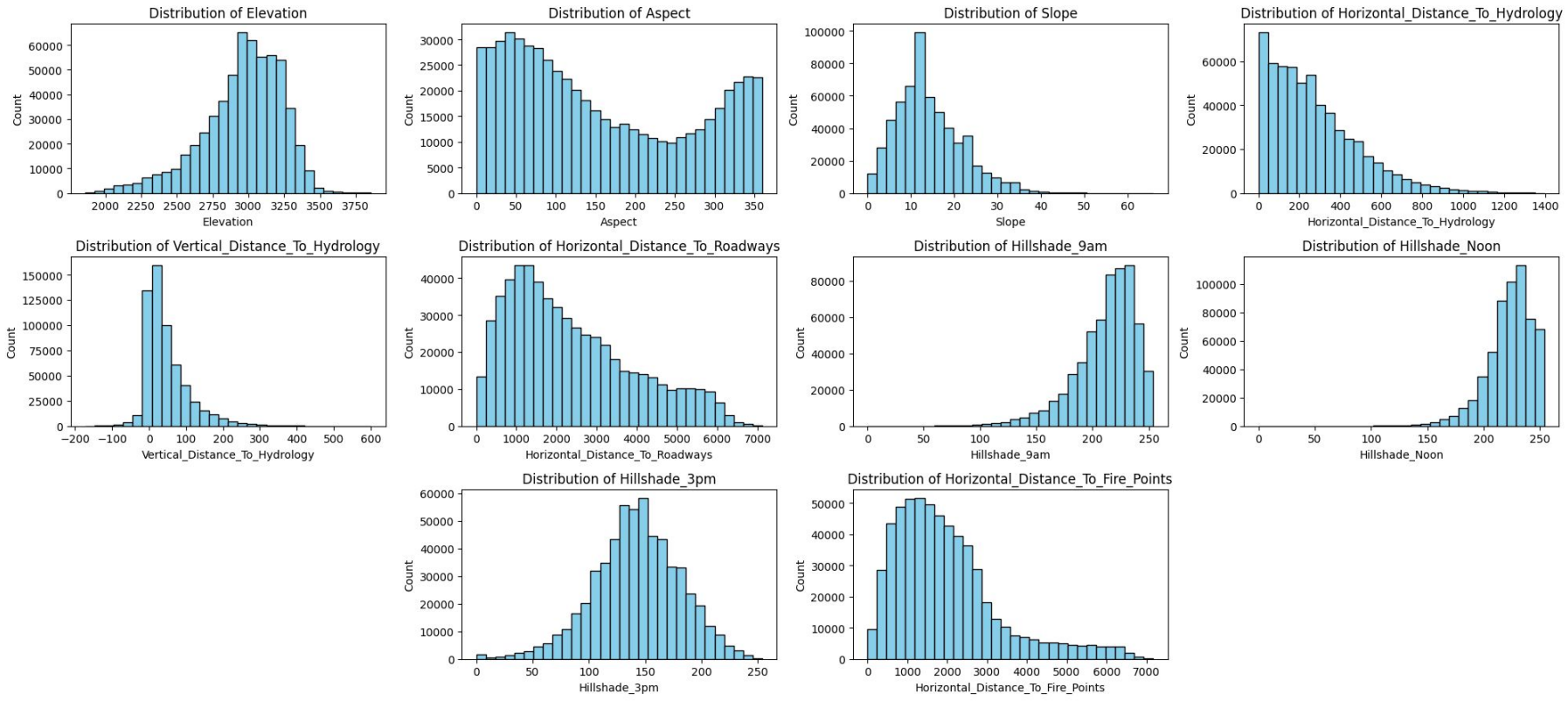
- Imbalanced classes (some forest types less frequent)







