Dataset Introduction

The following analysis was conducted on a Forest Cover Type dataset sourced from the University of California, Irvine Machine Learning Repository [5]. This dataset, compiled in 1998, originated from the Remote Sensing and GIS Program conducted by the Department of Forest Sciences in Colorado State University's College of Natural Resources.

The "Covertype Data Set" contains 54 attributes and 581,012 instances. Each instance represents a 30 x 30 meter plot of land in Roosevelt National Forest, Colorado. The dataset specifies one of 7 possible cover types for each square plot with these classifications originating from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. The independent variables, acquired from US Geological Survey (USGS) and USFS data, are cartographic, including elevation, aspect, slope, horizontal and vertical distance to nearest water source, distance to nearest road, hillshade index at 9:00am, 12:00pm, and 3:00pm, distance to nearest wildfire ignition site, wilderness area (4 possible wilderness areas represented as indicator variables), and soil type (40 possible soil types represented as indicator variables). Therefore, the dataset includes categorical, numerical (both discrete and continuous), and indicator variables.

Previous Usage

This dataset was originally compiled for and used in a Ph.D. dissertation [4], later presented at the Second Southern Forestry GIS Conference at the University of Georgia [3], and then published in *Computers and Electronics in Agriculture* [2], an academic journal. The original study evaluated the relative accuracy of classification when using an artificial neural network versus a statistical model based on discriminant analysis. For all subsets of the independent variables examined, the artificial neural network performed better than the discriminant analysis in correctly classifying plots of land. The dataset has also been cited in papers investigating decision trees and clustering [1][7]. Therefore, the dataset has been primarily used to test different classification and clustering models, while very few have conducted much exploratory analysis on the data.

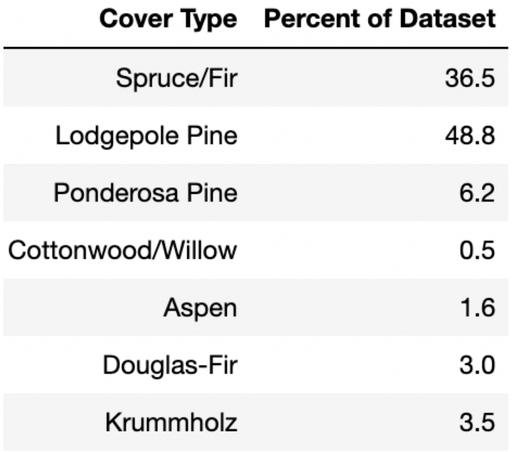
Applications of Classification

The original usage of this dataset was focused on determining the accuracies of different classification methods used to make predictions. The original dissertation [4] identifies that these predictions could act as an aid to natural resource managers in providing them with a model which can accurately predict the cover type for land outside of their immediate jurisdiction, as the data might not be readily available for these areas. The codebook for the dataset notes that the four wilderness areas included in the dataset are areas with relatively few human-caused disturbances, so the observed instances are "more a result of ecological processes rather than forest management practices" [4].

Since 1998, Roosevelt National Forest has seen an increase in wildfires with a total of five significant wildfires occurring in 2020. Two were the largest in Colorado history. According to the National Forest Service, these wildfires burned more than 25% of the land in the Arapaho and Roosevelt National Forests [6]. The wilderness areas in the dataset, largely unaffected by human disturbances, possess the same cover types naturally present in the areas of the Colorado forests which have been destroyed. With recovery projected to take years and planting efforts beginning for hundreds of the burned acres of forest, it is important for recovery efforts to restore these areas to their original cover types, as natural cover types are more likely to survive and flourish than those not natural to the areas. Using a predictive model, those working to replant these affected areas can determine the best cover type for a given plot of land, saving time and money in their efforts to help the forests recover as quickly and efficiently as possible.

Attribute Distributions

First, we created a table displaying the breakdown of what percentage of the dataset each cover type makes up, as seen in Figure 1. Spruce/Fir and Lodgepole Pine make up almost 85% of the instances in the dataset, while Cottonwood/Willow only made up 0.5%. This speaks to the relative abundance of each species in the wild in the environment from which the data was collected.



