Project : Final Report

Team: my (ana)conda don't
Members: Aditya Srivastava, Harsh Gupta, Rachna Konigari

Objective

Given topological and cartographic data from four different areas of the Roosevelt National Forest in Colorado, perform a predictive analysis of the type of forest cover that can be observed in an area.

Motivation

Global forest cover over the past 60 years has declined by 81.7 million hectares which has been correlated with many negative impacts on the environment. To preserve and restore forest cover and biodiversity without causing a detrimental effect on the ecology of a region, it is essential that we analyze the type of flora that is endemic to it, and ensure that we replicate the same in conservation efforts.

One of the ways we can do this is by using data to model the environmental patterns that are specific to certain kinds of plants, and use it to predict which plants would be suitable given the topological and cartographic features.

Dataset Description

The dataset presents records of the types of forest cover found in four different wilderness areas located in the Roosevelt National Forest of Northern Colorado. Each observation is from a 30x30m^2 patch.

The dataset contains **581012** samples and **55** different features. Some of the interesting features are elevation, distance from water sources, distance from roadways and soil type.

Features

There are 55 different features in the dataset;

Feature	Description
Elevation	Elevation in meters
Aspect	Aspect in degrees azimuth

Slope	Slope in degrees
Horizontal_Distance_To_Hydrology	Horizontal Distance to nearest surface water features
Vertical_Distance_To_Hydrology	Vertical Distance to nearest surface water features
Horizontal_Distance_To_Roadways	Horizontal Distance 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	Horizontal Distance to nearest wildfire ignition points
Wilderness_Area (4 binary columns)	Wilderness area designation (Rawah, Neota, Comanche Peak, Cache la Poudre)
Soil_Type (40 binary columns)	Soil Type (Cathedral, Vanet, Bulwark, etc)
Cover_Type (7 types)	Forest Cover Type designation (Spruce/Fir, Aspen, etc)

We can split the features into numerical and categorical features below;

Numerical Features	Categorical Features
Elevation	Soil Type
Aspect	Wilderness Area
Slope	Cover Type (Prediction Target)
Horizontal Distance to Hydrology	
Vertical Distance to Hydrology	
Horizontal Distance to Roadways	
Hillshade at 9 am	
Hillshade at Noon	
Hillshade at 3 pm	
Horizontal Distance to Fire Points	

We also observe that none of the data points have any duplicate values and none of them have an N/A value as well.

Sample Data

Five randomly selected rows are displayed below. The binary fields, *Wilderness_Area 1* through *4* and *Soil_Type 1* through *40* have been omitted.

	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways
520204	3241	103	16	285	45	108
326736	2860	245	20	443	28	899
148101	2901	133	16	192	46	3397
452647	3139	191	1	85	6	967
408927	2883	108	5	90	2	2143

Hillshade_9am	Hillshade_Noon	Hillshade_3pm	Horizontal_Distance_To_Fire_Points	Cover_Type
245	217	96	601	2
176	251	209	1739	5
244	230	109	2764	2
219	239	157	1910	1
229	234	138	2696	2

Data Analysis

Distributional Analysis

Figure 1. shows a scatter plot of the 2D **t-SNE embeddings** extracted from the dataset. T-Distributed Stochastic Neighbor Embeddings (t-SNE) is a nonlinear dimensionality reduction technique to visualize data in a two or three dimensional space. From the visualization we can see that the dataset is very homogenous with only clusters of cover type 4 and 3 showing any significant separation from the rest.

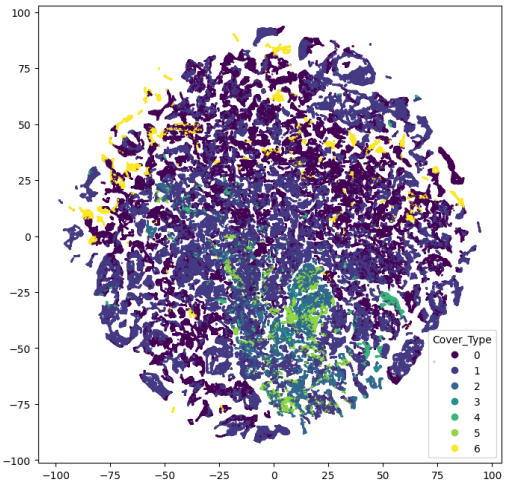


Fig 1: t-SNE plot

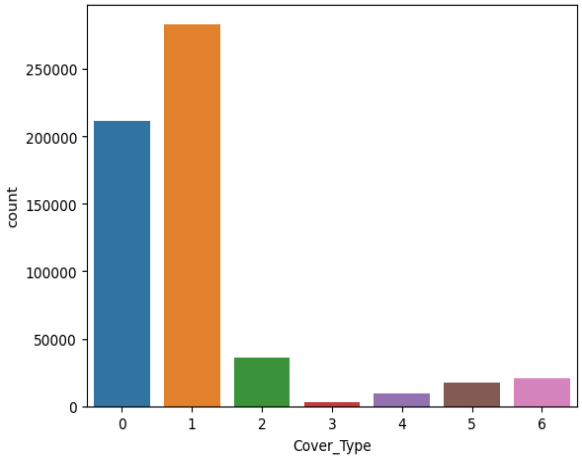


Fig 2: Data Distribution across different

cover types

The dataset demonstrates considerable skew; in the classwise sample distribution in figure 2. we can see that the dataset is biased rather heavily towards cover types 0 and 1. Thus, we opted to use stratified sampling when creating our train and test splits, so that we could maintain the same distribution of samples across our splits.