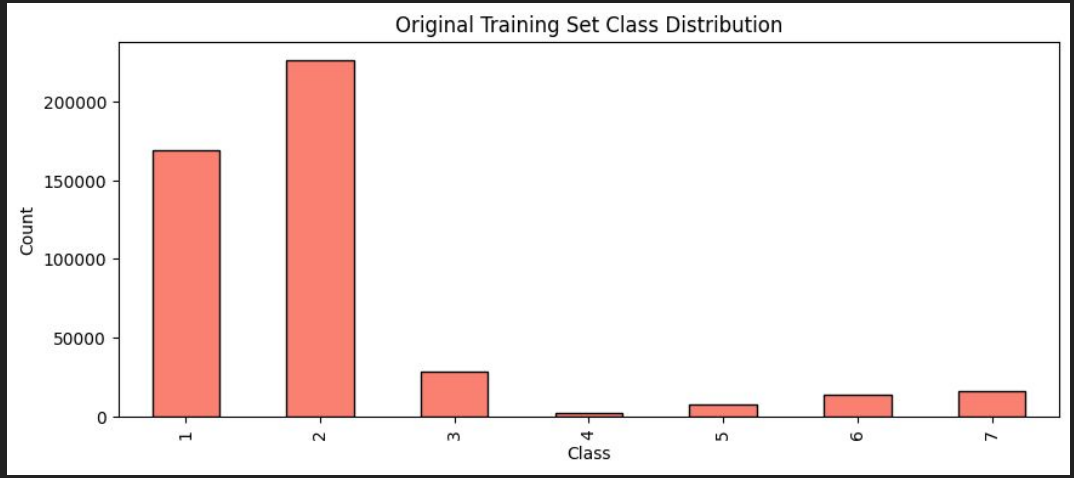
Dataset


Distribution of Elevation
Count
Elevation
Distribution of Aspect
Count
Aspect
Distribution of Slope
Count
Slope
Distribution of Horizontal_Distance_To_Hydrology
Count
Horizontal_Distance_To_Hydrology
Distribution of Vertical_Distance_To_Hydrology
Count
Vertical_Distance_To_Hydrology
Distribution of Horizontal_Distance_To_Roadways
Count
Horizontal_Distance_To_Roadways
Distribution of Hillshade_9am
Count
Hillshade_9am
Distribution of Hillshade_Noon
Count
Hillshade_Noon
Distribution of Hillshade_3pm
Count
Hillshade_3pm
Distribution of Horizontal_Distance_To_Fire_Points
Count
Horizontal_Distance_To_Fire_Points

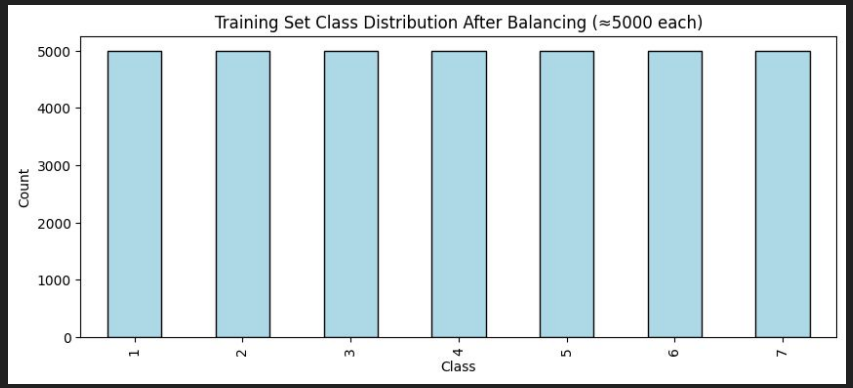
Data Preprocessing

Steps taken:

- Handling missing values
- Encoding Categorical Values
- Train/Test Split
- Addressing Class Imbalances
- Feature Engineering / Selection
- Data Scaling

