Predictive Model of Spatial Distribution of Forest Fires Driving Factors: A Case Study in Portugal, Spain

Report Info:

Brendan B. Pham
University of Minnesota - Twin Cities
Created 11/04/2023

Keywords: Wildfire Feature Selection
Neural Networks

ABSTRACT

The ability to predict the spatial distribution of Forest Fires is an important tool for forest fire management services that could prove helpful in preventing the volatility of forest fires in the United States. Wildfire poses an escalating threat to ecosystems, resources, and communities in high-risk-prone areas. An analysis was done on Portugal Monteshino's wildfire events, to simulate the meteorological and temporal features of Colorado.

A comprehensive exploration of Artificial Neural Networks and Random Forest regressive machine learning models, to primarily use in the wildfire forest protection services toolbox. A fast, interoperable, and accurate model is essential for early detection and response of wild Fire burns. In our comparison, the fastest, most accurate, and most interoperability model selected was the Random Forest Regressive model

To determine the feature deterministic importance of volatile wildfires, we looked at groups of features ranging from meteorological, temporal, geographical, and calendar features. Performed factor interpretation and recommendations by partial dependence plot analysis and feature importance. It shows that temporal features prove more valuable in determining the area of wildfires than meteorological ones. To prove this point, using the model, there were simulations run of Colorado temperature and meteorological features.

1 INTRODUCTION

Forest Fires pose multifaceted threats to ecosystems, communities, and resources. The challenges of managing wildfires compound from the volatility of forest fires, forest fire services conducively state that wildfires are an unpredictable force. A forest fire phenomenon can get out of control easily, the initial spread can grow to a state that is no longer controllable simply by the sheer size and violence of a forest fire. US forest policy initiated U.S. After 1910 August a terrifying firestorm set Yellowstone Park ablaze destroying three million acres of land and leaving brimstone, 90 deaths, and Towns Destroyed (Diaz, H. F., & Swetnam, T. W.,2013). Forest policy since then has evolved into quick and effective reactionary responses, but today wildfires are getting increasingly more catastrophic due to climate changes and increasingly large dense forests (Flannigan & Stocks & Wotton, 2000)

Although forest fires have seen a larger occurrence, forest fires have immediate impact and indirect impacts. An indirect impact of Forest fires while still being researched could have direct impacts on human lives (Fann, 2018). Direct forest fire impacts are not restricted only to the area but could impact air pollution and soil contamination (Kala, 2023). Overall forest fires have impacts on human lives, land destruction, air quality, and soil contamination

Machine learning offers an accurate, detectable, scalable, and highly beneficial approach to wildfire management teams. These tools can interpret vast amounts of data, analyze complex patterns, and provide forest fire management teams with rapid decisions based on the tool's prediction and interpretation of feature impacts and signs of differing levels of volatile forest fires. Furthermore, the integration of Machine Learning tools can enhance the speed and accuracy of decision making which could prove valuable in enhancing the firefighter's emergency response efforts. Utilization can improve resilience and preparedness in the face of the everchanging unpredictable force of forest fires.

Development of models in this research, we have decided to utilize Neural Networks and a Random Forest model both for their ability to interpret complex relationships of the the data, feature learnings, and predictive power for this particular data set. Below is the solution map of the research model to come to our discussions about the dataset.

