Forest Cover Type Prediction

Final Presentation of 2023 Spring Machine Learning

HaeLee Kim, Nayaeun Kwon, Upmanyu Singh

Introduction

- The study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado.
- Each observation is a 30m x 30m patch.
- Our project will predict an integer classification for the forest cover type (7 types).

    1 - Spruce/Fir
    2 - Lodgepole Pine
    3 - Ponderosa Pine
    4 - Cottonwood/Willow
    5 - Aspen
    6 - Douglas-fir
    7 - Krummholz

Forest Cover Type 1 - Spruce/Fir

Dataset Description

Number of Instances : 15120 observations

Number of Features : 13 features

Types of Features : integer

Data Source URL: Forest Cover Type Prediction | Kaggle

Forest Cover Type 2 - Lodgepole Pine

The Implemented Features

Elevation - Elevation in meters

Aspect - Aspect in degrees azimuth

Slope - Slope in degrees

Horizontal_Distance_To_Hydrology - Horz Dist to nearest surface water features

Vertical_Distance_To_Hydrology - Vert Dist to nearest surface water features

Horizontal_Distance_To_Roadways - Horz Dist to nearest roadway

Hillshade_9am (0 to 255 index) - Hillshade index at 9am, summer solstice

Hillshade_Noon (0 to 255 index) - Hillshade index at noon, summer solstice

Hillshade_3pm (0 to 255 index) - Hillshade index at 3pm, summer solstice

Horizontal_Distance_To_Fire_Points - Horz Dist to nearest wildfire ignition points

Wilderness_Area (4 binary columns, 0 = absence or 1 = presence) - Wilderness area designation

The wilderness areas are:

1 - Rawah Wilderness Area
2 - Neota Wilderness Area
3 - Comanche Peak Wilderness Area
4 - Cache la Poudre Wilderness Area

Soil_Type (40 binary columns, 0 = absence or 1 = presence) - Soil Type designation

The soil types are, for example:

1 - Cathedral family - Rock outcrop complex, extremely stony.
40 - Moran family - Cryorthents - Rock land complex, extremely stony.