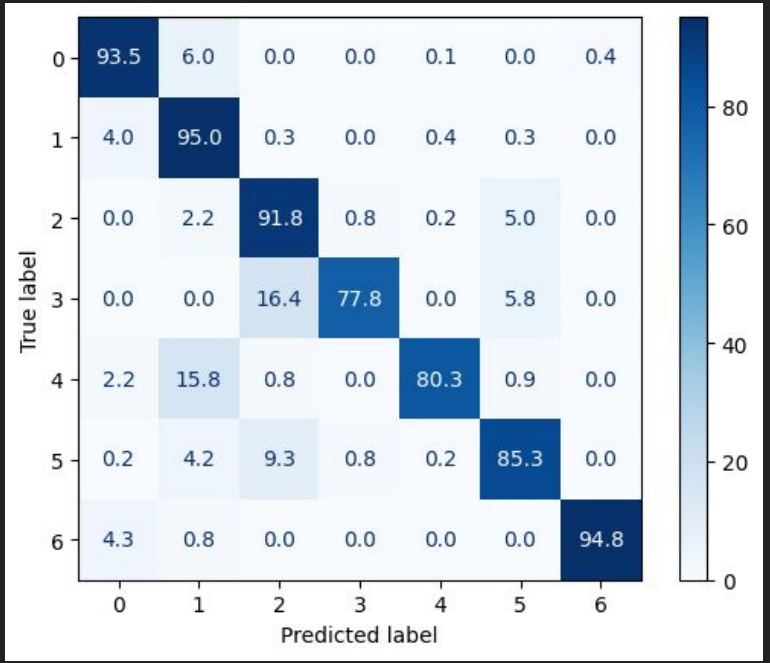


Addressing Class Imbalances



Model Evaluations

- Accuracy
  - Initial Assessment
  - Misleading
- Matthews Correlation Coefficient (MCC)
  - Balanced Evaluation
- Classification Reports
  - Individual Classification
- Confusion Matrices
  - Visualizations



Model #1: K-Nearest Neighbors Algorithm (KNN)

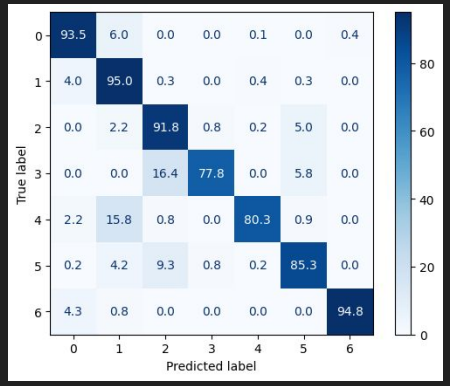
Reason for Choice:

- Non-parametric, instance-based learning
- Captures local patterns in feature space

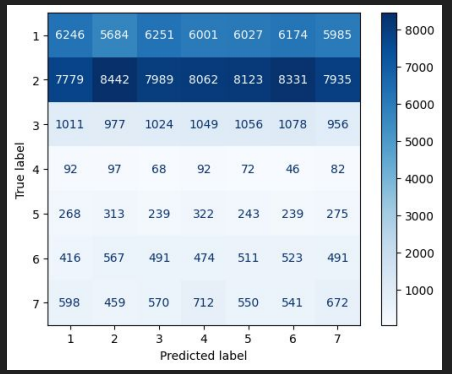
Hyperparameters Tuned:

- k (number of neighbors) - [3, 5, 7, 11, 15]
  - Overfitting/underfitting
- Distance metric (euclidean, minkowski)
- Weighting function (uniform, distance)
- SMOTE / Sampling / No Sampling

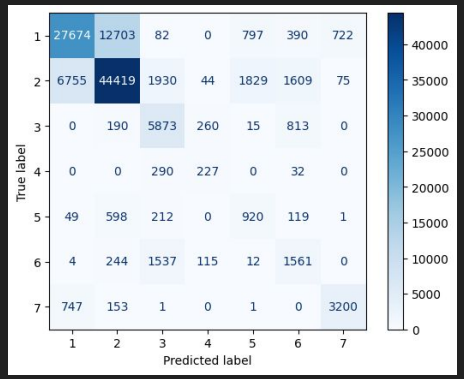
No Sampling



Undersampling (2198 per class)