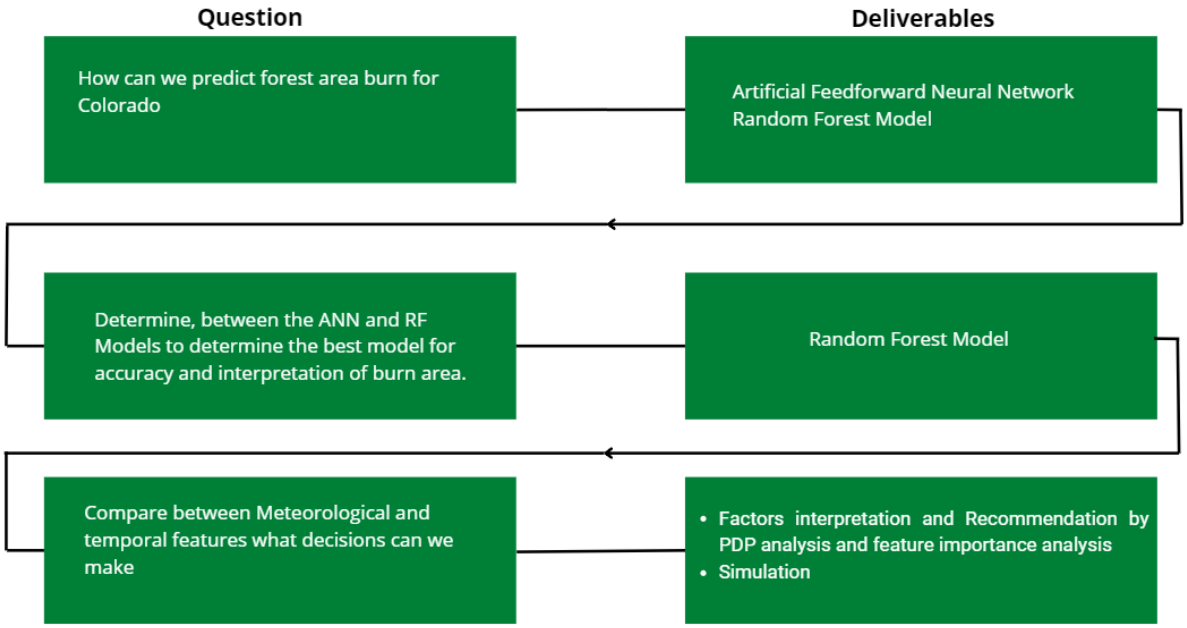


Figure 1, (Solution map)

To employ the machine learning method we have chosen an Artificial Neural Network and a Random Forest model to best predict Portugal's Montesino Area burn events. After finding the best model for the use case, we want to compare Meteorgical and Temporal Features to determine in a forest fire management incident which features are significantly important to the area of forest fires, utilizing factors interpretation and recommendations by partial Dependence plots, features importance analysis, and a simulation of the Colorado's Forest Fires.

2 METHODS

The data set was obtained from the Center for Machine Learning Repository at the University of California, Irvine. The dataset has 517 burn events and 13 variables that were collected from January 2000 to December 2003. The data was collected in the Northeast region of Portugal of Montesinho National Park. Data aims to predict the area of Forest Fires.

Forest Fire Table

Variable code	Description	Unit	Min	P25	P50	Mean	P75	Max
X	East-West Coordinate (1 to 9)	Integer series	1	3	4	4.67	7	9
Y	North-South Coordinate (1 to 9)	Integer series	2	4	4	4.30	5	9
Month	(1 to 12)	Integer series	1	7	8	7.48	9	12
Day	(1 to 7)	Integer series	1	2	5	4.26	6	7
FFMC	Fine Fuel Moisture Code	Index	18.7	90.2	91.6	90.64	92.9	96.2
DMC	Duff Moisture Code	Index	1.1	68.6	108.3	110.87	142.4	291.3
DC	Drought Code	Index	7.9	437.7	664.2	547.94	713.9	860.6
ISI	Initial Spread Index	Index	0.0	6.5	8.4	9.02	10.8	56.1
Temp	Temperature	°C	2.2	15.5	19.3	18.89	22.8	33.3
RH	Relative Humidity	%	15	33	42	44.29	53	100
Wind	Wind Speed	kmh^{-1}	0.4	2.7	4	4.02	4.9	9.4
Rain	Rainfall	mm	0	0	0	0.02	0	6.4
BA	Total Burned Area	Ha	0	0	1.52	3.04	7.57	1091.8
BA (Ln)	Natural Logarithm of BA	Ln(Ha)	0	0	0.42	1.11	2.02	7.00

Note: A burned area of zero refers to any event for which the BA is less 100 m^2.

(Figure 2. Wood, D, 2021)

Discussions:

Some notes about the data set for interpretation X and Y are geographical locations for Montesihno Park. Month and Day are calendar features. FFMC, DMC, DC, and ISI are Fire Weather Indices (FWI) otherwise known as meteorological features. Temperature, RH, Wind, and Rain are temporal features.