Cover Type	Percent of Dataset
Spruce/Fir	36.5
Lodgepole Pine	48.8
Ponderosa Pine	6.2
Cottonwood/Willow	0.5
Aspen	1.6
Douglas-Fir	3.0
Krummholz	3.5

Fig 1: Cover Type Distribution

Grouping the data by cover type reveals relationships between some of the independent cartographic variables and the different cover types. Figure 2 shows that Ponderosa Pine, Cottonwood/Willow, and Douglas-fir have mean elevations between 2,000 and 2,500 m, Lodgepole Pine and Aspen have mean elevations between 2,500 and 3,000 m, and Spruce/Fir and Krummholz have mean elevations between 3,000 and 3,500 m. Figure 3 shows that Spruce/Fir, Lodgepole Pine, and Krummholz have mean slopes between 10 and 15 degrees, Cottonwood/Willow, Aspen, and Douglas-fir have mean slopes between 15 and 20 degrees, and Ponderosa Pine has a mean slope above 20 degrees.

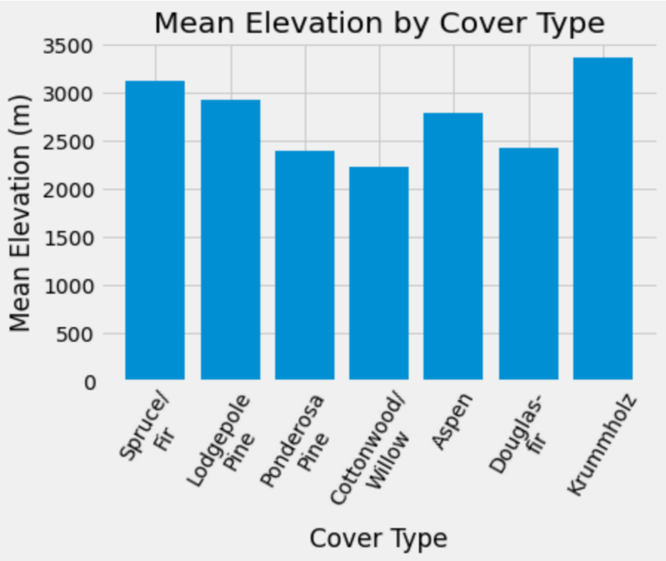


Fig 2: Mean elevation by cover type

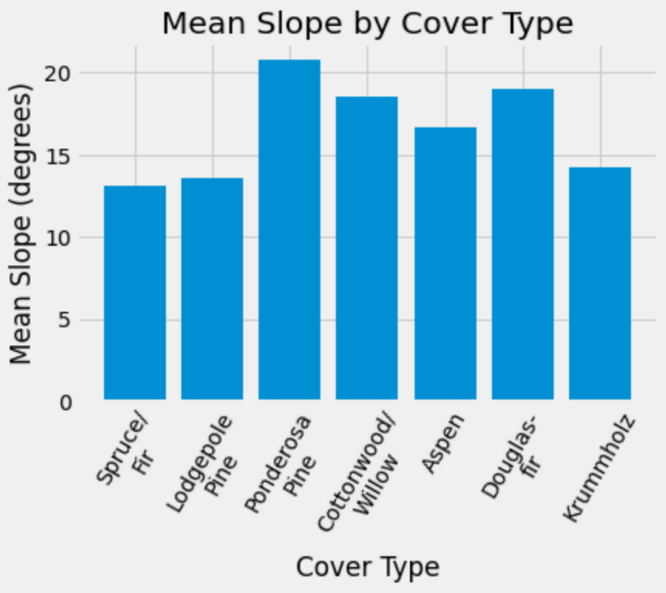


Fig 3: Mean slope by cover type

We created a histogram of the slope of all instances (Figure 4). However, we can look at Figure 4 in the context of Figure 3 to see that the mean slope for Spruce/Fir and Lodgepole Pine was smaller than many of the less common cover types (from Figure 1). However, because Spruce/Fir and Lodgepole Pine made up the majority of the dataset, they dominated the histogram in Figure 4, creating a peak at their mean slopes, even when the majority of the cover types had greater average slopes.