Cover_Type (7 types, integers 1 to 7) - Forest Cover Type designation

Expected Outcomes

Goal

Predict an integer classification for the forest cover type

Target

Cover_Type (7 types) - Forest Cover Type designation

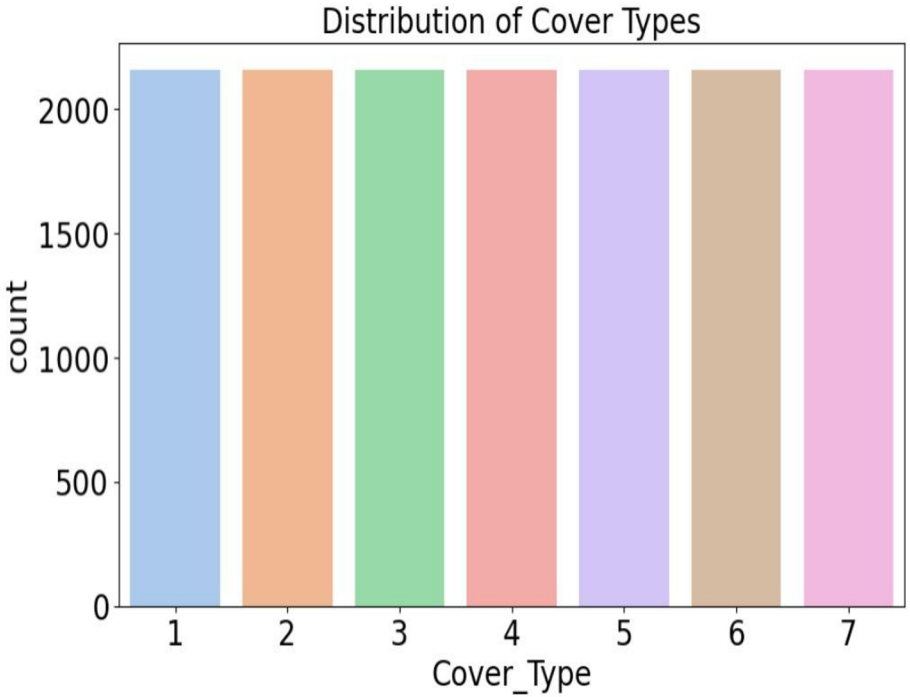
ML Models

Random Forest, KNN, MLP, Decision Tree

Data Preprocessing

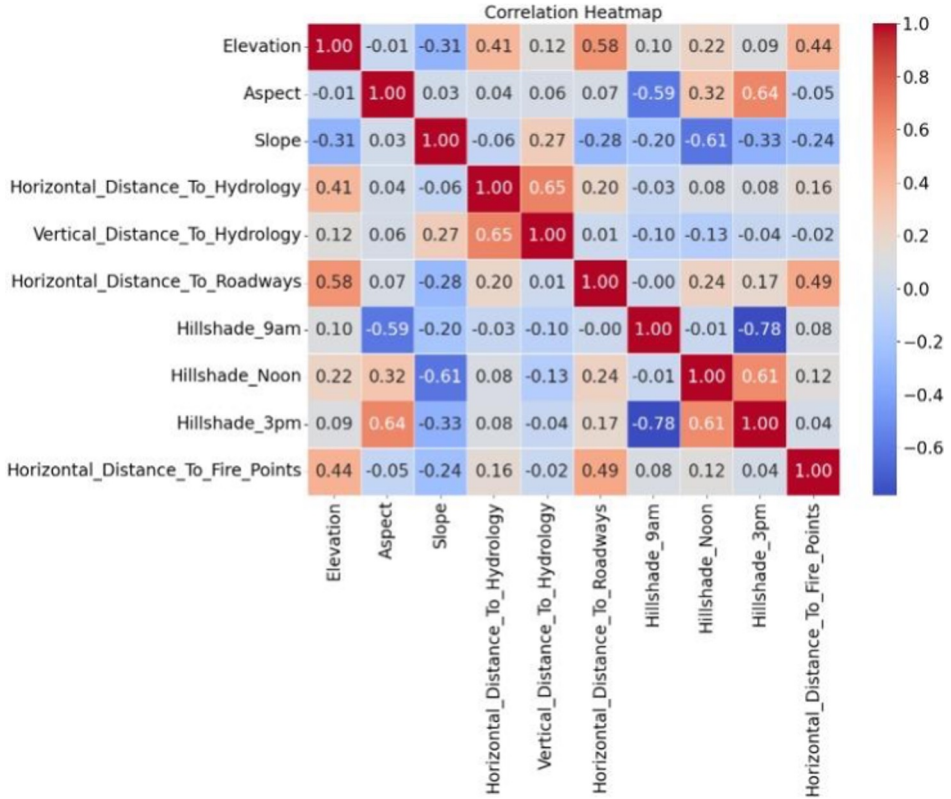
- Split the data - 20% test, 80% train
- No uncommon features
- Handed missing variables - No nulls
- Removed outliers
- Encoded the data - all the categorical variables are already label encoded.
- Scaled the data - standard scaler (from sklearn.preprocessing import StandardScaler)

EDA



We have an equal number of data points for all forest cover types in our dataset, which means that there is no imbalance present.