Data Preparation:

After looking at the distribution of burn area we can see that there is a large right skew distribution of zero burn areas for the dataset, in the interpretation of the model when training, there will be biases towards lower burn areas. Whereas a burn area in the area 0 represents a bun that is less than $100m^2$, to deal with this we set a log(BA +1). We can see there is a bias in area 0, leading to the model not having enough data to predict for areas larger than 40 hectares.

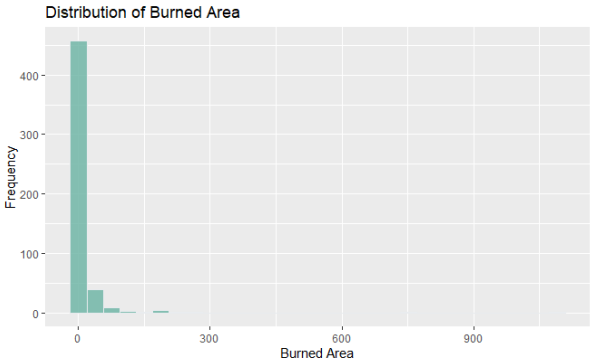


Figure 3. Distribution of burned area

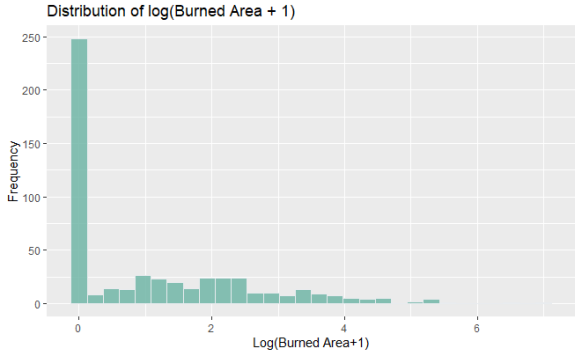


Figure 4. Distribution of log(burned area+1)

Model Comparison

To compare regression models we look at multiple metrics to compare the accuracy of models

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

Mean absolute error is defined as the average of the absolute error values, between predicted and actual values. A lower MAE indicates a model with higher precision and accuracy. Where n is the number of samples and the absolute is the actual minus the predicted value of the sample Schneider, P., & Xhafa, F. (2022)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Root Mean Square Error is the mean of the sums of squares of observations of the sample. Where it is the sum of squares of the average predicted observations minus actual observations. (Patil, A., Soni, G., & Prakash, A. 2022)

3 METHOD

Random Forest Regression is a statistical ensemble learning machine learning technique that employs the usages of multiple decision trees, decision trees, that use simple decision trees to infer the importance of each data feature for the final model. The random forest model randomly samples features with replacements for multiple decision trees, all decision trees are averaged across all predictions to produce a final prediction. The procedure of averaging decision tree results reduces overfitting. Therefore random forest can handle large datasets with complex relationships because of the averaging of random decision trees. The decision trees can be interpreted as a node otherwise parent leaves while its child leaves is a decision made upon the node. The random forest model: uses a mean square error to measure the accuracy of each tree prediction for each node

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

One note of each tree, to reduce overfitting and minimize variance, bootstrapping is employed which randomly selects N samples of replacement from the dataset. Once all decision trees are grown, predictions of the tree are made for the out-of-bag sample prediction. Random forest models are less prone to overfitting models based on random sampling and have interoperable applications of the feature importance of the dataset (Gao, C., Lin, H., & Hu, H. 2023)

Neural Networks

Artificial Neural Networks is a deep learning method that is inspired by the structure and functioning of the human brain. During the data flow, it consists of multiple terms, it consists of interconnected notes otherwise neurons, and organized into layers. The first layer is the input layer typically features of the dataset, the second layer is the hidden layer which is how the artificial neural network propagates information to each neuron to function and pick an important

neuron for the next layer. The final layer is the output layer, in the case of a regression task it consists of one neuron the final output. Neural Networks are a fundamental component of deep learning models, a subfield of machine learning models. Deep learning involves training deep neural networks to automatically learn the data flow representation (Kartalopoulos, 1997).