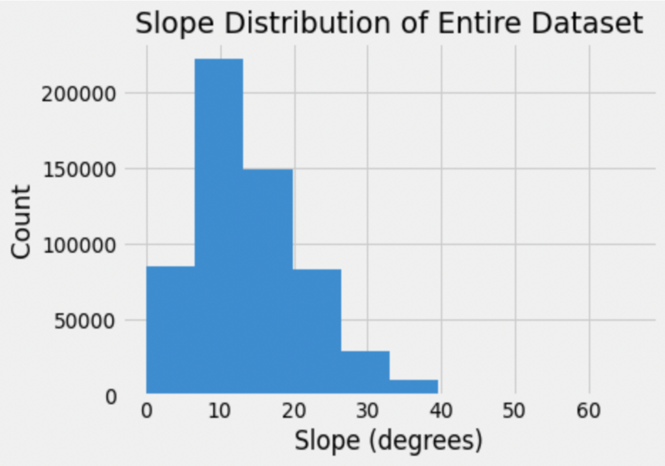


Fig 4: Histogram of slope for all cover types

Figure 5 shows mean distance to water (calculated by finding the euclidean distance based on horizontal and vertical distance to water). It shows that Cottonwood/Willow has a mean distance to water of under 125 m, Douglas-fir has a mean distance of under 175 m, and Krummholz has a mean distance to water of over 350 m. The remaining cover types have mean distances to water between 200 m and 300 m. These distributions suggest that there exist correlations between elevation and cover type, slope and cover type, and distance to water and cover type, as the different cover types have varying mean elevations, mean slopes, and mean distances to water.

We also grouped by cover type and wilderness area. This revealed that Lodgepole Pine was found in all four wilderness areas, Spruce/Fir and Krummholz were found in only the Rawah, Neota, and Comanche Peak wilderness areas, Aspen was found in only the Rawah and Comanche Peak wilderness areas, Ponderosa Pine and Douglas-fir were found in only the Comanche Peak and Cache la Poudre wilderness areas, and Cottonwood/Willow was only found in the Cache la Poudre wilderness area. This suggests that there is a correlation between wilderness areas and cover types, as only certain cover types are found in certain wilderness areas. In other words, it is likely that the conditions in each wilderness area best support specific types of forest cover.

