

# Forest Cover Type Classification

Presented by:  
Shahad Aati  
Abdullah Alshami



# TABLE OF CONTENT

**1**

**Introduction**

**2**

**Problem  
Statement**

**3**

**Objective**

**4**

**Data Set**

**5**

**Target  
Imbalance**

**6**

**Tools**

**7**

**Project  
Workflow**

**8**

**Correlations  
with Target**

**9**

**Validation  
Scores**

**10**

**Conclusion**



# INTRODUCTION

- The United States of America has many diverse biomes (deserts, forests, mountains...etc.).
- Forests have many trees that form a cover over the forest. The types of these covers can be determined from certain factors; such as **elevation** and **soil types**.



# PROBLEM STATEMENT

What are the factors that determine the types of forest covers?

# OBJECTIVE

Predict forest cover types based on certain factors



# DATA SET



**Data Source:**  
[archive.ics.uci.edu](http://archive.ics.uci.edu)



**Before Cleaning:**

581012 rows

55 columns



**After Cleaning:**

536431 rows

53 columns

# EXPLANING DATA

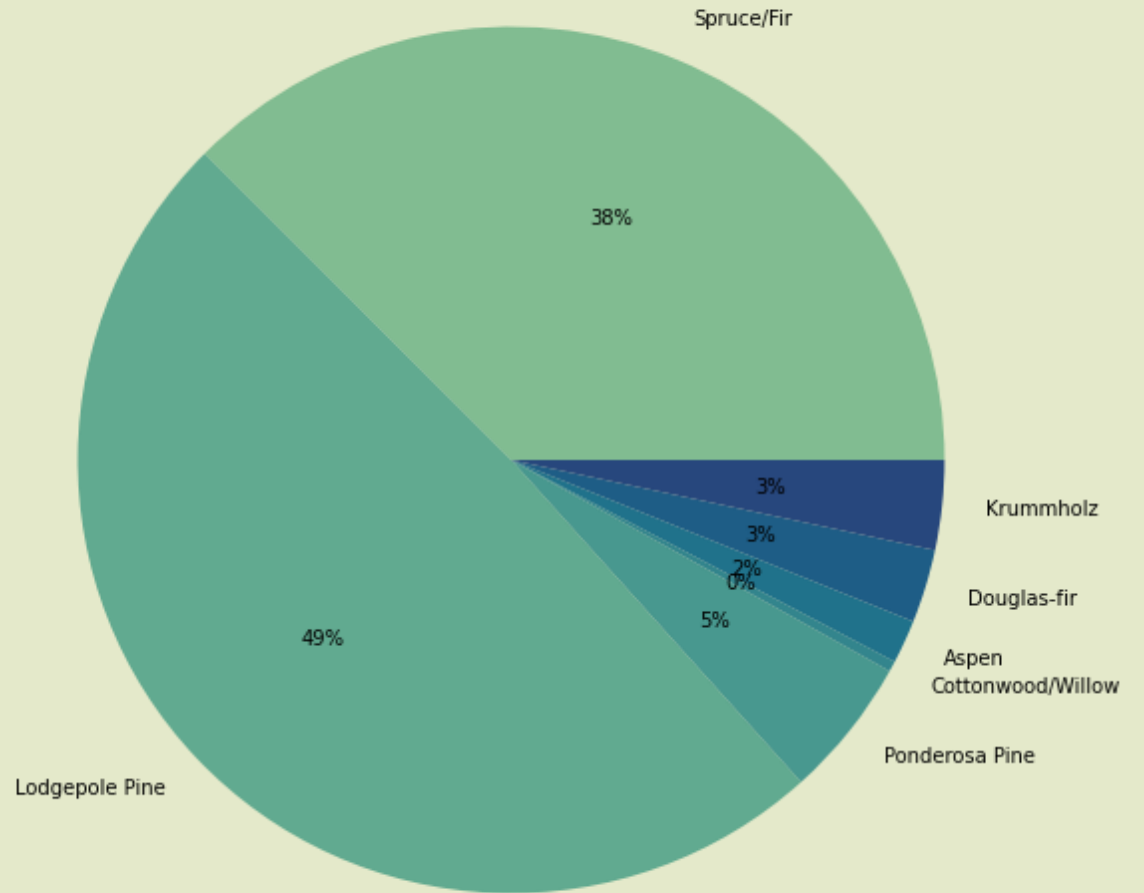
- The dataset contains **54** features and **1** target variable.
  - **10** numeric features.
  - **44** categorical (dummies) features.
- The target has **7** classes represented as numbers ranging from **1** to **7**.
- Each number represents a type of forest cover.