To propagate through each layer, a forward propagation is utilized the measure the weights and biases of each neuron. The weight and biases represent the strength of the connection between each model's neurons, based on the threshold that is selected by the model it will choose which neuron connection is strongest for the interpretation of the next layer of neurons. The neurons are propagated through an output function and a sigmoid function.

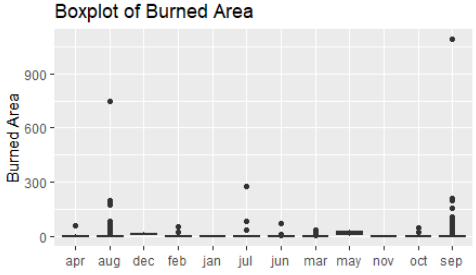
$$Y = W_1 x_1 + w_2 x_2 + w_n x_n + b \quad \text{Activation } \sigma(z) = \frac{1}{1+e^{-z}}$$

The output function can best be represented by the weights, connections, and biases used to measure the strength of the connection of each neuron. Based on the model, a threshold determines the importance of each strength. To receive the final output a linear activation function is utilized to get the output. Neural networks have impressive machine learning capabilities in understanding complex data relationships, specifically for forest fires to understand relationships of data features (Han, H., 2018)

4 Initializing Models

To compare models, utilized a combination of meteorological models, temporal, and spatial. Switching between meteorological and temporal features for the model. After looking at Figure 5 we can see that there is a seasonality difference between winter, spring, fall, and summer. Data engineer a new variable season that is numeric and calendar dates to numeric

Figure 5. boxplot



4 Random Forest Model

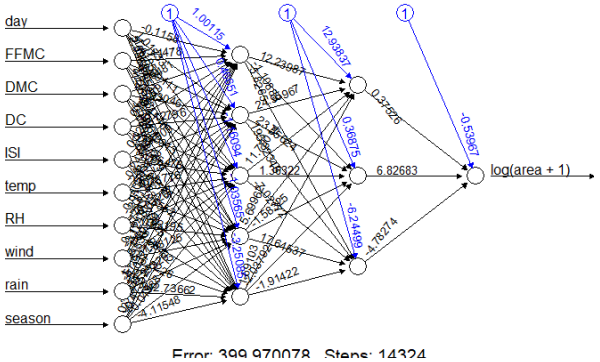
Employing a cross-validation technique which is a resampling technique to train and test on a subset of data that uses multiple splits to average the result to reduce oversampling we use 5 folds and 5 repeats. For ntree we set it to 500, to prune the model we would grid search against multiple ntrees but in this case, we want to set 500 to a ntree parametrization. To find the number of variables tried at each split we employ a grid search of the features from 2 to 12 features (Ramadhan, 2017). In model searching it is best to use $mtry = \sqrt{p}$ where p is the number of features. In the case of cross-validation, we can all randomly select predictors.

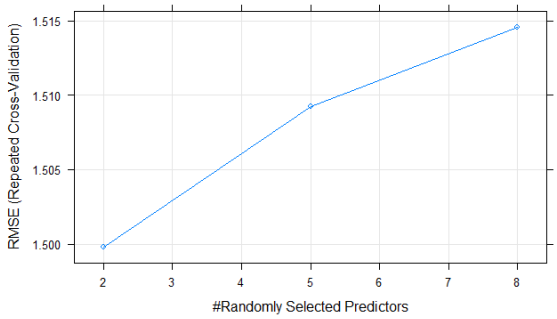
5 Neural Network

Comparing between features, we can change the parameters for neurons, layers, stepmax, and learning rate. Cross-validation wasn't utilized for the model due to the limitations of the model not being able to converge. The rule of thumb for parameter setting for neurons, for a higher complex dataset we set neurons above 5, and for a higher complex dataset we set layers to 3. Typically for layers, if the task was a classification situation we would set layers to 3, in this case, we want to train and test against the layers of 2 and 3. The step max is the maximum step for the model to reach the final output, the ANN's prediction is independent of the step max, for ease we set it to 2e05. The issue with model convergence is the amount of step max, to reduce issues with the step max we could set it higher, but there are limitations of time, as setting a higher step max leads to 30+ minutes of computation. The last parameter was the learning rate, where the learning rate determines the size of each step through each connection of neurons, we set it to 0.01, typically the result of the dataset's complexity can be set between 0.001 and 0.01 (Koutsoukas, 2017).