The features themselves also vary in symmetry and spread, necessitating the need for scaling, standardization and normalization. The skewed features before and after treatment are shown in the table below.

Skew	Before Treatment	After Treatment
High	HDHydrology VDHydrology Hillshade9am HillshadeNoon HDFirePoints	None
Moderate	Elevation Slope HDRoadways	HDHydrology VDHydrology HDRoadways HDFirePoints
Fair	Aspect Hillshade3pm	Elevation Aspect Slope Hillshade9am HillshadeNoon Hillshade3pm

Feature Correlation

We studied feature correlation using Pearson's R for the numerical values. The heatmap for the correlation scores is shown in figure 3. The Pearson correlation measures the strength of the linear relationship between two variables. There is a high correlation between features Aspect and Hillshade. Elevation also has a high correlation with the target variable Cover_Type, suggesting that certain species of plants are better adapted to survival at certain heights. The relationships can also be observed in the pair plots.

Some other inferences from our pair plots shown in figure 4:

- Hillshades have an ellipsoid pattern with each other
- Hillshades and Aspect attributes show a sigmoid relationship
- Wilderness_Area_1, Wilderness_Area_3 map to a very small subset of Cover_Types. They are highly correlated with Cover Type 0, where the absence of both the wilderness areas signals the presence of the latter.

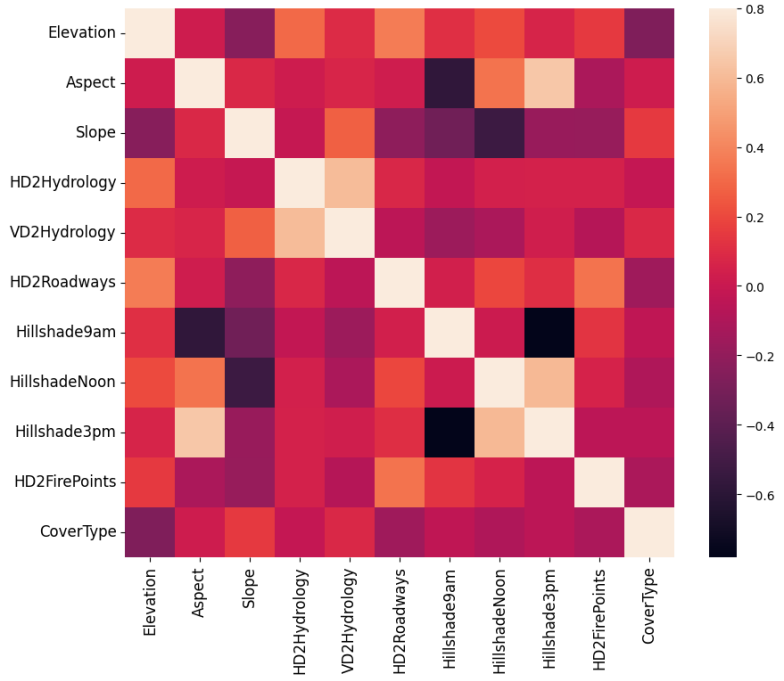
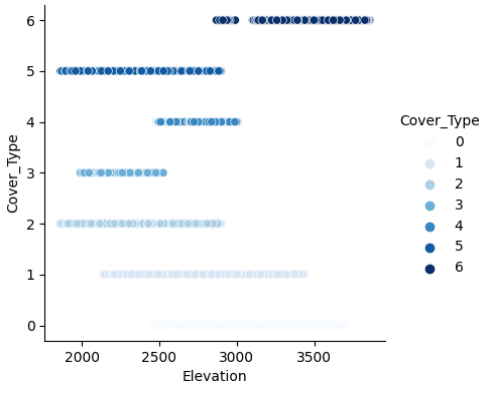
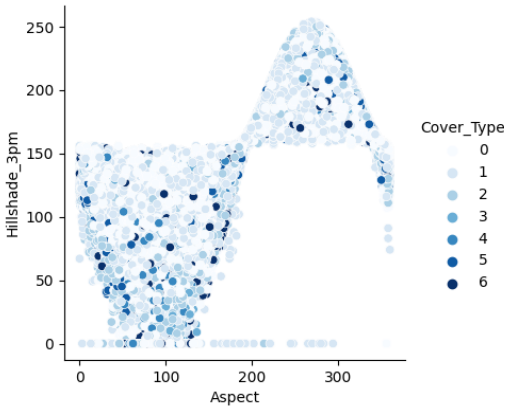


