## Introduction

In this project, we were tasked with using nonlinear methods for regression and classification. The data that we used contains variables related to Algerian Forest Fires, including temperature, relative humidity, wind speed, millimeters of rain, Fine Fuel Moisture Code Index, Duff Moisture Code Index, Drought Code Index, Initial Spread Index, Fire Weather Index, month, year, region, and whether or not there was a fire.

The prediction problem that we solved with *regression* was to predict rainfall, in millimeters, in certain regions of Algeria. This was done using Decision Trees and Random Forests. The prediction problem that we solved with *classification* was to predict whether or not certain regions in Algeria would have a forest fire. This was also done using Decision Trees and Random Forests. In this report, we will discuss the methods used for our analysis, as well as our results and conclusions.

## Methods

**Commented [1]:** Summarize the models used and parameter tuning approach

Before using any methods, we prepared the data by implementing data splitting and data standardization. First, we split the data into 75% training data and 25% testing data by random sampling. Next, we standardized the data by using the preProcess() function from the caret package with the method = c("center", "scale") option. This procedure calculated the mean and standard deviation of each predictor in the training set, subtracted the mean, and divided by the standard deviation so that each numeric variable had a mean of zero and a standard deviation of one.

The first regression model entailed using a decision tree to predict the amount of rainfall that would occur given other explanatory variables. To perform this analysis, we used the tree function from the tree package in r and then used the decision tree to predict rainfall. The parameter tuned was the maxdepth parameter, where it was determined that the ideal value for this parameter was one. This indicates that the decision tree will work best when there is only one split and two leaves. After doing this, we evaluated the performance of the model by using the mean squared error, mean absolute error, and the r-squared value.

For the second regression model, we decided to use random forests. We set the seed to 4630 and use the randomForest function on the training data. The model predicted our response variable which was rainfall in millimeters. The parameter used included an ntree = 500, and the mtry value chosen was about 1/3rd of the predictor values. After training the model, the predictions were generated towards the testing data. Then the performance of the model was evaluated using the regression metrics of MSE, MAE and R^2 which were computed by comparing the predicted rainfall values to the real rainfall measurements.

For the first classification model, we used decision trees to predict whether or not there would be a forest fire. First, a decision tree was created using the tree function with the standardized training data in order to predict the variable “Classes”, which represents whether or not there will be a forest fire. After creating this original decision tree, we used cross validation to tune the decision tree. First, the cv.tree function was used to find the optimal tree size in order to balance complexity and accuracy. Appendix 3 shows the output of this cross validation, showing that the optimal number of leaves is 3. After finding the optimal tree size, the tree was pruned using the optimal size found through cross validation. The final decision tree is displayed in Appendix 3.

For the second classification model, we used random forests. The code trains a Random Forest model for binary classification using cross-validation. First, it ensures that the target variable (Classes) in both the training and test datasets is treated as a factor with two levels, "0"

and "1." A cross-validation method is defined using trainControl, and a hyperparameter tuning grid (tuneGrid) is specified for the mtry parameter (determined that mytry=2 was optimal). The model is trained on the standardized training dataset, and predictions are made for both class labels and probabilities on the test dataset. The model's performance is evaluated using a confusion matrix, metrics such as accuracy, precision, recall, and F1-score, and the area under the ROC curve (AUC) to summarize its classification capability.

## Results

Commented [2]: Highlight key findings, comparing linear and nonlinear methods

In order to analyze our results, we calculated performance metrics for each method. The decision tree model with one node involved predicting .100 mm of rain if the drought code index was above 9.9 and predicting 3.187 mm of rain if the drought code was below that value. For the decision tree component of our regression analysis, the mean squared error was 1.598. This result indicates that the squared error between our predicted and actual values for rainfall was very large. The mean absolute error was .73, and the fact that this is significantly lower than our mean squared error means there are some significant outliers that are pushing the MSE up. Nonetheless, this indicates that rainfall estimates are, on average, off by .73 millimeters. Finally the R-squared of this method was .357, indicating that 35.7% of the variance of the test data is explained by this model, a disappointing result.

