Applying Machine learning to predice tree cover in the Roosevelt National Forest in Colorado

Final Project Proposal for Data Science Professional Certificate
Harvardx Professional Certificate Program

Felipe Vergara, M.Sc

Submission Date
December 23th, 2021

Summary

1 Introduction

This is the final report to obtain the professional certificate grade from Hardardx. The goal of this final exercise is to apply a machine learning method to a free database or a personal provided database. In this case, it was decided to use the database **Forest Covertype data**, provided by the *Remote Sensing and GIS program Department of Forest Sciences College of Natural Resources Colorado State University*. The data was released free in 1998, and it can be downloaded by this link.

The purpose of this project is to predict forest covers from cartographic variables such as elevation, slope, distance to roads and others. The unit analysis are cells with a size of 30 x 30 meter determined by the US Forest Service (USFS), while independent variables were obtained from US Geological Survey (USGS) (Blackard et al., 1998). The research area is compounded by four wilderness areas in the Roosevelt National Forest Colorado characterised by being with minimal human-caused disturbances (Blackard et al., 1998).

It was selected this problematic, because it is a breakthrough in the thematic improving the tree classification to a high accuracy level. This undoubtedly can help policymakers and environmental policies to identify priorities areas to protect certain biodiversity or trees as in this case. Furthermore, the selected machine learning method was **random tree forest**. This decision is based on the amount and type of data, which frames the scope of the application. In the next steps it is justified this decision.

The next section shows starting how was the data imported and the corresponding exploration to determine what variable could light up a clue for the prediction. Then it is presented the machine learning methodology. Finally, it is displayed the results and the conclusions.

2 Preprocessing

2.1 Libraries

Diverse tools are necessary to run the project, thus the libraries here below where used through the project:

```
library(tidyverse)# organization and visualization data
library(kableExtra) #for table presentations
library(caret)# Used in the script, not in the report
library(purrr)# to extract data info
library(scales)# to add big mark point as thousand separator
library(treemapify) # to create treemaps
library(gridExtra) # to arrange plots in a grid
library(randomForest) # to run random forest
```

2.2 Importing Forest Covertype data & Exporting data

2.2.1 Starting processing

First of all, it is necessary to create a link to connect R with the raw data to download it. Thus, it is created 2 temporal vectors, one for the dataset and another one for the data info, which contains the raw column names.

```
options(width = 60) #to extend screen width to 10 characters per line
#Download datasets
df <- tempfile()#temporal datatable
download.file(
  "https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.data.gz", df)
```

```
infod <- tempfile()#temporal data info
download.file(
  "https://archive.ics.uci.edu/ml/machine-learning-databases/covtype/covtype.info", infod)

trees_c <-str_split_fixed(readLines(gzfile(df)),",",55) #dataset
info<-readLines(infod)# data info to assign column names to trees_c dataset
```

The dataset contains 31955660 registers, which it can be understood as a big dataset (Table 1):

Table 1: Raw data dimension

Rows	Columns
581,012	55

The relevant from this data is that there are 12 measurements, but represented with 54 columns. 10 are quantitative variables and 2 qualitative (wilderness area and Soil type), which are represented by 4 and 40 columns correspondingly. The last column is the dependent variable which it will be predicted, the cover type. The table below indicated the attributes and characteristics, but not the cover type variable (Table 2):

Table 2: Table data description

Name	Data type	Measurement	Description
Elevation	quantitative	meters	Elevation in meters
Aspect	quantitative	azimuth	spect in degrees azimuth
Slope	quantitative	degrees	Slope in degrees
Horizontal_Distance_To_Hydrology	quantitative	meters	Horz Dist to nearest surface water features
Vertical_Distance_To_Hydrology	quantitative	meters	Vert Dist to nearest surface water features
Horizontal_Distance_To_Roadways	quantitative	meters	Horz Dist to nearest roadway
Hillshade_9am	quantitative	0 to 255 index	Hillshade index at 9am, summer solstice
Hillshade_Noon	quantitative	0 to 255 index	Hillshade index at noon, summer soltice
Hillshade_3pm	quantitative	0 to 255 index	Hillshade index at 3pm, summer solstice
Horizontal_Distance_To_Fire_Points	quantitative	meters	Horz Dist to nearest wildfire ignition points
Wilderness_Area	qualitative	0 (absence) or 1 (presence)	Wilderness area designation
Soil_Type	qualitative	0 (absence) or 1 (presence)	Soil Type designation

2.2.2 Special consideration:

2.2.2.1 Aspect

Table 2, shows that the *Aspect* variable is measured in *Azimuth* and is quantitative. Notwithstanding, for this topic it cannot be understood in that way, instead as a nominative data type. For example, when is treated this value as quantitative, the difference between 357 and 1 is 356. However, as a nominative in this respect, both values only have a difference of 4 and both indicate a north direction.

Due to this situation is aggregated the aspect values in values of 90° in a new column. So:

- 315°-45°: North
- 45°-135°: East
- 135°-225°: South
- 225°-315°: West

```
#create new 4 columns to identify the north, east, south and west correspondingly.
#they are dummy variables
trees_c<-trees_c%>%mutate(
              Aspect_1= (Aspect>=0 & Aspect<45 | Aspect>=315 & Aspect<=360)*1,#North
              Aspect_2= (Aspect>=45 & Aspect<135)*1,#East
              Aspect_3= (Aspect>=135 & Aspect<225)*1,#South
              Aspect_4= (Aspect>=225 & Aspect<315)*1)#West


trees_c$Aspect_label<-factor(
  ifelse(
    trees_c$Aspect >=0 & trees_c$Aspect<45 | trees_c$Aspect>=315 & trees_c$Aspect<=360,
    "North",
                            ifelse(trees_c$Aspect >=45 & trees_c$Aspect<135, "East",
                                   ifelse(trees_c$Aspect >=135 & trees_c$Aspect <225,
                                          "South",
                                          "West"))),
  levels = c("North","East","West","South"))
```

Additionally, it wass added **4** dummy columns for each orientation mentioned above, to facilitate the procedure.