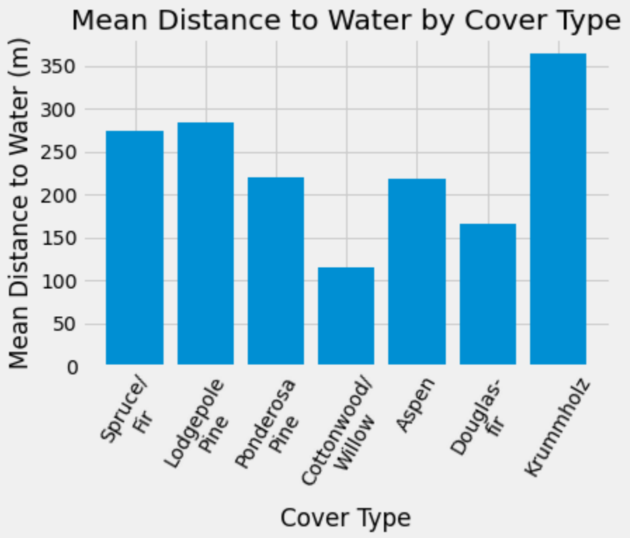


Fig 5: Mean euclidean distance to water by cover type

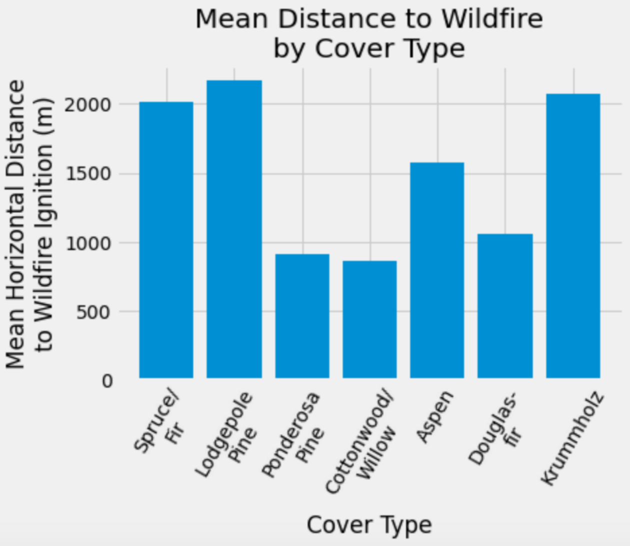


Fig 6: Mean euclidean distance to wildfire ignition site by cover type

Figure 6 shows the mean distance to the nearest wildfire ignition site by cover type. It should be noted that these ignition sites are from before 1998, and more recent wildfires have very likely changed these distances. According to Figure 6, Ponderosa Pine, Cottonwood/Willow, and Douglas-fir are all found at around a mean distance of 1,000 m from wildfire ignition sites, implying that these species may be more likely to reemerge and survive following a wildfire than other species.

We also examined the distribution of hillshade indices (an index calculating the amount of shade a location receives based on its orientation and the sun's position). We created a histogram of hillshade index at three different times throughout the day, 9:00am, 12:00pm, and 3:00pm, which is shown in Figure 7. The spread of the distribution is the largest at 3:00pm, indicating that the sun is low in the sky and hitting geophysical features angled both towards and away from it which would give some points very low indices and others very high ones. The lower mean suggests the sun was stronger in the afternoon than morning. Meanwhile, the distributions for 12:00pm and 9:00am had smaller spreads with higher mean index scores suggesting that the sun was less strong and closer to being directly overhead, causing less variation in index score.

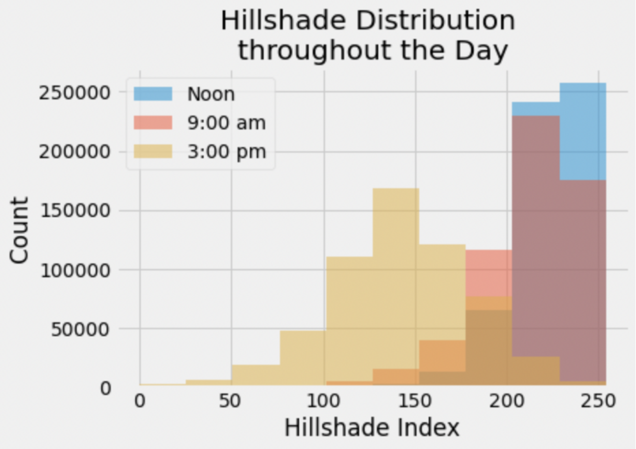


Fig 7: Histogram of hillshade indices at different points of the day

Significance of Attributes

In order to determine the significance of attributes, we first calculated the pairwise correlations for all attributes in the dataset (after splitting the cover type classification column into columns of indicator variables). We then determined the 10 variables with the strongest correlations for each cover type (not including other cover type indicator variables). Taking the set of all of these variables left us with 31 variables with which to predict cover types. These variables included elevation, slope, hillshade at 9:00am, 12:00pm, and 3:00pm, distance to wildfire ignition, horizontal distance to water, horizontal distance to the nearest road, all four wilderness areas, and 19 of the 40 total soil types.

However, we wanted to determine if any of the variables with the highest correlations were related so as not to compound their influence in our classification. So, we created scatter plots of multiple variables, including aspect vs. hillshade index at 3:00pm in Figure 8. The scatter plot makes a well-defined sinusoidal shape, which leads us to determine that aspect was used in calculating the hillshade index, making the variables dependent on one another. Specifically, this plot most likely depicts a hill with the sun shining on one side, making half the points have larger indices while supplementary aspects have lower index scores.

