Machine Learning analysis with the ERA Database

A) Summarize and subset the Dataset

The most recent version of the ERA dataset contains **107952** instances and **142** attributes. I will carry out the experiment on this whole data but for now, let's focus on the LER outcomes. The new subset has **808** instances and **142** attributes. For this case, we will focalize on the most important attributes (factors) leading to this. Let's first see the type of attributes we have:

Type of Attributes

We can have some examples from the first 6 attributes listed below

```
    Index        Code      Author        Date     Journal         DOI
"integer" "character" "character"   "integer" "character" "character"
```

The 3 types of attributes we have here are:

```
"integer"   "character" "numeric"
```

The 8 most important factors are

-MeanC

-MeanT

-Yield

-Mean annual precipitation (MAT)

-Total seasonal precipitation (TSP)

-Soil organic carbon at start of experiment (SOC)

-Soil pH

-Duration

First, our target here will be the product we get out of this. In total, we have 59 type of products and this is a **multi-class or a multinomial classification problem.**

B) Create a Validation Dataset

We need to know that the model we created is any good.

Later, we will use statistical methods to estimate the accuracy of the models that we create on unseen data. We will split the loaded dataset into two, **80%** of which we will use to train our models and **20%** that we will hold back as a validation dataset.

This doesn't seem to be interesting since some classes have a single record. Let's carry out our analysis in the other direction and target the possible outcomes. In this case let's target the full dataset with the attributes listed above. We get 86 types of outcomes.

A view of the class distribution gives us the following where we have their frequencies (number of instances of each class) and percentage

	freq	percentage
Aboveground Biomass	32	0.029642804
Animal Mortality	89	0.082444049
Animal Survival	186	0.172298799
Belowground Biomass	24	0.022232103
Beneficial Organisms	675	0.625277901
Benefit Cost Ratio (GRTC)	233	0.215836668
Benefit Cost Ratio (GRVC)	242	0.224173707
Benefit Cost Ratio (NRTC)	59	0.054653920
Benefit Cost Ratio (NRVC)	105	0.097265451

C) Statistical Summary

We can take a look at a summary of each attribute. We can also see the amount of each NA value

```
     MAT                TSP              SOC              Soil.pH           yi                MeanC              MeanT
 Length:107952      Min.   :   0.0   Min.   :    0.00   Min.   : 3.20   Min.   :  -Inf   Min.   : -270380   Min.   :   -46415
 Class :character   1st Qu.: 321.4   1st Qu.:    0.74   1st Qu.: 4.90   1st Qu.:-0.069   1st Qu.:       1   1st Qu.:        1
 Mode  :character   Median : 472.0   Median :    1.70   Median : 5.60   Median : 0.141   Median :       3   Median :        4
                    Mean   : 548.0   Mean   :    8.45   Mean   : 5.67   Mean   :   NaN   Mean   :    2825   Mean   :     4751
                    3rd Qu.: 736.0   3rd Qu.:    6.20   3rd Qu.: 6.20   3rd Qu.: 0.483   3rd Qu.:      25   3rd Qu.:       27
                    Max.   :1935.0   Max.   :21200.00   Max.   :57.50   Max.   :  Inf    Max.   :16380000   Max.   :139713846
                    NA's   :96950    NA's   :72163      NA's   :59338   NA's   :3191     NA's   :2004
    Duration       Out.SubInd
 Min.   : 0.000   Length:107952
 1st Qu.: 1.000   Class :character
 Median : 1.000   Mode  :character
 Mean   : 2.437
 3rd Qu.: 3.000
 Max.   :78.000
 NA's   :3358
```

D) Visualize Dataset

We now have a basic idea about the data. We need to extend that with some visualizations.

We are going to look at two types of plots:

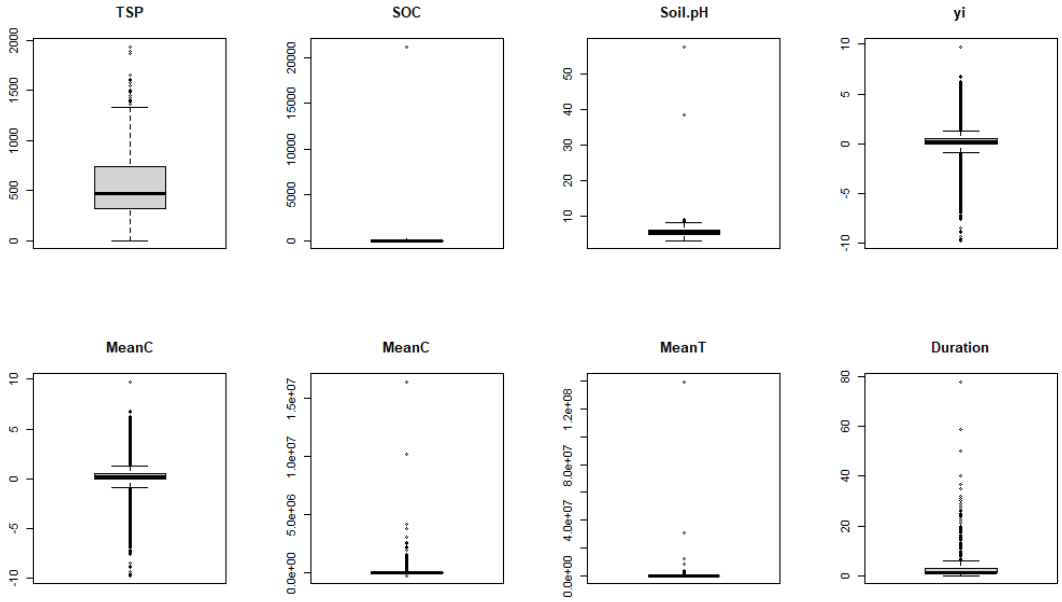
Univariate plots to better understand each attribute.
Multivariate plots to better understand the relationships between attributes.