7 Results

To compare regression models we use different models of the meteorological and temporal features removing X and Y for the case study of Colorado as it relates to Montesino Park, and removing rain as indicated later with feature importance. Below are the plots of NN and RF.





Model Metrics *Figure 6, Model Metrics*

Model	Percent Variance Explained	MAE	RMSE
ANN Default	0.004	19.1065	1.5078
ANN: FFMC + DMC + DC + ISI + season	-0.014	7.4201	1.2850
ANN: season + temp + RH + wind	-0.098	7.43372	1.3372
RF Default	-0.159	8.0256	1.3728
RF: FFMC + DMC + DC + ISI + season	-0.088	19.5869	1.5299
RF: season + temp + RH + wind	-0.200	19.6341	1.6064

We can see that the ANN default and Meteorgical Random forest default had the highest predictive power, however due to interpretation and computation we will select Random Forest default discussed in the insights and discussions section

6 Variable Importance & Simulation

To feature select and remove rain, utilize a default random forest model to calculate the feature importance of the random forest model. The average of the decision trees for splits, the InCNodePurity increases if a split of the decision tree is purer than the parent node, so we can determine the feature's importance in the prediction

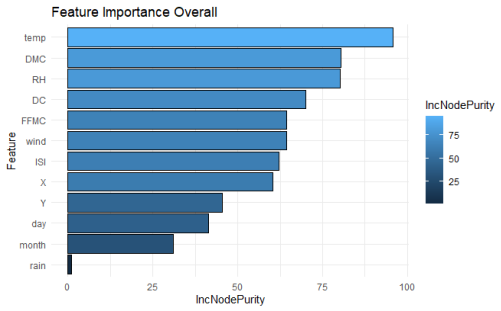


Figure 7, Feature Importance Overall

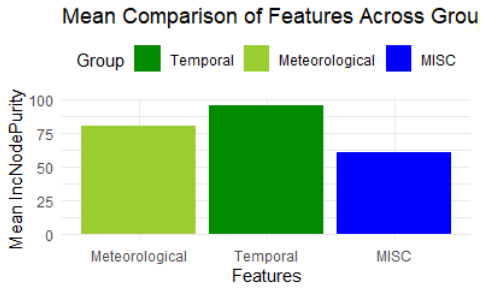


Figure 8, Mean Comparison of Features

To simulate the differences between temperature and FWI (meteorological features), simulating two scenarios with n = 1000 observations, while keeping the temperature constant to Coloardos average temperature of seasons, and insights of changing FWI between simulations

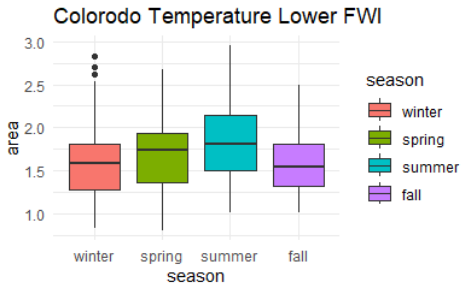


Figure 9, Feature Importance Overall

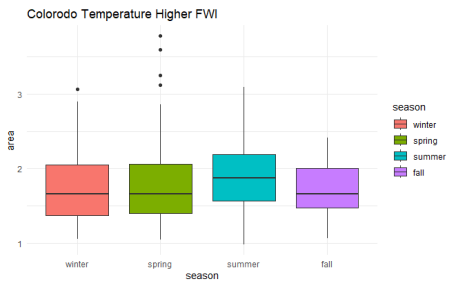


Figure 8, Mean Comparison of Features

From the mean comparison and simulations, we can see that temporal features and seasons still have a greater impact on the area of the forest fire burns as seen between figure 9 and figure 10.