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



# TARGET IMBALANCE

- To reduce the imbalance in the data, all the classes were grouped together except for **Spruce/Fir** and **Lodgepole Pine** which make up 87% of the data.



# TOOLS





# PROJECT WORKFLOW

Data Cleaning

Feature  
Engineering

Training-Validation

Hyper Parameter  
Tuning

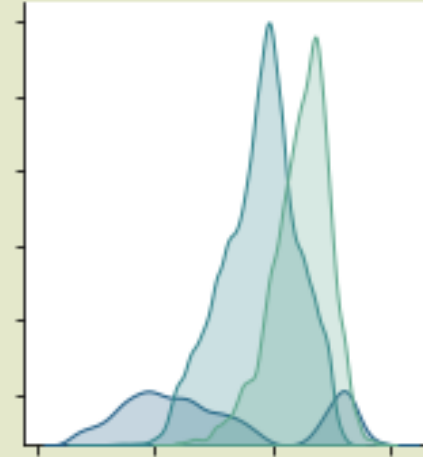
Testing Model



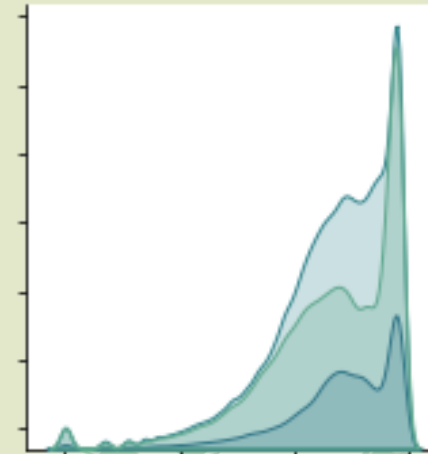
# FEATURE ENGINEERING

- Tried many feature engineering techniques but most did not affect the model scores, or affected them negatively.
- Only 2 columns were modified, **Elevation** and **Aspect**.
- Log value was taken for each column which helped with the skewness and distribution of the data.