Univariate Plots

We start with some univariate plots, that is, plots of each individual variable.

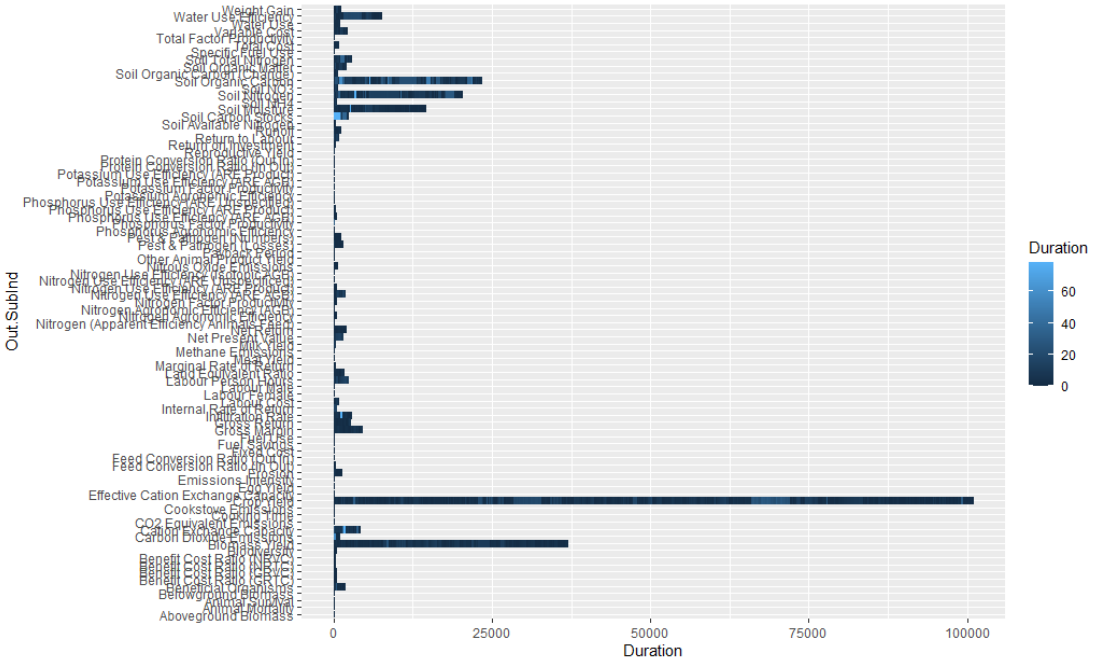
It is helpful with visualization to have a way to refer to just the input attributes and just the output attributes. Let's set that up and call the inputs attributes x and the output attribute (or class) y.

This gives us a much clearer idea of the distribution of the input attributes:



Box and Whisker Plots in R

We can also create a barplot of the outcomes class variable



Multivariate Plots

Now we can look at the interactions between the variables.

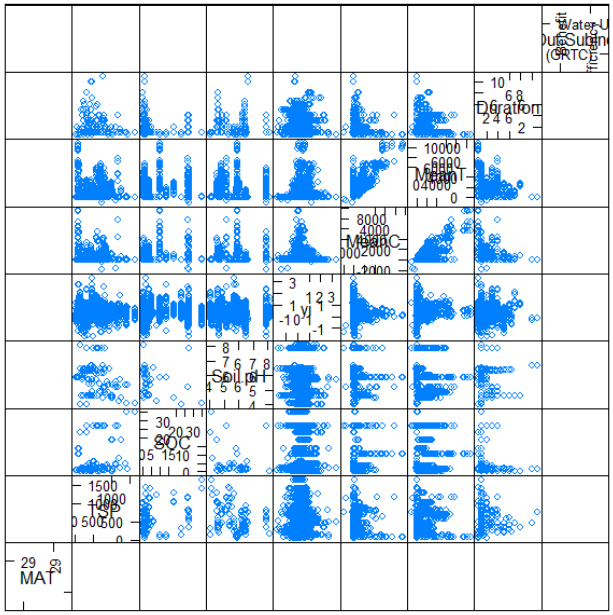
Let's first study the missing data or null values

Our data contain a lot of missing values and NA values. The column with missing data are

"MAT" "TSP" "SOC" "Soil.pH" "yi" "MeanC" "Duration"

The new dataset contains **2366** rows compared to **107952** with the original dataset.