2.2.2.2 Soil type

This type is the most populous in the dataset, containing more than 40 classes. According to the results of (D. Fernandes Terra Machado, 2019), it seems that is more relevant to find the proper dataset, rather than evaluating great number of variables. Thus, taking into account the soil data characteristics, it was performed a classification of the soil type according to its firsts digits codes, in which the first and second digit indicates climatic and geological zone correspondingly of the USFS Ecological Landtype Units (Blackard et al., 1998). As a result, it was split the soil type in soil climate zone (abb:s_cli) and soil geological zone (abb:soil_geo), decreasing the classes from 40 to 11 classes (Table 3).

Table 3: Soil classifications

Soil Climatic Zone	Soil Geological Zone
lower montane dry	alluvium
montane dry	glacial
montane	mixed sedimentary
montane dry and montane	igneous and metamorphic
montane and subalpine	-
subalpine	-
alpine	-

2.2.3 Final adjustments

To facilitate the analysis and visualization, it was added the following columns:

- Id column
- Wilderness area factor column
- Soil climatic zone factor column
- Soil geological zone factor column

- Cover type factor column

In the end, the final raw dataset consists of the following dimensions (Table 4):

Table 4: Raw data dimension

Rows	Columns
581,012	36

2.2.4 Exporting data: Training and Test data

Random forest is a machine learning application, so it must be created a training data and testing data. **the former is for the algorithm creation, and the latter is for assessing our final algorithm**. Generally, during the procedure is used mostly the training data which for this project was the 90% of the dataset. The partialization was done through the **caret package**: **create DataPartition**.

From now on the training and testing data will be processed.

3 Analysis

3.1 Data description

In order to reduce hardware processing, the training, testing data and other resources were loaded in this step.

Then, the visualized data of the data description section, **which corresponds to the training data**, is presented in the following order: dependent variable, qualitative variable and quantitative variable.

3.1.1 Cover type

The starting point is to know what it wants to be predicted. In this sense, the cover type variable is the dependent variable which consists in 7 categories.

```
## [1] "Aspen"             "Lodgepole Pine"
## [3] "Spruce/Fir"        "Krummholz"
## [5] "Ponderosa Pine"    "Douglas-fir"
## [7] "Cottonwood/Willow"
```

The first question here is to know how they are represented on the sample. Let see the pixel number for each cover type (Figure 1):

The Figure 1 unveils that there are 2 cover types (Lodgepole Pine, Spruce/Fir) predominates on the sample, while Cottonwood/Willow presence is marginal. Let's take a deeper view on the proportions (Figure 2).

This view clarifies that the outcome prevalence should be considered, when the analysis is performed. Approximately, the 85% cover type classification is concentrates in 2 types.

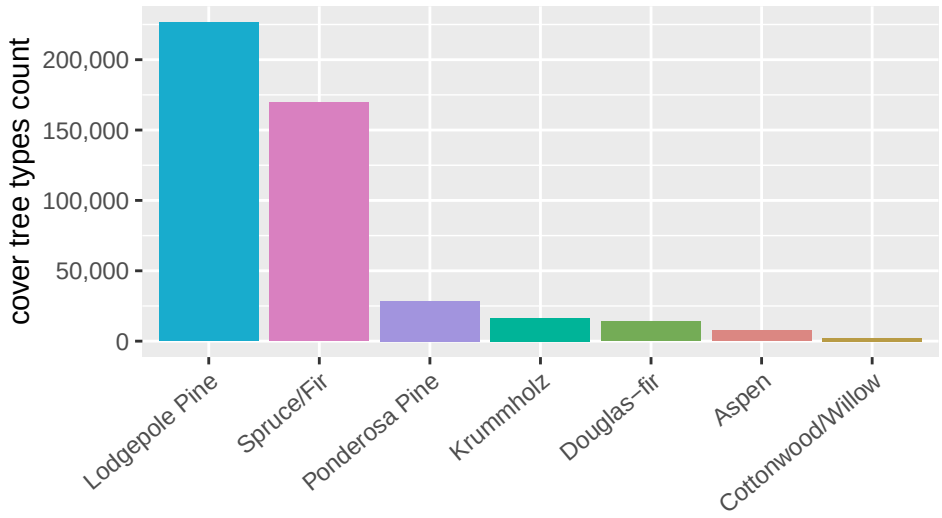


Figure 1: Cover tree types count (top) and tree type (bottom)

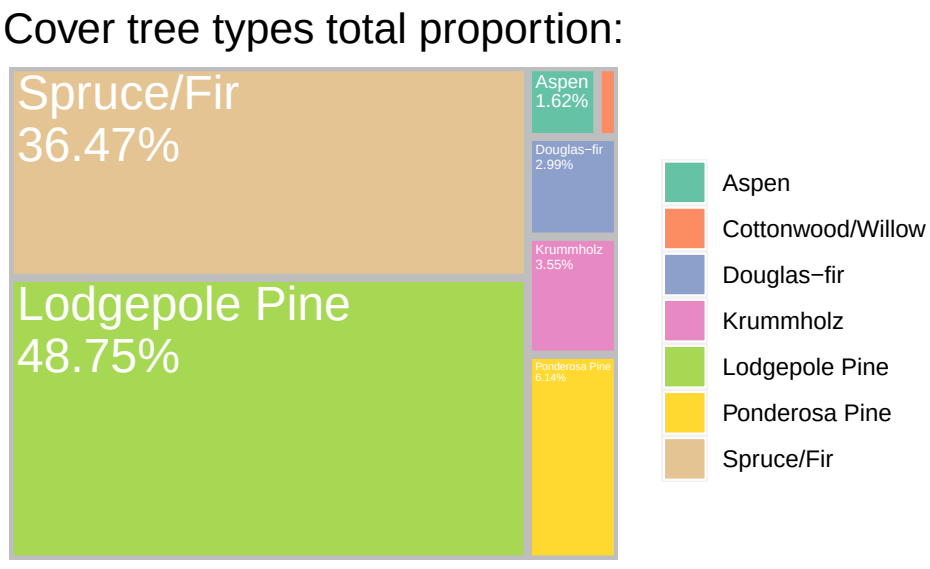


Figure 2: Treemap plot

3.1.2 Aspect

Regarding this variable, let's see its distribution according to the cover type (Figure 3).