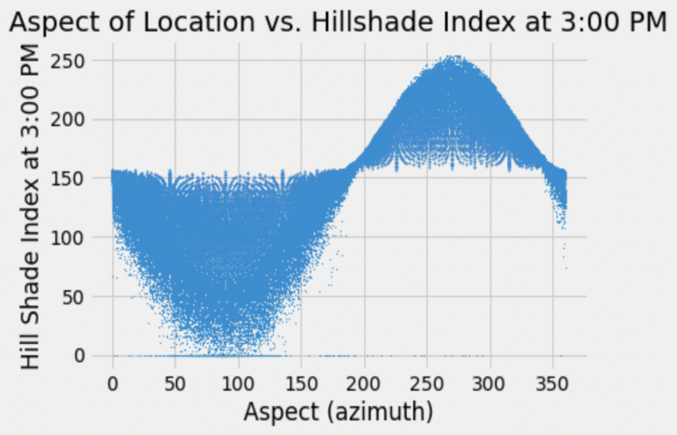


Fig 8: Hillshade index at 3:00pm by aspect

We also determined that slope was most likely used to calculate the hillshade index because, when we plotted hillshade index at 12:00pm with slope, there was an abnormal distribution with a finite lower and upper bound that can be seen in Figure 9.

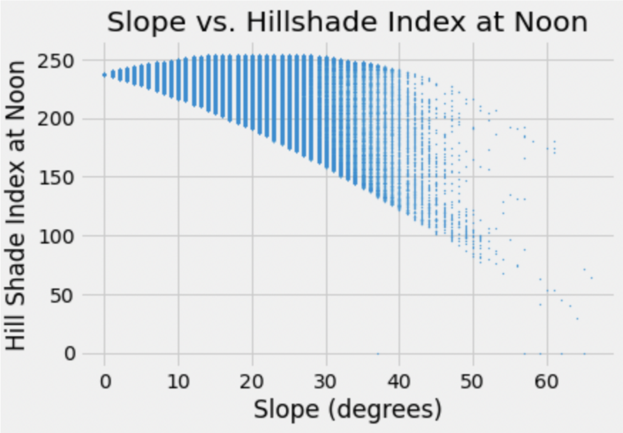
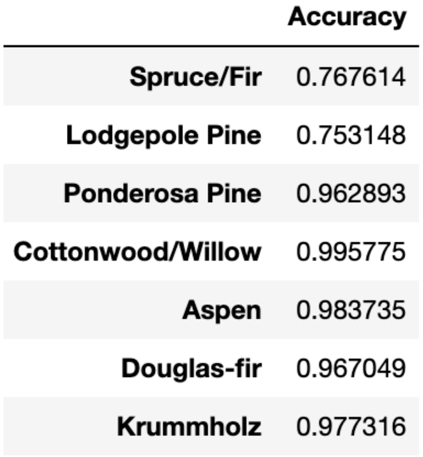


Fig 9: Hillshade index at noon by slope

Therefore, hillshade index at multiple times of day was one of the most correlated determinants of cover type, and multiple other variables are most likely used to calculate hillshade index which could be considered in our classification. However, because the relationship was nonlinear we chose to keep these variables in our classification algorithm.

Logistic Regression

The first classification method we tested was logistic regression. This required repeating for each cover type the process of defining and standardizing the predictor variables (narrowed down to the 31 with the strongest correlations) and then standardized using MinMax scaling. Then we defined the outcome variable as one of the cover types (as an indicator variable), splitting this data into training (80%) and test (20%) datasets, fitted the logistic regression model to the training data, and finally predicted outcomes for the test data. We measured accuracy as the proportion of correctly classified instances in the test data. Figure 10 reports these accuracies by cover type. Notably, Spruce/Fir and Lodgepole Pine were less accurately classified using this method than the other cover types. However, Spruce/Fir and Lodgepole Pine made up over 85% of the dataset while other cover types (like Cottonwood/Willow) made up less of the data (from Figure 1). This could explain why cover types with smaller sample sizes were more accurate because they only grew under specific (less common) conditions, making them easier to classify.