For the Random Forest regression performance metrics, we received a Mean Squared Error of 0.01962642 meaning that the predicted rainfall values are close to accurate measurements and that on average the squared deviations between predicted and real rainfall are relatively small. Next, the Mean Absolute Error of 0.03208470 indicated to us that our models rainfall predictions differ from the actual rainfall values by about 0.032 millimeters which is a minor deviation with all things considered. Next, when plotting the final random forest model, FFMC, ISI and FWI stand out as the top variables for prediction as can be seen in the feature importance plot in appendix 2. Lastly, we received an R-Squared score of 0.92147323 which means that more than 91% of the variation in rainfall is explained by many different variables in our model which makes it strong in predicting rainfall levels in the forested areas of Algeria.

For decision trees for classification, we calculated Accuracy (0.9672131), Precision (1), Recall (0.9333333), F1-Score (0.9655172), and ROC-AUC (0.9634409). These values are also displayed in Appendix 3. These values show that the model performed well because all of the values are greater than 0.9. Additionally, we created a ROC Curve for the pruned decision tree, which is also displayed in Appendix 3. This ROC Curve further confirms the good performance of the model because of the shape of the curve.

For Random Forests for classification, there was a reported accuracy of 0.9508197, precision of 0.9375, recall of 0.9677419, F1-Score of 0.952381, and a ROC-AUC of 0.9956989. Based on the above values being all close to 1, the model performed well. These values are displayed in Appendix 4 along with the ROC curve which further confirms the successful performance of the model.

Of the non-linear methods we employed, the decision tree for the regression model was by far the weakest. We suspect this is because a decision tree is not suited for this particular regression task. We think it is quite difficult for a decision tree to predict rainfall outcomes given the other data points in this dataset. Rather a decision tree could better be used in a classification task to predict whether or not a fire occurred, where it fared much better than the regression model. Other non-linear models predicted the test data with high accuracy for both regression and classification. While this decision tree did not lead to great results, it did slightly outperform the linear regression model in part 1 in terms of mean squared error and r-squared values,

indicating that it is a better prediction model for this data. However, the random forest model is clearly the best at predicting the amount of rainfall based on these data points.

In project part 1, only 32% of the variance of rainfall was explained in our linear regression model. That value significantly increased to 92% when using a non-linear model of random forest. Additionally, for classification, our linear discriminant analysis had a 90% accuracy whereas the non-linear model increased its accuracy to 95% with random forests. This jump may not seem significant, but when that 5% is the difference of predicting 10-15 forest fires, small percentage differences matter. These improvements in performance of the non-linear models is expected due to their ability to capture more complex and non-linear relationships in the data. With more complicated data sets, non-linear methods will tend to perform better.

## Conclusion

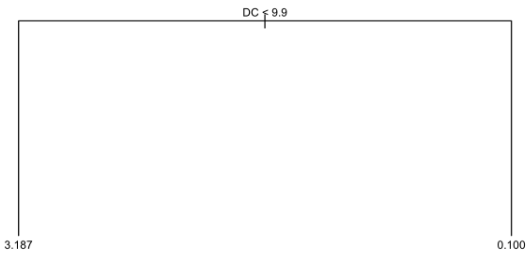
Overall, our results when performing regression and classification analysis demonstrated to us that certain methods performed considerably better than others.

In regards to regression, with the task of predicting rainfall, the Random Forest model outperformed the Decision tree model. With an R^2 value of 0.91 and lower MSE and MAE values, the Random Forest provided accurate predictions and explained the high proportion of the variation in rainfall.

For classification, both models performed well with the accuracy, precision, recall, F1-score, and ROC-AUC values all being greater than 0.9 in both models.

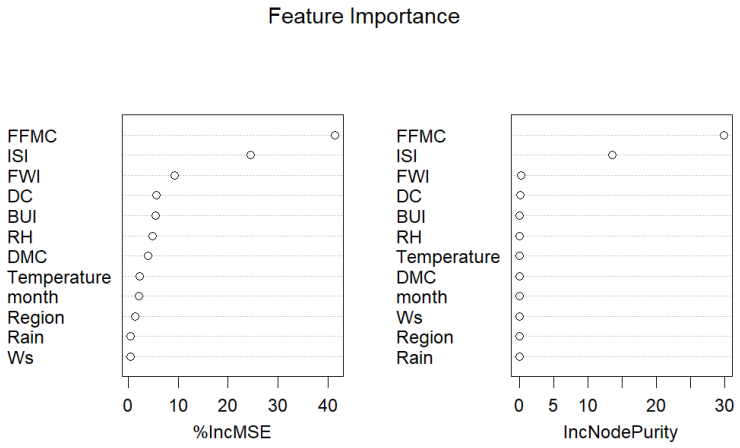
**Commented [3]:** State which methods performed better and under what conditions