Oversampling with SMOTE



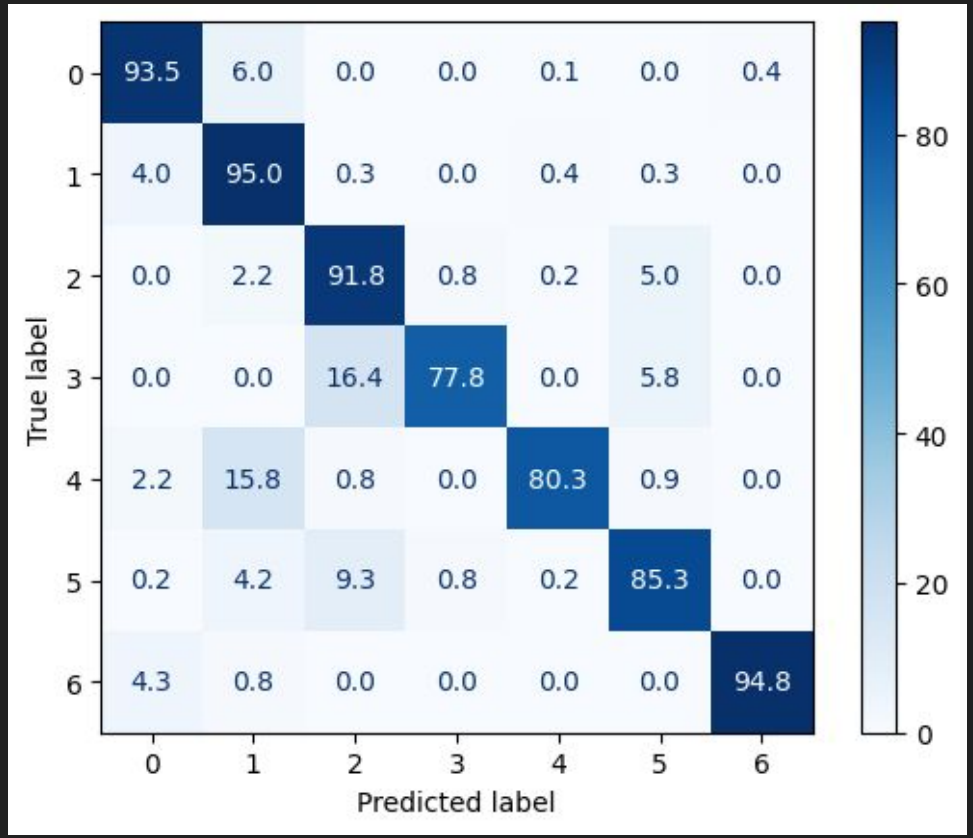
Model #1: K-Nearest Neighbors Algorithm (KNN)

Final Model Parameters:

- k = 3
- Distance metric = Minkowski
- Weighting function = Distance
- No Sampling

Performance:

- MCC Score on training: 86.66%
- MCC Score on test set: 89.80%
- Classification Report (F1-scores):
  - Class 1: 0.94
  - Class 2: 0.95
  - Class 3: 0.92
  - Class 4: 0.81
  - Class 5: 0.82
  - Class 6: 0.85
  - Class 7: 0.95



Model #2: Decision Tree

Reason for Choice:

- Interpretable and easy to visualize
- Handles both numerical and categorical features well
- Captures non-linear decision boundaries

Hyperparameters Tuned:

- max_depth - [None, 5, 10, 20, 30, 35, 40, 45]
- criterion - (gini, entropy)

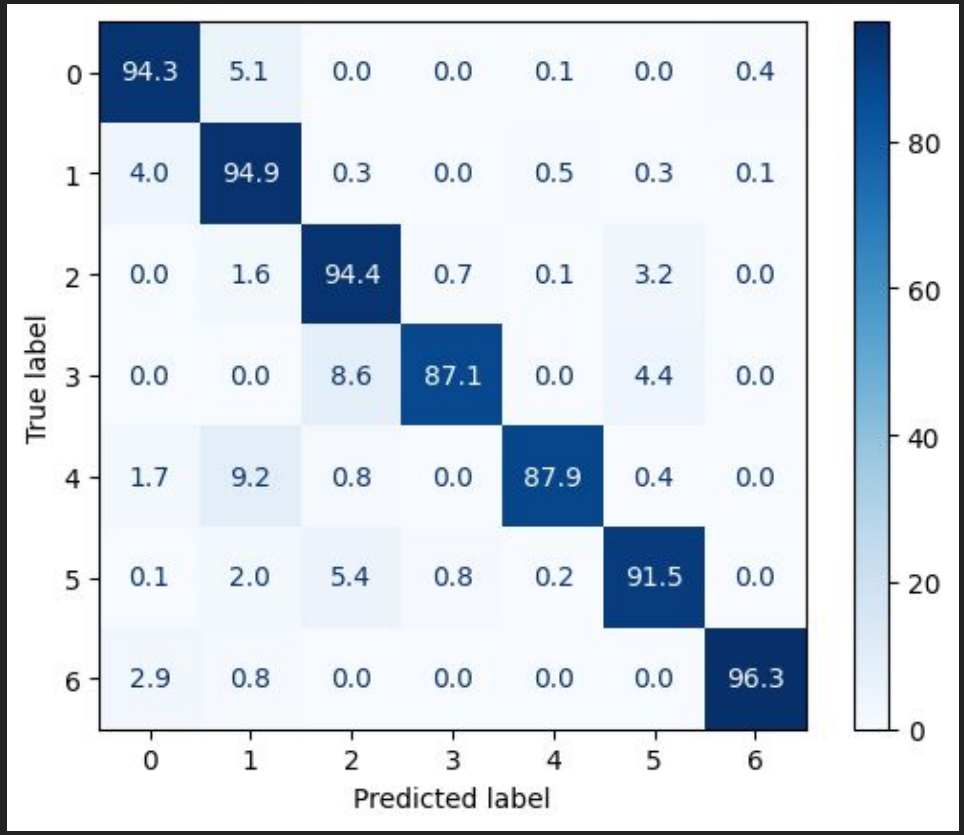
Final Model Parameters:

- max_depth = 40
- criterion = entropy

Model #2: Decision Tree

Performance:

- Accuracy on training set: 97.90%
- Accuracy on test set: 94.46%
- Classification Report (F1-scores):
  - Class 1: 0.94
  - Class 2: 0.95
  - Class 3: 0.94
  - Class 4: 0.87
  - Class 5: 0.86
  - Class 6: 0.90
  - Class 7: 0.95



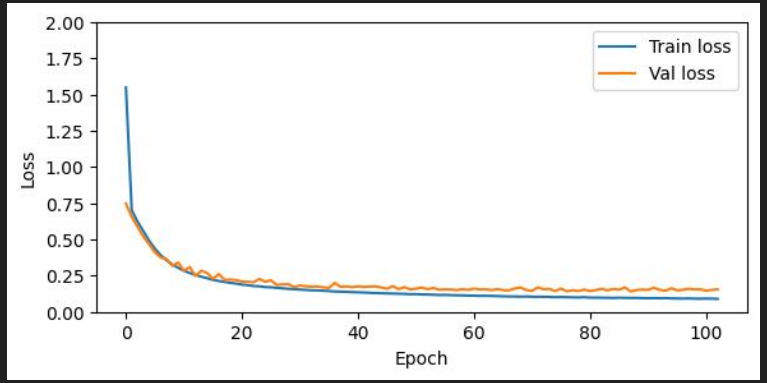
Model #3: Neural Network

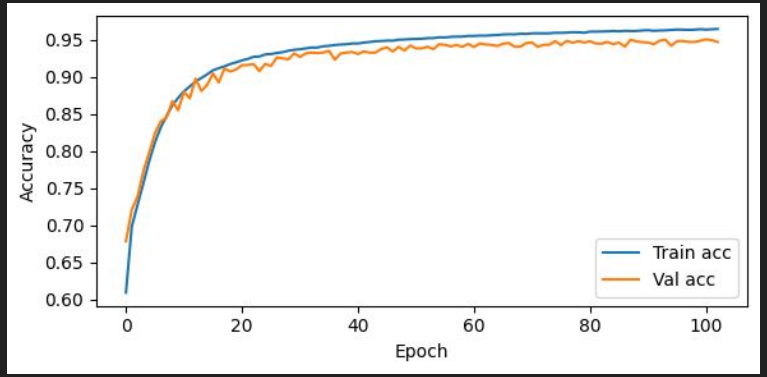
Reason for Choice:

- Extremely powerful training
- Can discover new nonlinear patterns

Model Description:

- 4 Hidden Layers (512)
  - ReLU activation
- Softmax for output layer
- Adam Optimizer: 0.00025 learning rate
- Early Stopping: 15 patience
- 400 epochs, batch size of 256
- Imbalanced data used