	Accuracy
Spruce/Fir	0.767614
Lodgepole Pine	0.753148
Ponderosa Pine	0.962893
Cottonwood/Willow	0.995775
Aspen	0.983735
Douglas-fir	0.967049
Krummholz	0.977316

Fig 10: Logistic regression accuracy by cover type

kNN Classification

Finally, we used a kNN classification (with k=5) to predict the cover type of each instance based on all the variables in the dataset. To do this, we split the data such that 80% was used to train the model and 20% was used to test it. When we tested the classification on the withheld 20% of the data, we predicted the cover type with an accuracy of 92.8%. We then repeated the classification with cross validation, randomly splitting the training dataset into four equally-sized parts. We retrained the model, withholding each of the four validation subsets and averaging the resulting accuracies. With cross validation, the average accuracy was 93.9%.

We then performed the same classification procedure again, narrowing the included predictor variables down to the 31 most correlated variables. We again split the data into training (80%) and test (20%) datasets. We first trained the model with the entire training dataset and then tested it with the test dataset to get an accuracy of 93.9%. We then performed the same classification with the cross validation detailed above to get an average accuracy of 93.0%. In summary, the accuracy without cross validation was higher when the predictor variables were limited, while the average accuracy with cross validation was lower when the predictor variables were limited. This implies that limiting the predictor variables to those with the strongest correlations better fits the training data, but does not improve the accuracy of the model in predicting the test data.

Conclusion

In conclusion, exploratory analysis on the dataset highlights how certain characteristics like slope and distance to wildfire ignition can be used to differentiate certain cover types' growing conditions from others. We can combine these factors to create a kNN classification that predicts cover type with a >90% accuracy rate.

Our exploration and classification model can be useful for landowners to help rehabilitate the environment in the aftermath of wildfires, making it particularly useful in recent years. So, while prior individuals used this dataset to compare predictive models in theory, our analysis has practical applications for the environment. Our findings that certain cartographic characteristics are good predictors in the classification of individual tree species suggests that this methodology could be applied to other environments with different tree species outside of Colorado, making it potentially useful in environments impacted by other natural disasters or habitat destruction.

Bibliography

[1] Arto Klami and Samuel Kaski and Ty n ohjaaja and Janne Sinkkonen. 2003. “Regularized Discriminative Clustering.” Helsinki University of Technology. Department of Engineering Physics and Mathematics. http://users.ics.aalto.fi/aklami/papers/regularized_dc.pdf.

[2] Blackard, Jock A. and Denis J. Dean. 2000. "Comparative Accuracies of Artificial Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables." *Computers and Electronics in Agriculture*. 24(3):131-151.

[3] Blackard, Jock A. and Denis J. Dean. 1998. "Comparative Accuracies of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types from Cartographic Variables." *Second Southern Forestry GIS Conference*. University of Georgia. Athens, GA. Pages 189-199.

[4] Blackard, Jock A. 1998. "Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types." Ph.D. dissertation. Department of Forest Sciences. Colorado State University. Fort Collins, Colorado. 165 pages.

[5] “Covertype Data Set.” *UCI Machine Learning Repository*. University of California, Irvine. 1998. https://archive.ics.uci.edu/ml/datasets/Covertype.

[6] “Fire Recovery Information.” *Forest Service National Website*. U.S. Department of Agriculture. https://www.fs.usda.gov/detail/arp/home/?cid=FSEPRD906576.

[7] Gama, João & Rocha, Ricardo & Medas, Pedro. 2003. “Accurate decision trees for mining high-speed data streams.” Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://www.researchgate.net/publication/221653305_Accurate_decision_trees_for_mining_high-speed_data_streams.