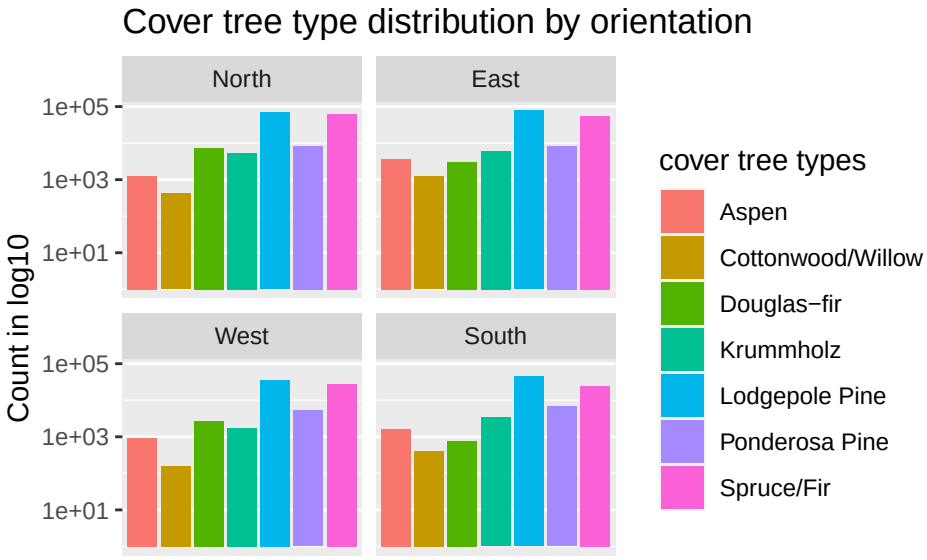


Figure 3: Tree distribution by Aspect

The Figure 3 is showing that practically the orientation does not have an implicit effect over the cover tree type distribution. Therefore, it was analysed the aspect variable to check if its values are independent or not by the chi-square test.

First, it was created a frequency table (Table 5).

Table 5: Frequency table Aspect

Cover_tree_type	North	East	West	South
Aspen	1,284	3,728	929	1,594
Cottonwood/Willow	417	1,212	159	419
Douglas-fir	7,488	3,003	2,657	766
Krummholz	5,398	5,977	1,750	3,358
Lodgepole Pine	68,261	77,888	35,739	44,711
Ponderosa Pine	8,200	8,404	5,236	6,717
Spruce/Fir	61,919	54,656	28,157	24,781

Then it was performed the chi-square analysis with a p-value of:

```
chisq_test <- faspc %>%dplyr::select(-Cover_tree_type) %>% chisq.test()
chisq_test$p.value #the values are correlated, not independent
```

```
## [1] 0
```

Under that result, the aspect variable was not considered as part of the following analysis, since its values are correlated.

3.1.3 Wilderness area

From the plot below (figure 4), it can be seen that there are cover types not presented in the 4 wilderness areas. Only Lodgepole Pine is presented in all the research areas. Therefore, this variable could be considered for the algorithm since there is a certain location dependency for the cover types.

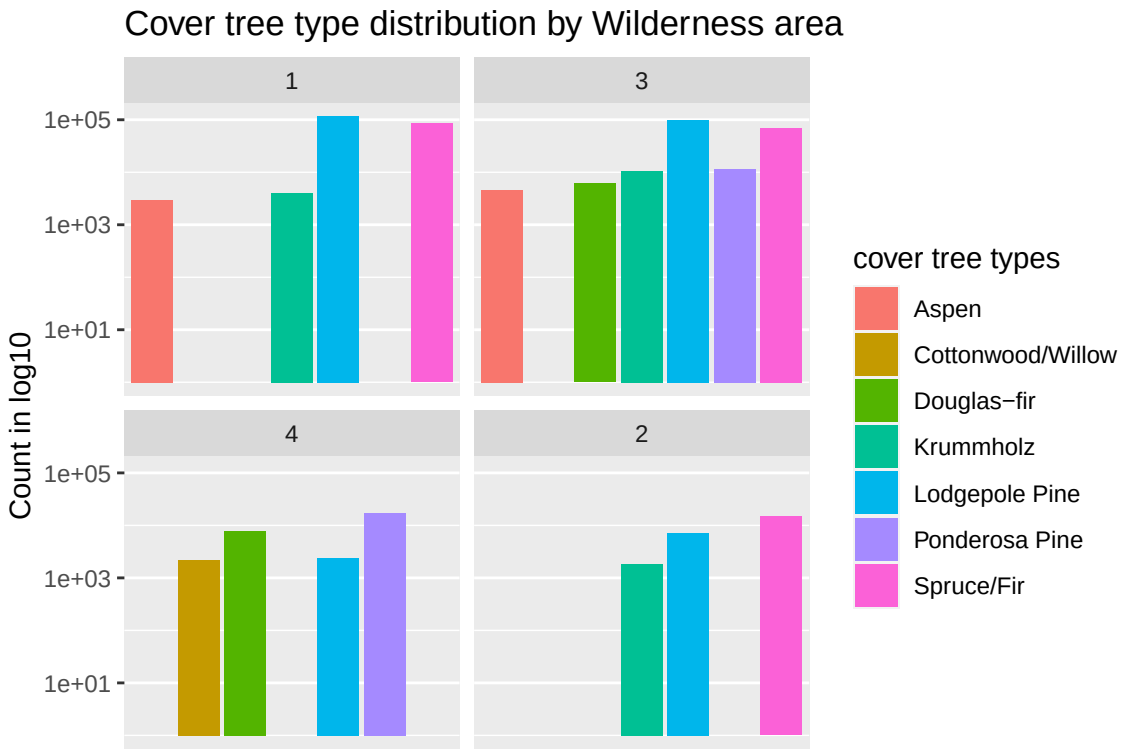


Figure 4: Tree distribution by Wilderness area

3.1.4 Soil type

By climatic zone, it can be seen that the those more present species are not present in *Monate dry and montane*. Also, there are soil climatic zones more populous than others. This is just a clue that this variable can give us information to determine a certain pattern (figure 5).