Fig 3: Heat Map showing correlation between Numerical features





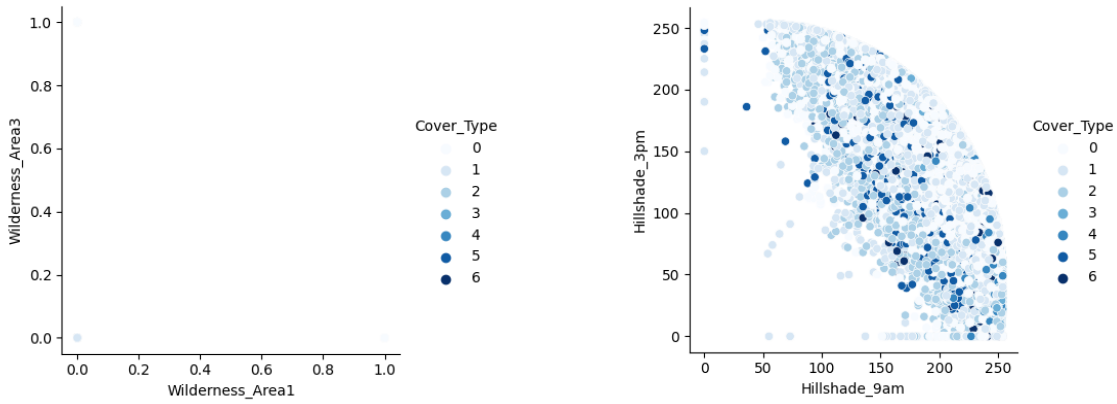


Fig 4: Pair plots between highly correlated features

Feature Engineering

Based on the correlation scores, we decided to drop the Aspect feature as it was highly correlated with both Hillshade_9am and Hillshade_3pm.

We also add a feature called Euclidean_Distance_From_Hydrology which combines the horizontal and vertical distance from hydrology into a single value. We then drop the horizontal and vertical distance features.

Modeling and Model Analysis

We tried three different models on our data, namely, KMeans, XGBoost and a Neural Network. The F1 scores for the models are in the table below. The KMeans model was unsupervised, whereas the other two were trained in a supervised fashion. The XGBoost algorithm was set to use 300 estimators and a subsampling ratio of 0.25. The Neural network consisted of ReLU activated linear layers followed by Cross Entropy loss and was trained using Stochastic Gradient Descent. The models were all evaluated by their F1 score on the test split.

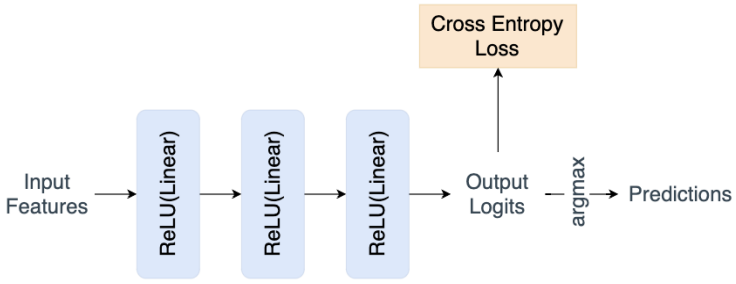


Figure 5: Neural Network Architecture

Model	F1 Score
KMeans	0.287
XGBoost	0.686
Neural Network	0.915

Table 3: Results

As expected from the homogeneity of the t-SNE visualization, the KMeans algorithm was unable to find distinct clusters in the data. The XGBoost algorithm performed better, but it too was outdone by the neural network which converged to a best score of ~91%.

Studying Feature Importance Through Ablation Tests

We also performed ablation tests, removing each feature in turn and examining the impact of the ablation on the final score. Accuracy of a XGBoost model when trained and tested on all the features of the dataset was 68.6%.

The table below shows the features that caused a drop in the accuracy score after dropping them from our model. Elevation was found to be the most impactful feature, followed by Euclidean_Distance_From_Hydrology and Slope.

Dropped Feature	Accuracy
Elevation	0.5610303858
Euclidean_Distance_From_Hydrology	0.6338500214
Slope	0.6355242158
Soil_Type4	0.6373966015
Horizontal_Distance_To_Fire_Points	0.639602783
Soil_Type32	0.6398948543
Soil_Type2	0.63999395

Conclusion

In this project we systematically present our methodology to predict forest cover types, given cartographic and topological data. We found that there is significant correlation between such features and the types of plants that will grow in these conditions. We demonstrate how we can deal with biases in the data, and model it using three different machine learning algorithms. Our experiments result in a neural model that can predict cover types with a high degree of success.