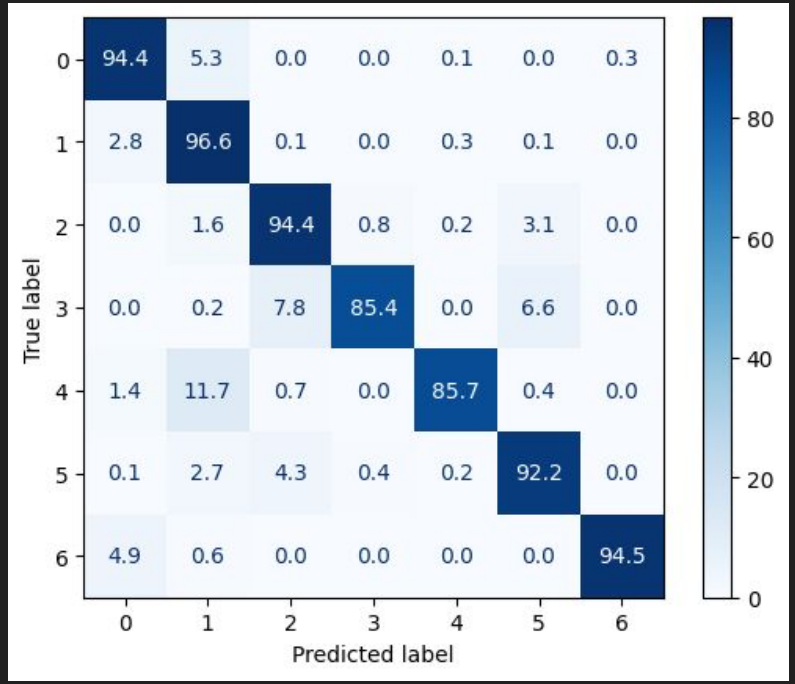




Model #3: Neural Network

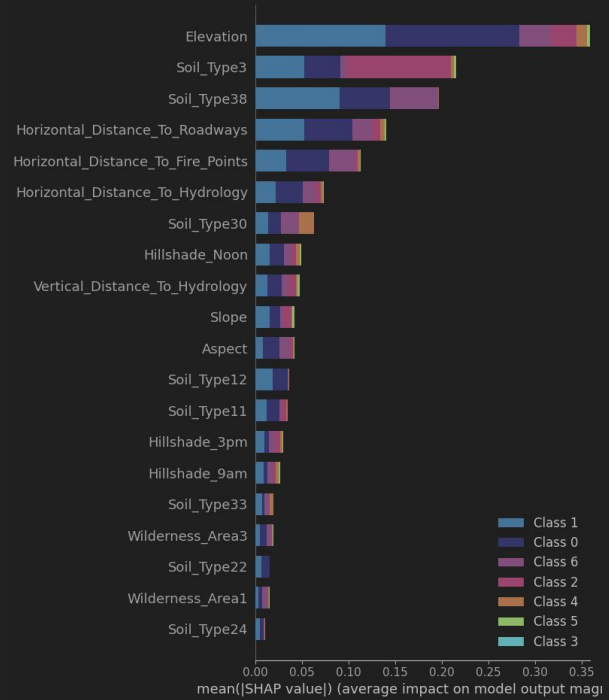
Performance:

- Test MCC: 92%
- Classification Report (F1-scores):
  - Class 1: 0.95
  - Class 2: 0.96
  - Class 3: 0.95
  - Class 4: 0.86
  - Class 5: 0.87
  - Class 6: 0.91
  - Class 7: 0.95

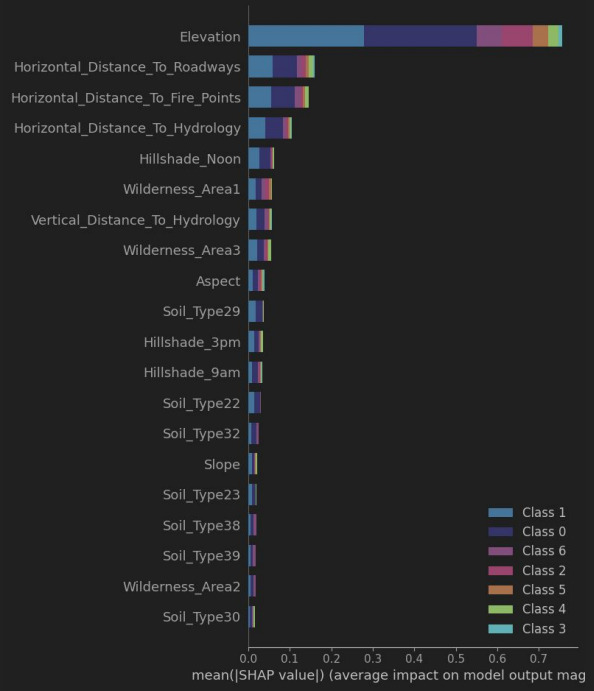


Interpretability - SHAP (SHapley Additive exPlanations)