Regarding to the soil geological zone, there is a certain effect since not all species are located in a equally rate, indicating a pattern. For example, mixed sedimentary soils are suitable for only Spruce and Lodgepole Pine species (figure 6).

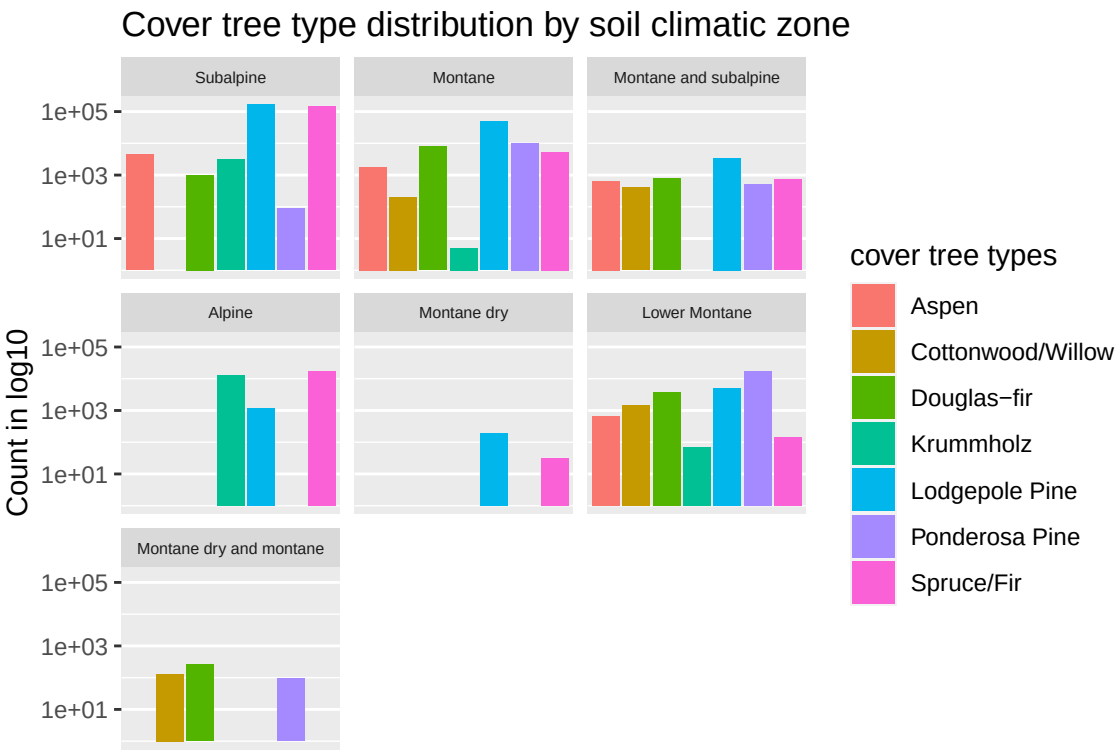


Figure 5: Tree distribution by soil climatic zone

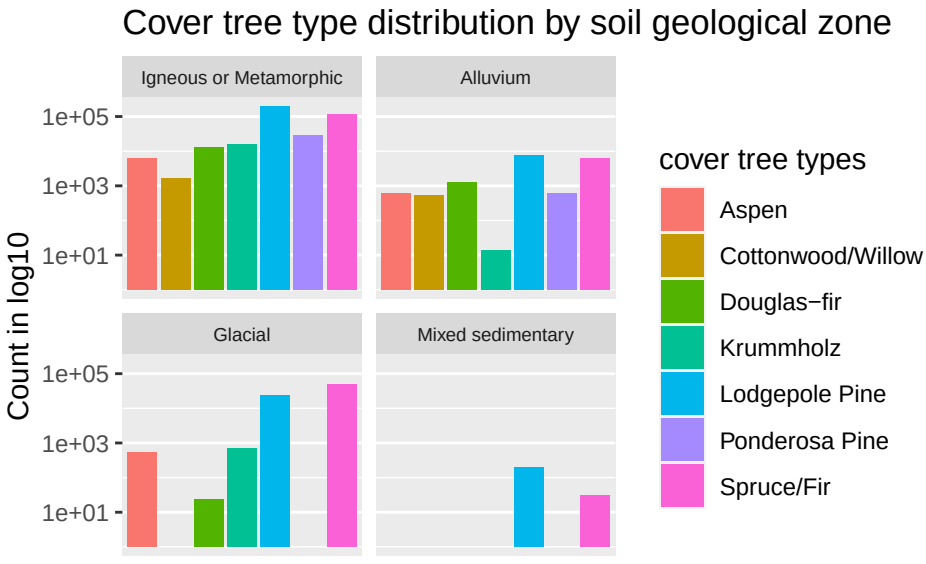


Figure 6: Tree distribution by soil geological zone

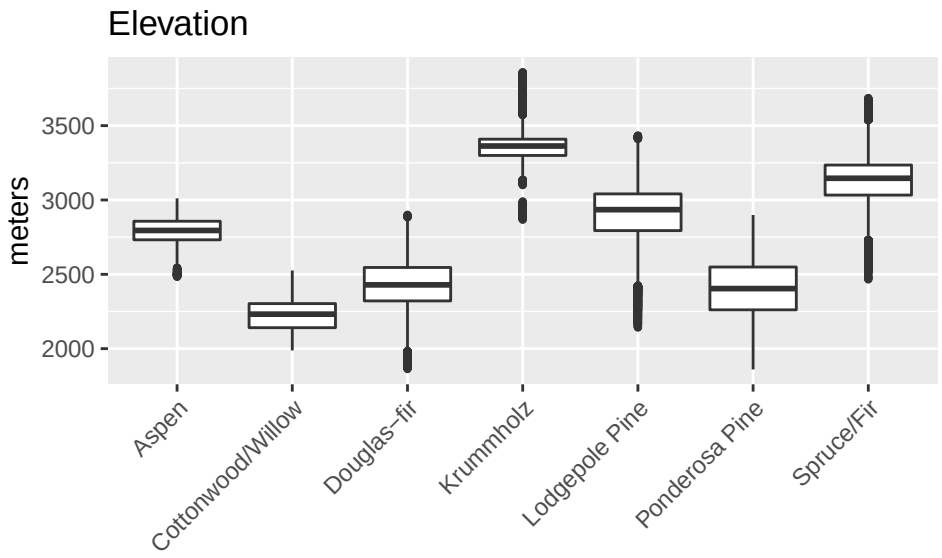
3.1.5 Quantitative variables

Afther the adjustment with the qualitative variables, there are 9 quantitative variables to analyze:

- Elevation
- Slope
- Horizontal distance to hidrology
- Vertical distance to hidrology
- Horizontal distance to roadways
- Hillshade 9 am
- Hillshade noon
- Hillshade 3 pm
- Horizontal distance to fire points

3.1.5.1 Elevation

Through a visual analysis it can be defined that the tree species have different distribution according to the elevation (figure **??**). Where it can be found similar distribution is with the Krummholz and Ponderosa Pine species. Therefore, is relevant to include other variables.



3.1.5.2 Slope

At difference with elevation variable, the species with the slope variable (figure 7),are more similar in their distribution, notwithstanding, is relevant that Krummholz and Ponderosa Pine species are different in this case, so this can help us to classify them better.

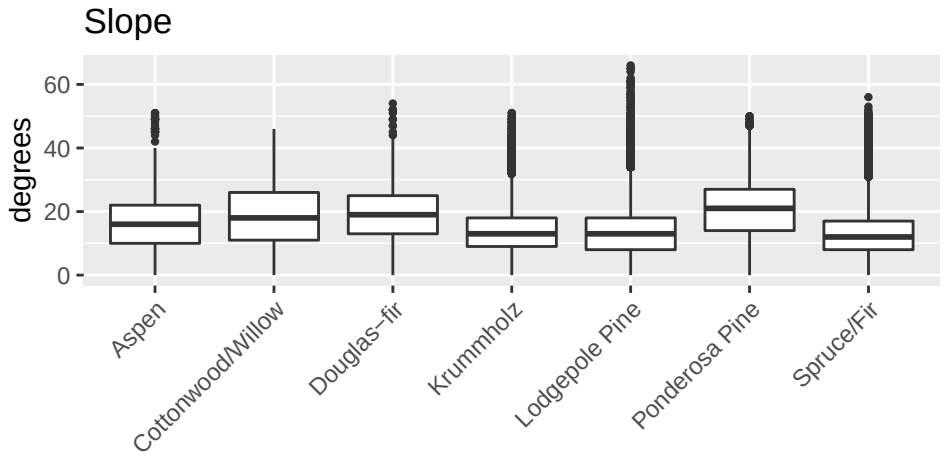
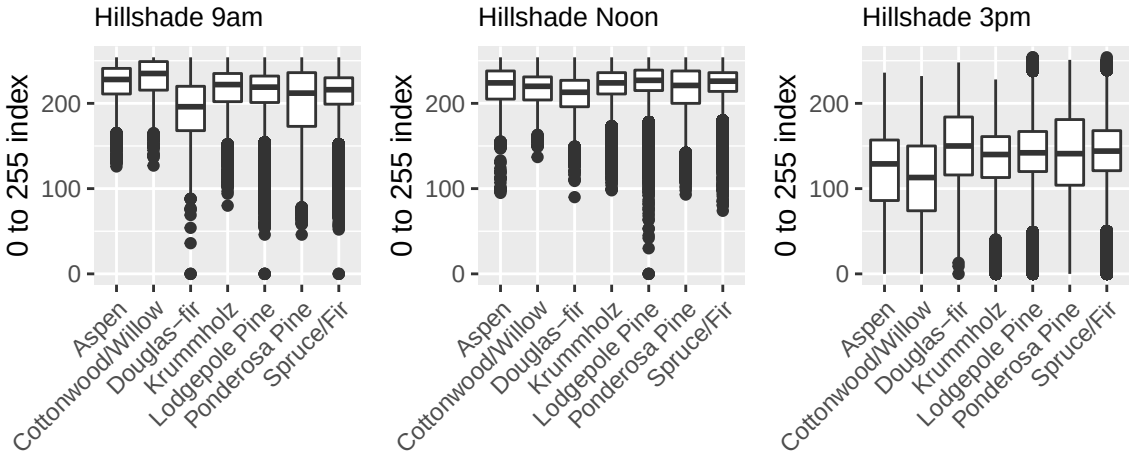


Figure 7: Treemap plot Wilderness area

3.1.5.3 Hillshades values

In this case, we can compare the values of these 3 graphs (figure **??**), indicating the chrome value in RGB units of the 7 species. In overall, they are quite similar but occurs a trend that the values start to become similar from the morning to the afternoon, being the *hillshade noon* the variable that provides less differences. As a result, it was decided that the variable *hillshade noon* be assessed through the ANOVA test (see ANOVA Test section). Another identified aspect was that a relevant difference was found in *Hillshade 9* am where Aspen and Cottonwood/Willow values are concentrated in higher values in comparison with the other species.



3.1.5.4 Horizontal and vertical distances

In relation to the horizontal distance to nearest surface water features (figure 8), there are similar distributions, so this variable does not provide much information. A consideration for this value is for the Cottonwood/Willow species which seems to crucial dependence to be near of waterbodies or waterlines.

According to the vertical distance to the nearest water surface features, the mean values are mostly extremely similar, but it can be confirmed that the Cottonwood/Willow depends more on water closeness. Also in general the species in terms of vertical distance are close to water features.

Regarding the horizontal distance to roadways, here it can be found relevant differences. Now how to justify this variable? In general the roads are located in areas where is much simpler to build, so it can be an aggregation of slope and closeness to waterbodies. However, these are suppositions and this variable is just one of the variables where it can be seen a human effect. In this case, it can be identified 2 groups Krummholz, Lodgepole Pine and Spruce/Fir in one side and the rest in the other group.