Elevation



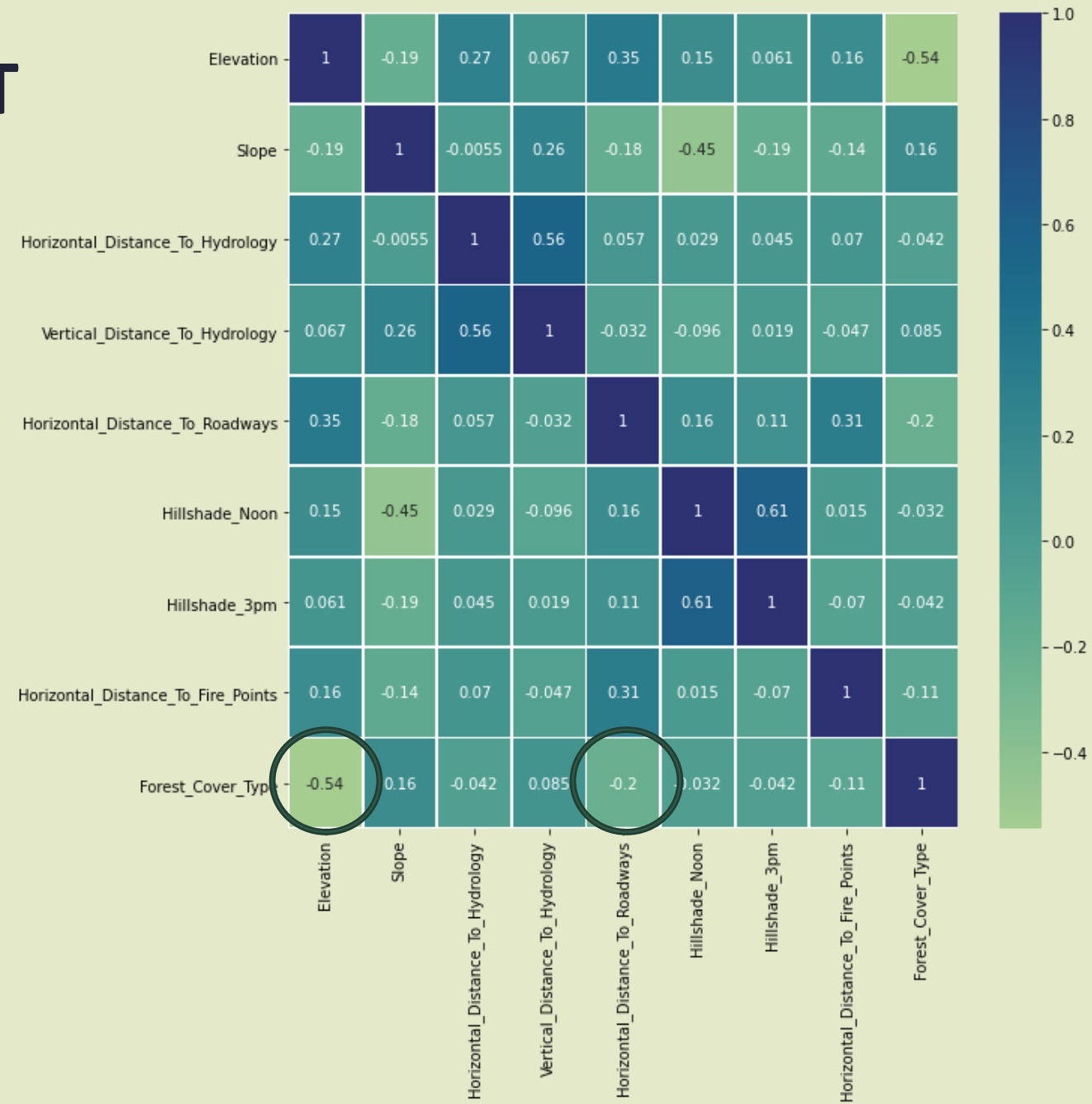
Aspect



# CORRELATIONS WITH TARGET

Highest correlated Features with target:

- The **Elevation** feature has high negative correlation (**-0.54**) with the target **Forest\_Cover\_Type**
- The **Horizontal\_Distance\_To\_Roadways** feature has high negative correlation (**-0.20**) with the target **Forest\_Cover\_Type**



# MODEL TRAINING & VALIDATION SCORES (OneVsRest)

Model	Train		Validation	
	Accuracy	F1 Score	Accuracy	F1 Score
K-Nearest Neighbor (Neighbors=5)	89.57%	89.57%	81.06%	81.09%
Logistic Regression (C=100)	66.32%	65.74%	66.24%	65.70%
Decision Trees (Depth=4)	71.38%	71.02%	70.76%	70.42%
Random Forest (Trees=100)	100%	100%	96.62%	96.62%
Extra Trees Classifier (Trees=100)	100%	100%	96.68%	96.68%
Gaussian Naive Bayes	58.80%	56.08%	58.48%	55.58%
Bernoulli Naive Bayes	64.86%	64.36%	64.60%	64.17%

# CONCLUSION

- As seen from the previous table, the best model in terms of performance is **Extra Trees**.
- Grid Search was used to tune and find the best hyper parameters for the chosen model.
- The test result after tuning was **97.85%**



# Thanks!