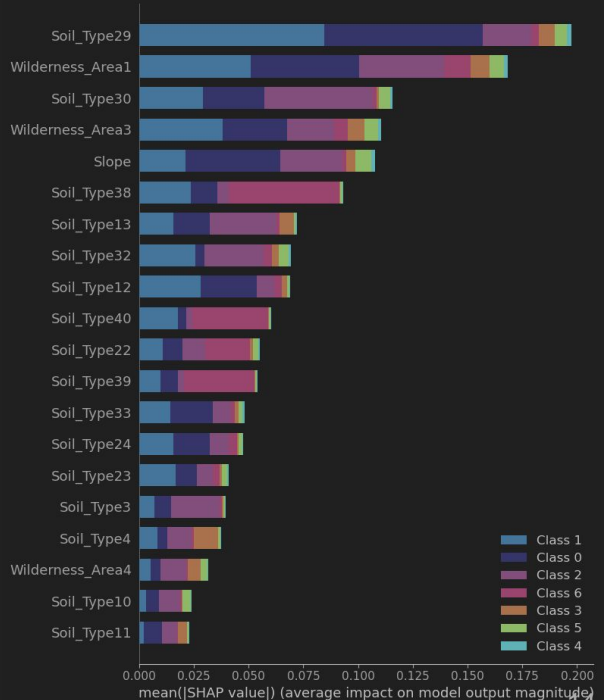
KNN



Decision Tree



Neural Network



Model Comparison

- The best model was our Neural Network
  - Highest individual F1 scores
  - Highest accuracy

KNN

Classification Report:

	precision	recall	f1-score	support
1	0.94	0.94	0.94	42368
2	0.94	0.95	0.95	56661
3	0.92	0.92	0.92	7151
4	0.84	0.78	0.81	549
5	0.85	0.80	0.82	1899
6	0.84	0.85	0.85	3473
7	0.95	0.95	0.95	4102
accuracy			0.94	116203
macro avg	0.90	0.88	0.89	116203
weighted avg	0.94	0.94	0.94	116203

Decision Tree

Classification Report:

	precision	recall	f1-score	support
1	0.94	0.94	0.94	42368
2	0.95	0.95	0.95	56661
3	0.94	0.94	0.94	7151
4	0.86	0.87	0.87	549
5	0.84	0.88	0.86	1899
6	0.88	0.92	0.90	3473
7	0.95	0.96	0.95	4102
accuracy			0.94	116203
macro avg	0.91	0.92	0.92	116203
weighted avg	0.94	0.94	0.94	116203

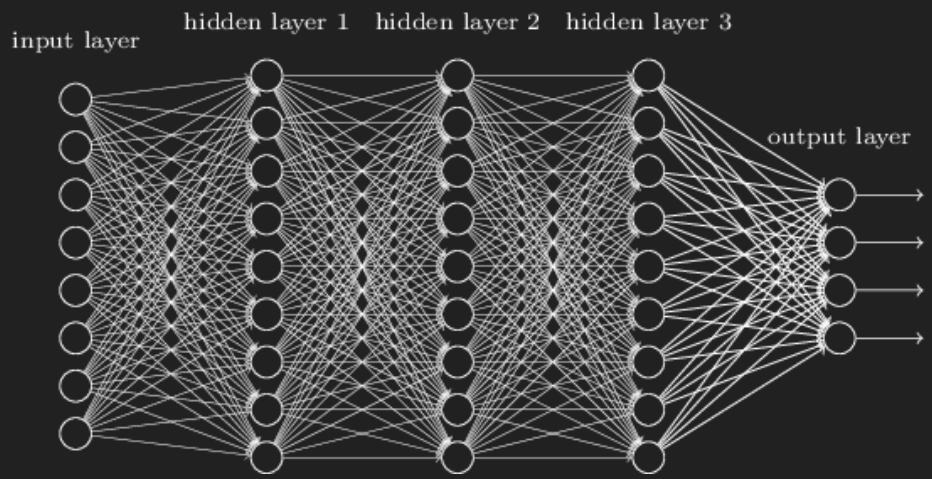
Neural Network

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.94	0.95	42368
1	0.95	0.97	0.96	56661
2	0.96	0.94	0.95	7151
3	0.87	0.85	0.86	549
4	0.88	0.86	0.87	1899
5	0.91	0.92	0.91	3473
6	0.96	0.94	0.95	4102
accuracy			0.95	116203
macro avg	0.93	0.92	0.92	116203
weighted avg	0.95	0.95	0.95	116203