Finally with the variable horizontal distance to fire points it can be seen a pattern. This can be justified in terms of preponderance to ignite much faster certain species than others, due to their shape and combination of chemical and structural plant (Blauw et al., 2017). Thus, it is assessed as a relevant variable.

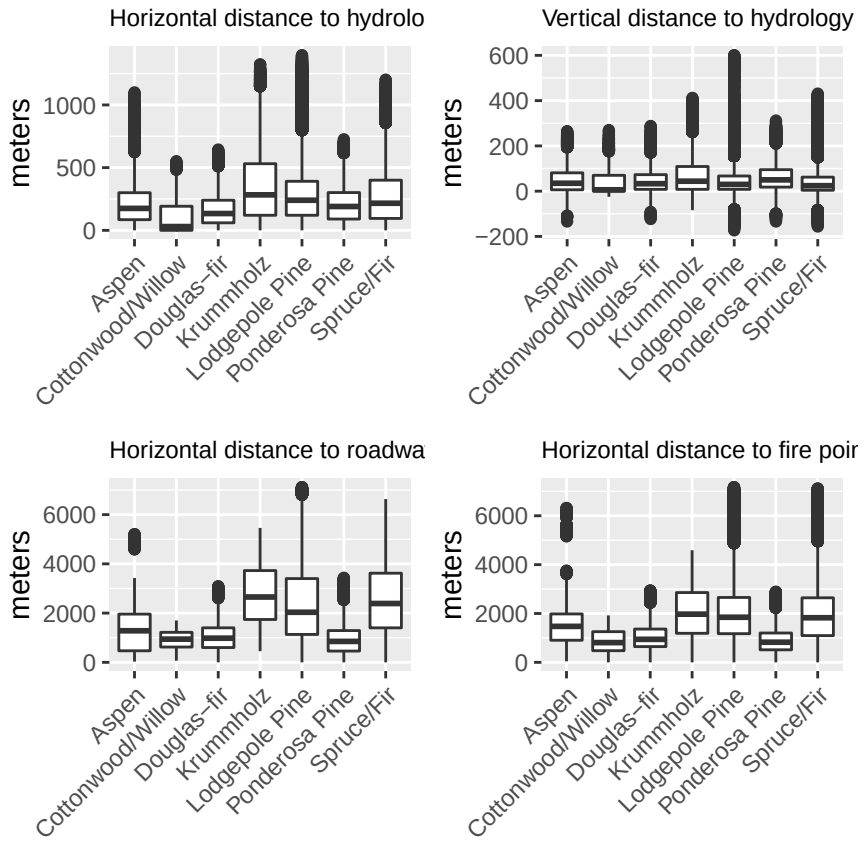


Figure 8: Horizontal and vertical distances boxplots

3.1.5.5 ANOVA Test

To provide a value to indicate that the variables are significant and different among them, it was applied a one-way ANOVA test. This allows to compare multiple groups of data of one independent variable, measuring the

mean of independent groups (Bevans, 2020). In this sense, the one-way ANOVA test was used for the variable *hillshade noon* to identify if the groups of this variable are different or not.

First, it was filtered out the outliers (figure 9).

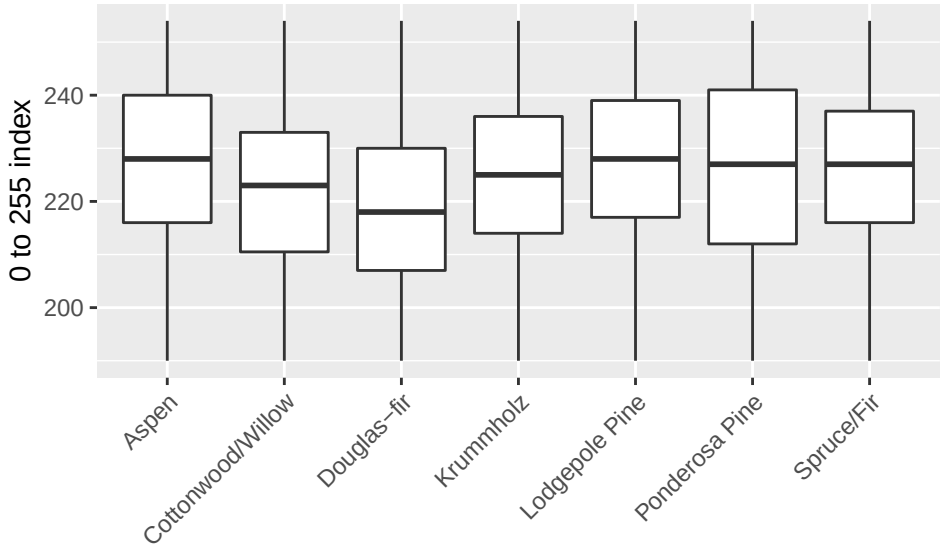


Figure 9: Hillshade noon filtered out outliers

Then, it was calculated the inter and intra-group variance. If the p-value from this calculation is below 0.05, means that the variable is not correlated (h_a), and viceversa if the p-value is above 0.05 (h_0).