EDA

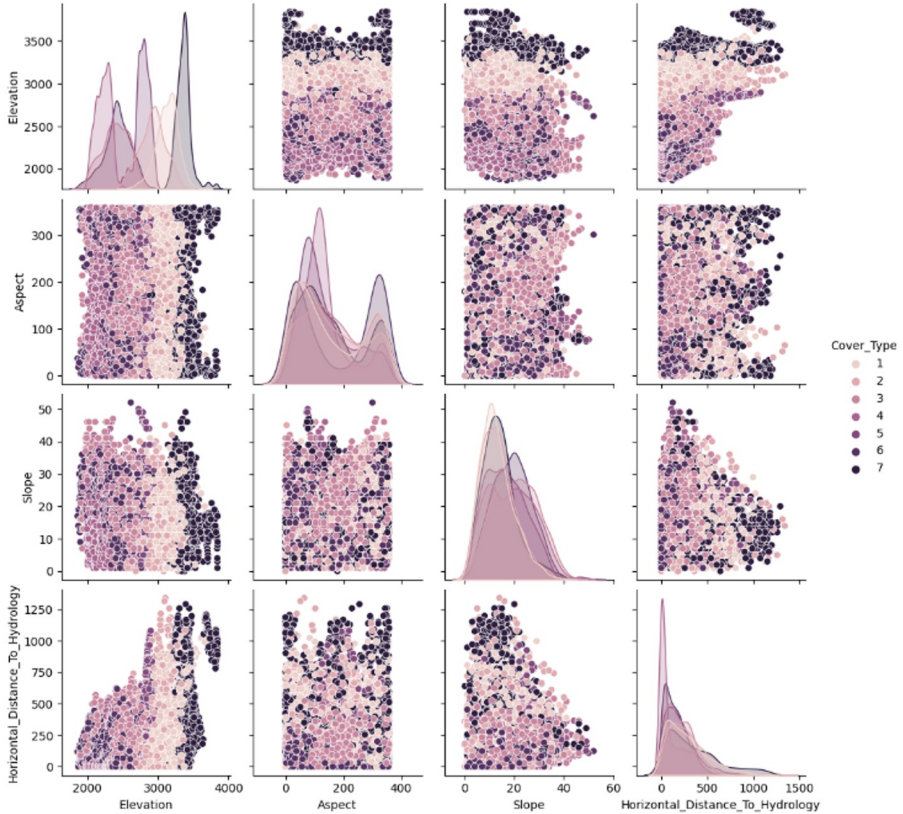


1. We created a correlation heatmap to visualize the relationship between all the continuous variables in the dataset.

1. One key insight we gained from the heatmap was that the elevation column has a significant correlation with all the features related to horizontal distance.

1. Another important finding was that Hillshade 9am and Hillshade 3pm have a high negative correlation. This makes sense, as the hillshade at 9am and 3pm will be opposite of each other.

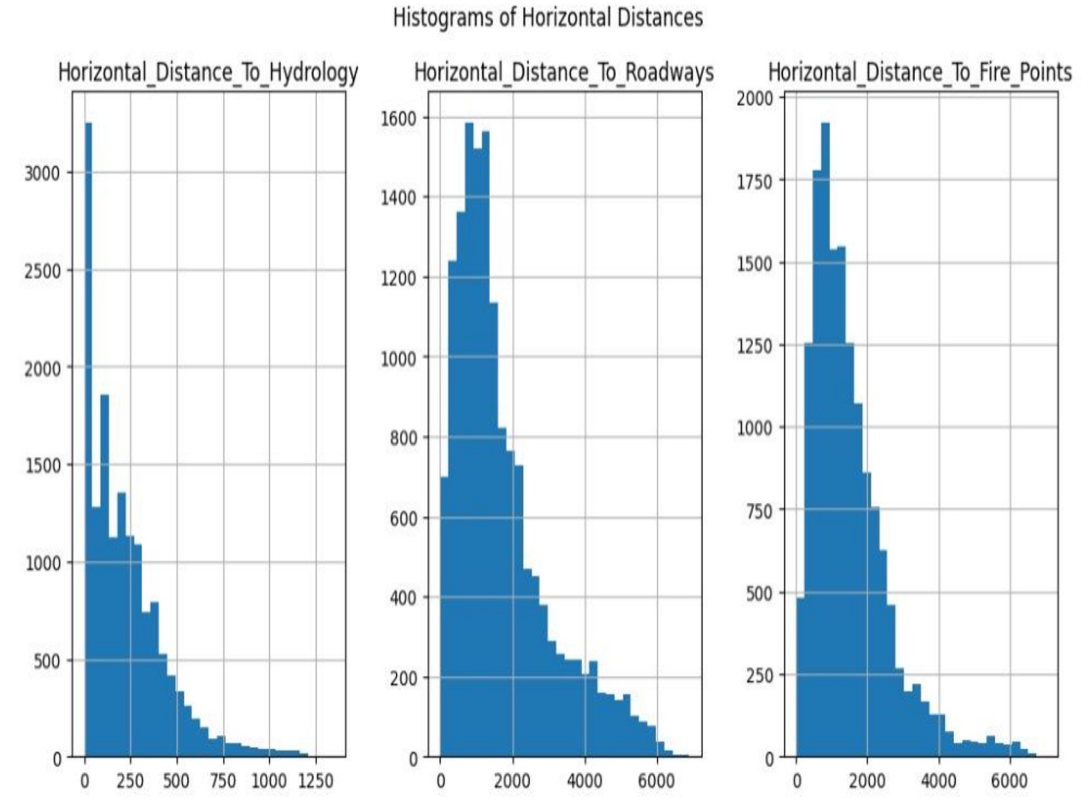
EDA



1. A pairplot to visualize the relationships between different continuous variables such as elevation, aspect, slope, and various distance measures (e.g. used horizontal distance to hydrology).