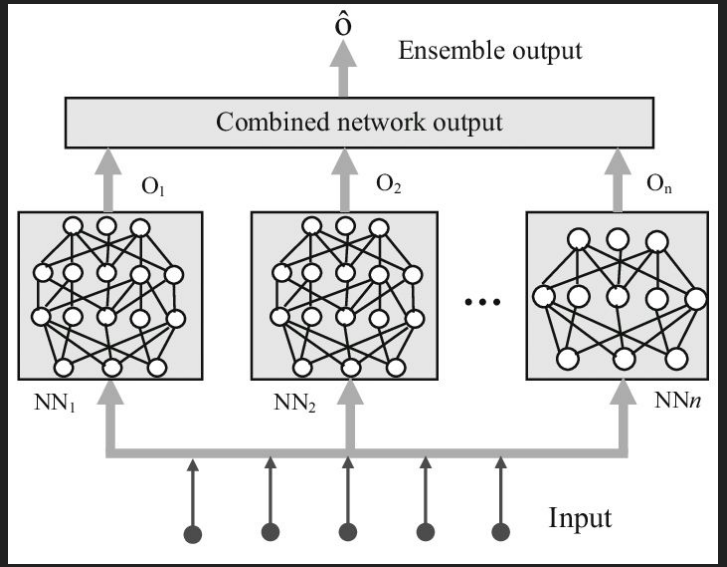
Limitations

- K-Nearest Neighbors (KNN):
  - Sensitive to noise and data quality
- Decision Tree:
  - Prone to overfitting
  - Limited generalization
- Neural Network:
  - Hard to choose correct parameters
  - Expensive training process
  - Automation difficulties



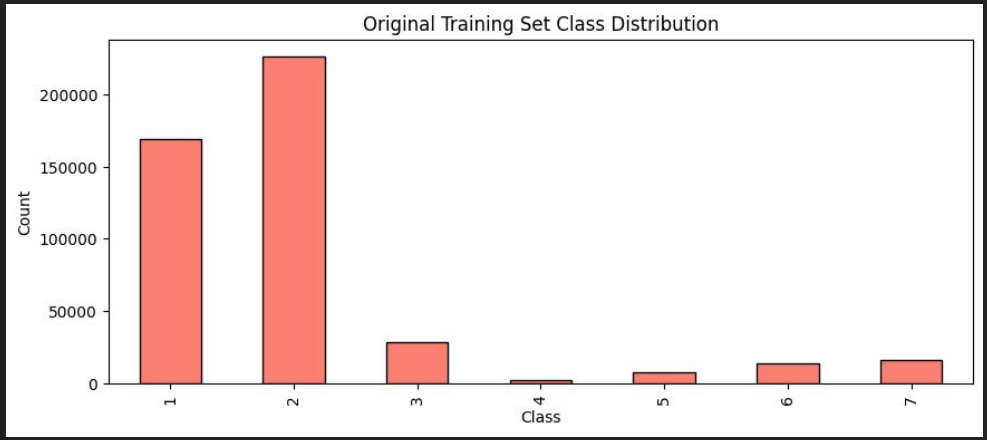
Possible Improvements

- Advanced Models
  - SVMs, ensemble neural networks
- New Techniques
  - Model stacking/blending
- Background Knowledge
  - Understand dataset better



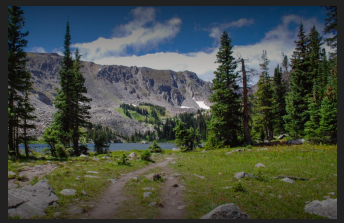
Model Conclusions

- Best Predictors:
  - Elevation
  - Distances to key locations
  - Differences in Neural Network
- Limits of Imbalanced Data
  - Class Weights
  - SMOTE
  - KNN



Acknowledgments

Github: https://github.com/ahm5348/CSCI-635-01-Group-3/tree/main



Dataset link: https://archive.ics.uci.edu/dataset/31/covertype

Questions?

Thank You!!!