In the results below it can be seen that p-value (Pr>F) is statistically significant, so that indicates the *hillshade noon* groups values are significantly different, and thus be included in the analysed variables.

```
##                 Df    Sum Sq Mean Sq F value Pr(>F)
## name_ctt         6   5252121  875354    2304 <2e-16 ***
## Residuals   464801 176617412     380
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After all this exploratory analysis and 2 applied tests (Chi-square and ANOVA test), **it was decided to use all the available data except the aspect variable** (Table 6).

Table 6: Selected variables

Variables	Variables
Elevation	Hillshade_Noon
Slope	Hillshade_3pm
Horizontal_Distance_To_Hydrology	Horizontal_Distance_To_Fire_Points
Vertical_Distance_To_Hydrology	s_clim_type
Horizontal_Distance_To_Roadways	s_geo_type
Hillshade_9am	Warea

3.2 Random forest approach

The random forest is the improvement of the decision tree approach. A decision tree consists of a flow chart of yes/no questions, which can be a classification or regression tree. The former refers to categorical vari-

ables, while the latter to continuous values. The method is to apply partitions, in simple words classify, to the predictors aiming to predict the outcome variable Y (Irizarry, 2021).

Meanwhile, the random forest is the modification of bagged decision trees to improve predictive performance (B. Boehmke, 2020). The difference with the bagged decision trees is that random forest reduces the tree correlation by injecting randomness into the tree-growing process (T. Hastie, 2008). Several decision trees are randomly created according to the predictor's variables. They can be classification and regression trees, being each decision tree independent and unique due to a bootstrap application. Afterwards, it is calculated the average of all of them producing a final prediction (Irizarry, 2021).

(T. Hastie, 2008) et al, founded that it is not necessar to create many trees to fit the random forest approach. They indicated that with 200 trees is sufficient. According to this, it was defined to use 300 trees for each random forest algorithm.

3.2.1 Interpretability

One disadvantage of random forest is that the interpretability is low. There is no clue which predictor was part of a decision tree. Notwithstanding, it is possible to identify how many times a predictor was used in the random forest through the *variable importance* function. Thus, being interpreted which variable is more important in the analysis. This is applied in the Results interpretability section.

3.2.2 Tackling Imbalance through subsamplings

It was decided to create subsamples from the training and test data since as it was seen in Cover type section, there are big disparities across the cover type classification. Thus, it was applied a down-sampling method, which match the observation of all the classes with the least prevalent class (Kuhn, 2019). In this case, the training subsample represents the 3% of the original training data, while the test subsample almost the same.

```
#Creation of subsamples for the training and test dataset,
#due to imbalance of the cover type classification
#Is applied a down-sampling tuning
set.seed(13, sample.kind="Rounding")# the seed is a random selection
down_train <- downSample(x = train_tc[, -ncol(train_tc)+1],
                         y = factor(train_tc$Cover_Type))
index<-sample(nrow(down_train))
x<-train_tc[index,]# 15449 units
y<-factor(train_tc$Cover_Type[index])

down_test <- downSample(x = validation[, -ncol(validation)+1],
                         y = factor(validation$Cover_Type))
test_rf_index<-sample(nrow(down_test))
x_rf_test<-validation[test_rf_index,] #3780 units
y_rf_test<-factor(validation$Cover_Type[test_rf_index])
```

Afterwards, these results were saved to do not repeat previous script processes.

3.2.3 Creating Algorithms

Since Random Forest applies a mix of combinations to find the best outcome, it was created 7 combinations according to its data characteristics. For all the algorithms it is present the quantitative variables, but for Wilderness Area and soil variables (climate and geological), it was considered if they are boolean or nominals, filtered out separately and jointly. The purpose of it is to find which data type is better for the Random Forest application. Also, it is important to remember that *Aspect* is not considered anymore.

- AL 1: Quantitative variables.
- AL 2: Quantitative variables + Nominal Wilderness Area and Soil variables.
- AL 3: Quantitative variables + Boolean Wilderness Area and Soil variables.
- AL 4: Quantitative variables + Nominal Soil variables.
- AL 5: Quantitative variables + Boolean Soil variables.
- AL 6: Quantitative variables + Nominal Wilderness Area.
- AL 7: Quantitative variables + Boolean Wilderness Area.

4 Results

4.1 Optimal Node and Cross validation

The next step was to identify the optimal number of nodes for each of the above algorithms. The applied method was the cross validation, creating 3 analysis groups. It was not used more, due to the large time necessary to compute that process.

```
control <- trainControl(method="cv", number = 3)#three cross validation
grid <- data.frame(mtry = c(1,seq(5,25,5)))#number of randomly selected predictors for each split.
train_rf <- train(x[, col_index], y,
                  method = "rf",#indicates random forest method
                  ntree = 150,#number of trees for each forest
                  trControl = control,
                  tuneGrid = grid,
                  nSamp = 10000)#random sample of observations for each tree

###For the rest was used the same logic
```

Then it was compared the algorithms finding that the algorithm 3 provides the most optimal accuracy (see below).