First let's look at **scatterplots** of all pairs of attributes and color the points by class. In addition, because the scatterplots show that points for each class are generally separate, we can draw ellipses around them.



Scatter Plot Matrix

E) Evaluate Some Algorithms

Let's create some models of the data and estimate their accuracy on unseen data.

Here is what we are going to cover in this step:

1. Set-up the test harness to use 10-fold cross validation.
2. Build 5 different models to predict outcomes from these agroforestry attributes.
3. Select the best model.

We will use the cleaned data with 2366 instances and 9 attributes

1 Test

We will use 10-fold cross validation to estimate accuracy.

This will split our dataset into 10 parts, train in 9 and test on 1 and release for all combinations of train-test splits. We will also repeat the process 3 times for each algorithm with different splits of the data into 10 groups, in an effort to get a more accurate estimate.

2 Build Model

Here we will evaluate 5 different algorithms:

- Linear Discriminant Analysis (LDA)
- Classification and Regression Trees (CART).
- K-Nearest Neighbors (kNN).
- Support Vector Machines (SVM) with a linear kernel.
- Random Forest (RF)

Let's build our five models

```
# a) linear algorithms
set.seed(7)
fit.lda <- train(Out.SubInd~., data=Mdat_drop, method="lda", metric=metric, trControl=control)
# b) nonlinear algorithms
# CART
set.seed(7)
fit.cart <- train(Out.SubInd~., data=Mdat_drop, method="rpart", metric=metric, trControl=control)
# kNN
set.seed(7)
fit.knn <- train(Out.SubInd~., data=Mdat_drop, method="knn", metric=metric, trControl=control)
# c) advanced algorithms
# SVM
set.seed(7)
fit.svm <- train(Out.SubInd~., data=Mdat_drop, method="svmRadial", metric=metric, trControl=control)
# Random Forest
set.seed(7)
fit.rf <- train(Out.SubInd~., data=Mdat_drop, method="rf", metric=metric, trControl=control)
```

3 Select Best Model

Unfortunately we can't compare because only one model works with our data Classification and Regression Trees (**CART**). We will have to figure this out later.

Here is the summary of the model and we can see the model doesn't perform well

```
> fit.cart
CART

2366 samples
   8 predictor
  25 classes: 'Benefit Cost Ratio (GRTC)', 'Benefit Cost Ratio (GRVC)', 'Benefit Cost Ra
tio (NRVC)', 'Biomass Yield', 'Carbon Dioxide Emissions', 'Cation Exchange Capacity', 'C
O2 Equivalent Emissions', 'Crop Yield', 'Erosion', 'Gross Margin', 'Gross Return', 'Labo
ur Cost', 'Land Equivalent Ratio', 'Methane Emissions', 'Net Return', 'Nitrogen Use Effi
ciency (ARE AGB)', 'Nitrous Oxide Emissions', 'Soil Moisture', 'Soil Nitrogen', 'Soil Or
ganic Carbon', 'Soil Total Nitrogen', 'Total Cost', 'Variable Cost', 'Water Use', 'Water
 Use Efficiency'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2129, 2128, 2130, 2128, 2133, 2130, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.02271114  0.5228390  0.3166095
  0.02555004  0.4991828  0.2666937
  0.04802460  0.4435037  0.1147268

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.02271114.
```

F) Make Predictions

Now we want to get an idea of the accuracy of the model on our validation set.

We can run the Classification and Regression Trees model directly on the validation set and summarize the results in a **confusion matrix**

```
Overall Statistics

               Accuracy : 0.5129
                 95% CI : (0.4664, 0.5593)
    No Information Rate : 0.4116
    P-Value [Acc > NIR] : 6.781e-06

                  Kappa : 0.2711

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Benefit Cost Ratio (GRTC) Class: Benefit Cost Ratio (GRVC) Class: Benefit Cost Ratio (NRVC)
Sensitivity                              0.000000                          0.00000                          0.00000
Specificity                              1.000000                          1.00000                          1.00000
Pos Pred Value                                NaN                              NaN                              NaN
Neg Pred Value                           0.997845                          0.98276                          0.99569
Prevalence                               0.002155                          0.01724                          0.00431
Detection Rate                           0.000000                          0.00000                          0.00000
Detection Prevalence                     0.000000                          0.00000                          0.00000
Balanced Accuracy                        0.500000                          0.50000                          0.50000
```

G) Let's carry out the experiment without dropping the NA values