**Appendix 1: Plots for Decision Trees (Regression)**



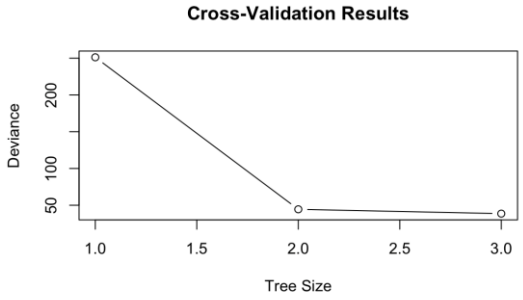

| | |
|---|---|
| Mean Squared Error | 1.5982398 |
| Mean Absolute Error | 0.7348045 |
| R-Squared | 0.3579862 |

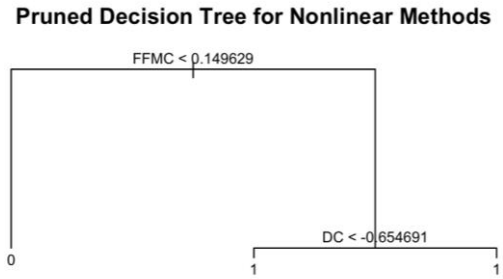
**Appendix 2: Plots for Random Forest (Regression)**



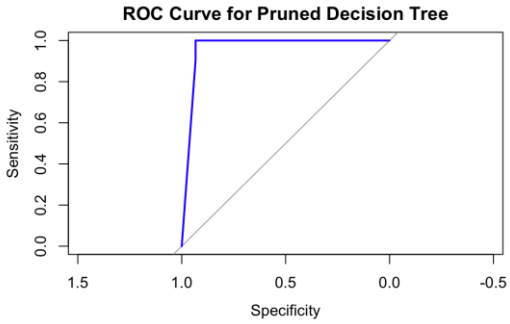

| | |
|---|---|
| Mean Squared Error | 0.01962642 |
| Mean Absolute Error | 0.03208470 |
| R-Squared | 0.92147323 |

**Appendix 3: Plots for Decision Trees (Classification)**

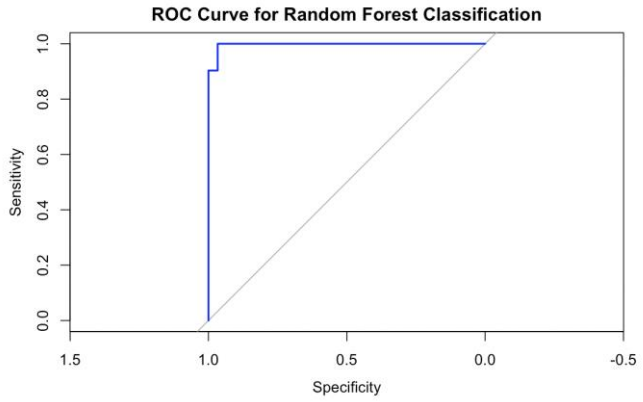



Accuracy: 0.9672131
Precision: 1
Recall: 0.9333333
F1-Score: 0.9655172
ROC-AUC: 0.9634409



**Appendix 4: Plots for Random Forest (Classification)**




```
Accuracy: 0.9672131
Precision: 0.9393939
Recall: 1
F1-score: 0.96875
Setting levels: control = 0, case = 1
Setting direction: controls < cases
ROC-AUC: 0.9967742
```

**Appendix 5: Individual Contributions**

Introduction - Sammy and Tanner
Data Splitting - Tanner
Data Standardization - Alex
Regression Analysis (Decision Tree) - Aidan
Regression Analysis (Random Forest) - Alex
Classification Analysis (Decision Tree) - Tanner
Classification Analysis (Random Forest) - Sammy
Comparison of Linear + Non-Linear Models - Aidan and Sammy
Conclusion - Alex and Tanner