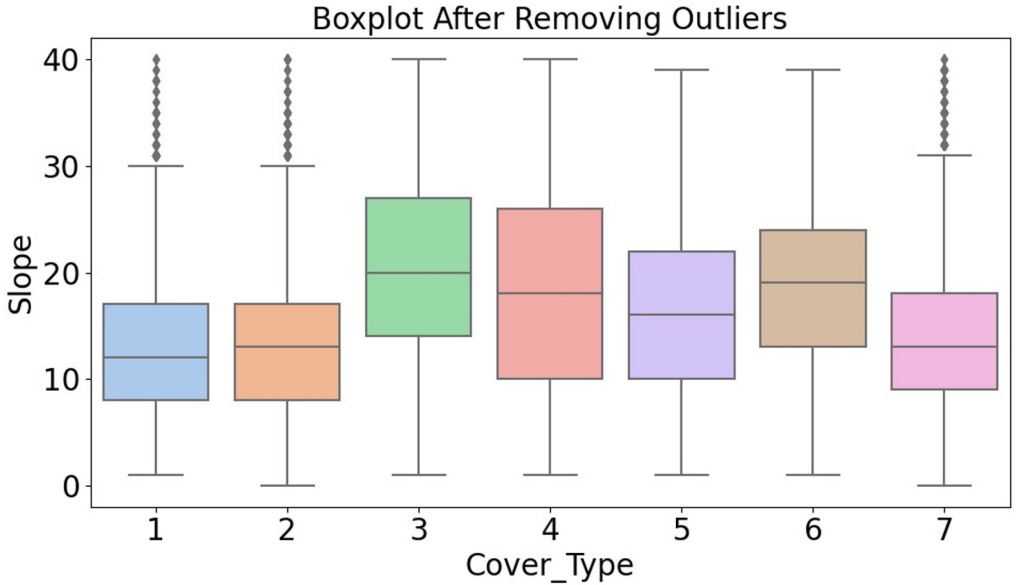
1. Higher elevations were more likely to have forest cover type 7. This was evident in plots comparing elevation to other features, where the highest elevation values consistently corresponded to type 7 forest cover.

EDA



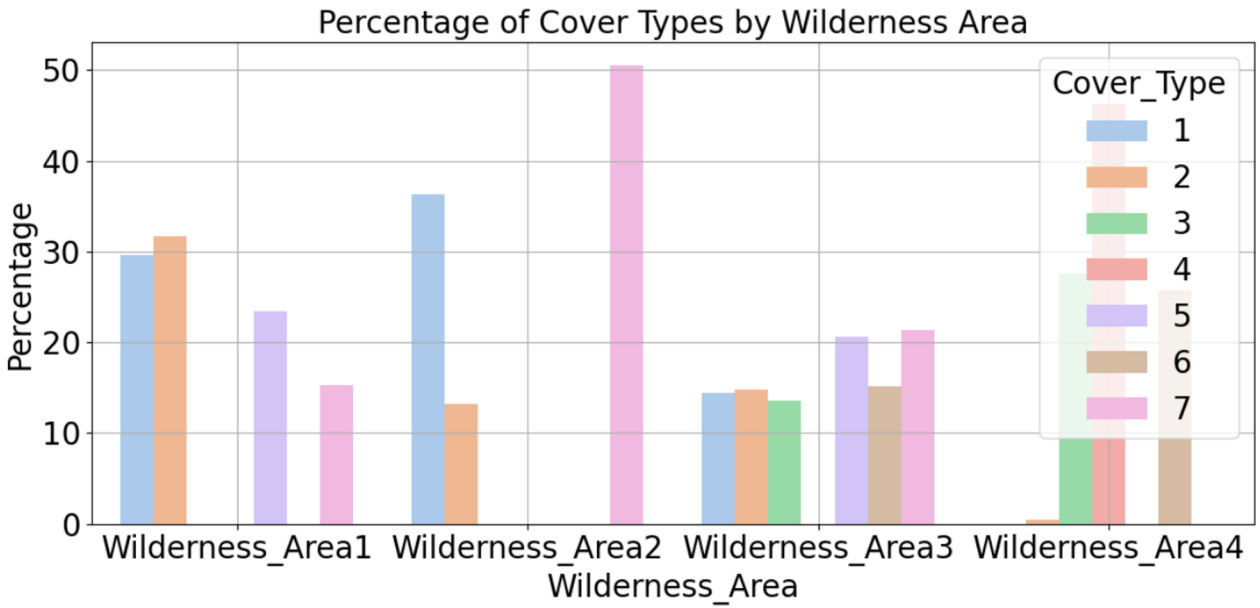
- The distribution of horizontal distance measures (to hydrology, roadways, and fire points) are all right-skewed, indicating that most of the forest cover types are located relatively close to these features.
- This suggests that proximity to water and accessibility via roadways may be important factors in determining forest cover type.

