Team Name: my (ana)conda don’t

Team Members: Aditya Srivastava, Harsh Gupta, Rachna Konigari

Objective

Given topological 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 our 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 features.

Dataset Description

Overview

The dataset has been sourced from the UCI Machine Learning Repository. 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 in a $30x30m^2$ patch.

The dataset contains **581012** samples and **55** different features. Some of the critical features include elevation, distance from water sources, distance from roadways and soil type. At the moment we do not plan to drop any features from the dataset as all the features look promising, but we plan to work more on feature selection and feature engineering.

The data does not come with any predetermined train-validation-test splits, and we plan to make our own splits in the ratio of 80:20:20 by stratified sampling.

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)