```
## [1] 0.8743789
```

As it can be seen in the table 7 the differences between algorithms are marginal, but in terms of Machine Learning, they are relevant. At first sight, it can be stated that the presence of the soil and wilderness area variables make a difference to increase the accuracy. Without them (AL 1), there is a difference of 1.8% with AL 2, and of 2.2% with the AL 3. Secondly, when is compared the variables individually in terms of Nominal-Boolean, for the case of soil it was obtained better accuracy with Nominals, and for the Wilderness Area case better accuracy with boolean. Therefore, to determine which approach is the most suitable one, it could depend on how many classes are in each variable. More classes could be better a nominal approach, and for less classes (e.g. 4) a boolean approach. Notwithstanding, this assumption was not tested and it can be refuted, due to in several tests the accuracy turned to the other value.

Table 7: Table accuracy algorithm values

tr_group	max. acc
1	0.8501525
2	0.8684701
3	0.8723537
4	0.8605089
5	0.8602500
6	0.8594726
7	0.8619328

Additionally, here is proved that is not necessary to use many variables to get better results (Figure 10). Generally for six ALs, the best accuracy was obtained with less than 15 variables. Only for the best AL (AL 3) it was obtained with 20 variables, which also was not the maximum implemented in this analysis.

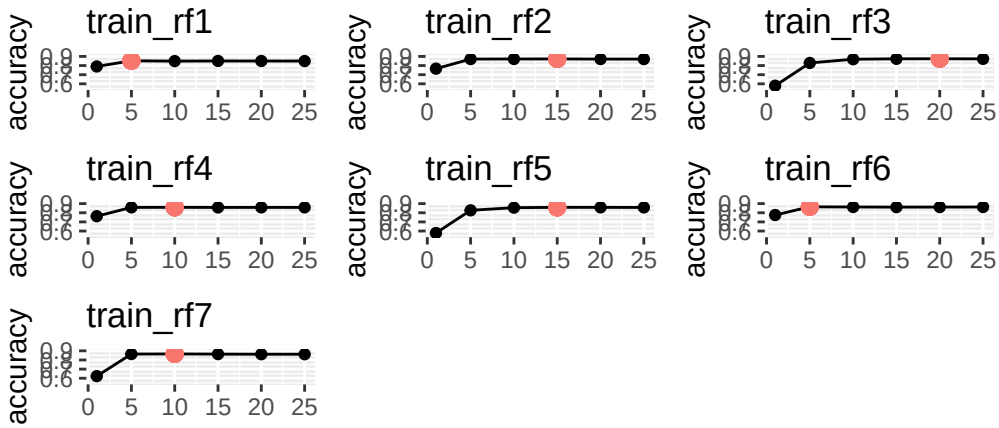


Figure 10: Comparison between algorithms

Afterwards, the optimized AL 3 was selected for the fitting phase. In the Figure 11 it can be seen that it was run enough trees to test the algorithm.

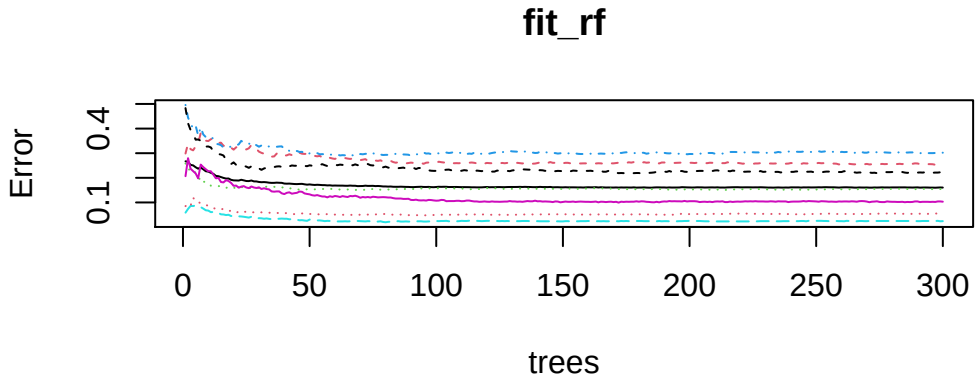


Figure 11: Evaluating Fitting algorithm 3

Finally, it was tested over the subsample test data, achieving an acceptable accuracy, and sensitivity-specificity for each class.

```
## Accuracy
## 0.839418
```

```
##          Sensitivity Specificity
## Class: 1   0.7446809   0.9641434
## Class: 2   0.8346883   0.9592218
## Class: 3   0.7120879   0.9747368
## Class: 4   0.9735577   0.9785969
## Class: 5   0.9046563   0.9795734
## Class: 6   0.8135198   0.9668756
## Class: 7   0.9333333   0.9863704
```

4.2 Results interpretability

To have an idea of which variable explained more for the algorithm, it was computed the variable importance (Table 8). As expected the variable Elevation explained by far the algorithm, nonetheless, the horizontal distance to fire points and to roadways are relevant too. This can highlight, that for future analysis to take a deeper look to those variables.

Table 8: Table Importance

	MeanDecreaseGini
Elevation	2,337.2065345
Horizontal_Distance_To_Fire_Points	1,120.2239796
Horizontal_Distance_To_Roadways	885.1647702
S_cli_8	695.4238438
Wilderness_Area_4	525.8160113
Horizontal_Distance_To_Hydrology	522.2549189
Hillshade_9am	511.4313015
Vertical_Distance_To_Hydrology	434.7938240
Wilderness_Area_1	434.3324279
Hillshade_3pm	426.8548482
S_cli_7	392.7404967
Hillshade_Noon	391.3576084
S_cli_2	304.3154971
Slope	291.8559269
Wilderness_Area_3	267.7400838
S_cli_4	235.8902189
S_geo_2	94.4594998
S_geo_7	93.8581193
S_geo_1	56.4118624
S_cli_6	53.5661188
Wilderness_Area_2	29.2127049
S_cli_5	3.7797304
S_cli_3	0.0090430
S_geo_5	0.0058636

5 Conclusions

Applying machine learning methods open a new way to predict phenomenon and or species distribution. The presented analysis illustrates how the random forest approach can identify tree species employing cartographic variables. In this term, this project dives into several techniques to get the final outcome. It was necessary to use several techniques acquired in the program Data Science certificate of HarvardX, such as statistical analysis, wrangling data and supervised machine learning. All these techniques and approaches are fundamental to providing a robust analysis method that can be useful for consultancy or research.

Finally, it is essential to remark that more variables and details could be included surpassing the cartographic ones. For example, samples of precipitation or wind area classification can help to have a deeper look at this analysis. Moreover, this analysis is still in its infancy since it can still add more parameters or apply other methodologies such as k-means or Singular Value Decomposition approach. The world of machine learning is broad and continuous, extending its frontiers. As a result, it can improve the presented project much more and provide an algorithm with accuracy levels reaching close to perfection.

References

B. Boehmke, B. G. (2020). Hands-on machine learning with R. https://bradleyboehmke.github.io/HOML/.

Bevans, R. (2020). An introduction to the one-way anova. https://www.scribbr.com/statistics/one-way-anova/.

Blackard, J., Dean, D., and C., A. (1998). Covertype data set. https://archive.ics.uci.edu/ml/datasets/Covertype.

Blauw, L. G., van Logtestijn, R. S., Broekman, R., Aerts, R., and Cornelissen, J. H. C. (2017). Tree species identity in high-latitude forests determines fire spread through fuel ladders from branches to soil and vice versa. *Forest Ecology and Management*, 400:475–484.

D. Fernandes Terra Machado, S. H. Godinho Silva, N. C. (2019). Soil type spatial prediction from random forest. *Scientia Agricola*.

Irizarry, R. (2021). Introduction to data science: Data analysis and prediction algorithms with R. https://rafalab.github.io/dsbook/.

Kuhn, M. (2019). The caret package. https://topepo.github.io/caret/index.html.

T. Hastie, R. Tibshirani, J. F. (2008). *The elements of Statistical learning. Data mining, Inference, and Prediction*. Springer. https://web.stanford.edu/~hastie/Papers/ESLII.pdf.