EDA



1. This box plot shows the distribution of slope among different forest cover types after removing outliers.
2. Forest cover types 3 and 4 have a higher median slope compared to other cover types.
3. Forest cover types 1, 2, and 7 have the lowest median slope, although there are a significant number of outliers among them.

EDA



- Percentage of cover types by wilderness area was analyzed.
- Wilderness area 2 only had forest cover types 1, 2 and 7.
- Wilderness area 4 had the highest percentage of forest cover type 4.

ML Analysis - Best Model: Random Forest Model

Classification Report:

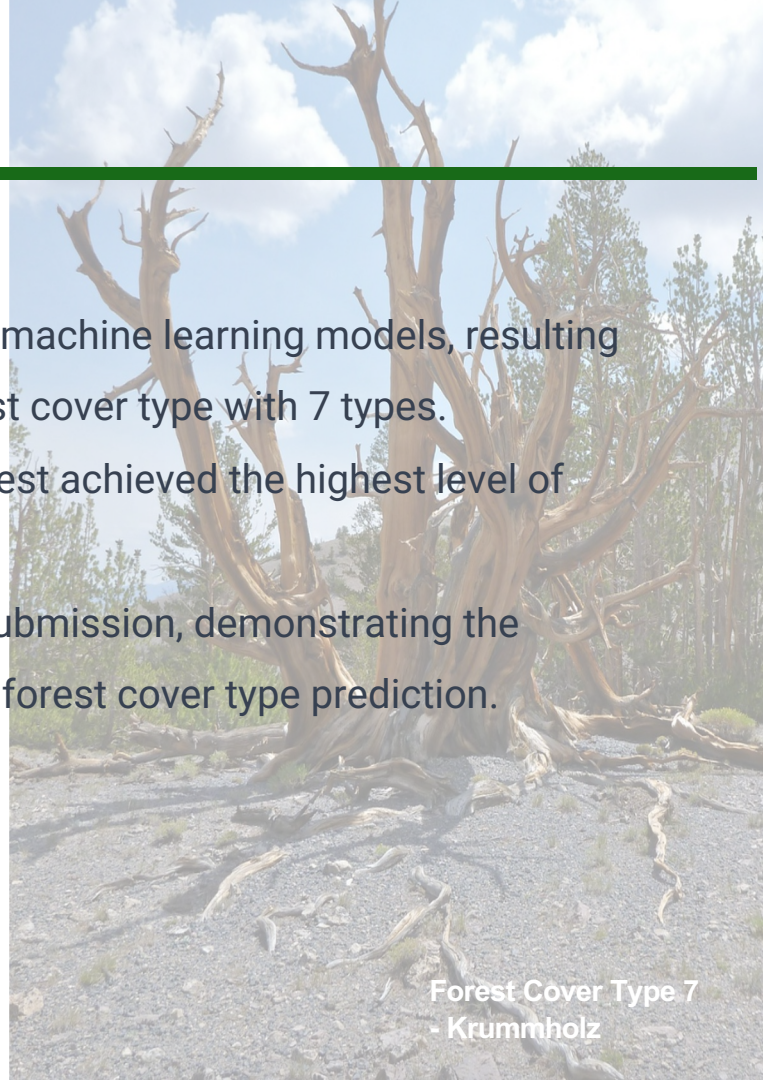
	precision	recall	f1-score	support
1	0.75	0.78	0.76	425
2	0.79	0.64	0.71	439
3	0.84	0.84	0.84	418
4	0.94	0.98	0.96	435
5	0.86	0.94	0.90	378
6	0.86	0.87	0.87	468
7	0.95	0.96	0.96	450
accuracy			0.86	3013
macro avg	0.86	0.86	0.86	3013
weighted avg	0.86	0.86	0.86	3013

```
Best model: Pipeline(steps=[('scaler', StandardScaler()),
                            ('model', RandomForestClassifier())])
Best F1 macro score: 0.8501251308768417
Test F1 macro score: 0.8565342004698275
```

- We evaluated multiple machine learning models, including Random Forest, KNN, MLP, and Decision Tree.
- We found that Random Forest was the best performer with an F1 score score of 0.86.
- Therefore, we chose it as our final model for predicting forest cover types.

Conclusion

- Our project aimed to predict the forest cover type using machine learning models, resulting in the prediction of an integer classification for the forest cover type with 7 types.
- Among the machine learning models used, Random Forest achieved the highest level of accuracy in predicting the forest cover type.
- Our project achieved a score of 0.73579 in the Kaggle submission, demonstrating the effectiveness of our approach in solving the problem of forest cover type prediction.




Thank you!