5 DISCUSSION AND SUMMARY

Addressing the research question in hand, what model is both fast, flexible, and accurate for predicting the fire area burn? By employing the Artificial Neural Network and Random Forest model we can see that ANN exhibits a higher accuracy and predictive power, ANN models are highly adaptive to the model. The issue with the ANN model is the replication and convergence of the data, due to the complexity or quality of the dataset there were issues with the convergence to replicate it and cross-validate to get a better prediction. The solution was to set a higher stepmax, the issue with setting a higher stepmax is that it was computationally large, taking upwards of an hour of computation. For both interpretation and computationally, Random Forest proved accurate and fast, as Random Forest was able to interpret the IncNodePurity between meteorological and temporal.

To check on the importance of features we did factor interpretation, analysis, and simulations to check the difference between meteorological and temporal features. Results show that features like Temperature and DMC show a higher importance overall. The mean comparison between temporal and meteorological show that on average, temporal features proves more useful in the partial dependence of the prediction of fire area burn also suggested by the partial dependence analysis. We can determine the in importance of time and even in the case of Colardo's Forest Fire Incident Management Team we should emphasize the importance of temporal features rather than meteorological features in the prediction of the area of forest fires.

For Colardo's Forest Fire Incident Management Team, I would suggest researching the X and Y coordinates relationship to forest fire burn area and classification methods of Hectare of burn area, as Portugal isn't the best representation of Colorado geographical features.

References

- Gao, C., Lin, H., & Hu, H. (2023). Forest-Fire-Risk Prediction Based on Random Forest and Backpropagation Neural Network of Heihe Area in Heilongjiang Province, China. Forests, 14(2), 170. MDPI AG. Retrieved from http://dx.doi.org/10.3390/f14020170
- Han, S. H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial Neural Network: Understanding the Basic Concepts without Mathematics. Dementia and neurocognitive disorders, 17(3), 83–89. https://doi.org/10.12779/dnd.2018.17.3.83
- Kartalopoulos, S. V., & Kartakapoulos, S. V. (1997). Understanding neural networks and fuzzy logic: basic concepts and applications. Wiley-IEEE Press.
- Wood,D (2021). Prediction and Data Mining of burned areas of forest fires: optimized data matching and mining algorithm provides valuable insights, Science Direct,https://www.sciencedirect.com/science/article/pii/S2589721721000118
- Cortez,Paulo and Morais,Anbal. (2008). Forest Fires. UCI Machine Learning Repository. https://doi.org/10.24432/C5D88D.
- Diaz, H. F., & Swetnam, T. W. (2013). The Wildfires of 1910: Climatology of an Extreme Early Twentieth-Century Event and Comparison with More Recent Extremes. Bulletin of the American Meteorological Society, 94(9), 1361-1370. https://doi.org/10.1175/BAMS-D-12-00150.1
- U.S. Department of Agriculture. (2023). Resource Management. Tahoe National Forest. https://www.fs.usda.gov/detail/tahoe/landmanagement/resourcemanagement/?cid=fseprd1024446
- Flannigan, M. D., Stocks, B. J., & Wotton, B. M. (2000). Climate change and forest fires. Science of The Total Environment, 262(3), 221-229. https://doi.org/10.1016/S0048-9697(00)00524-6
- Kala, H. P. (2023). Environmental and socioeconomic impacts of forest fires: A call for multilateral cooperation and management interventions. Natural Hazards Research, 3(2), 286-294. https://doi.org/10.1016/j.nhres.2023.04.003

- Fann, N., Alman, B., Broome, R. A., Morgan, G. G., Johnston, F. H., Pouliot, G., & Rappold, A. G. (2018). The health impacts and economic value of wildland fire episodes in the U.S.: 2008–2012. Science of The Total Environment, Volumes 610–611, 802-809. https://doi.org/10.1016/j.scitotenv.2017.08.024
- Koopmans, E., Cornish, K., Fyfe, T.M. et al. Health risks and mitigation strategies from occupational exposure to wildland fire: a scoping review. J Occup Med Toxicol 17, 2 (2022). https://doi.org/10.1186/s12995-021-00328-w
- Schneider, P., & Xhafa, F. (2022). Anomaly detection: Concepts and methods. Academic Press. 49-66. https://doi.org/10.1016/B978-0-12-823818-9.00013-4
- Koutsoukas, A., Monaghan, K.J., Li, X. et al (2017). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. J Cheminform 9, 42. https://doi.org/10.1186/s13321-017-0226-y
- Han, H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial Neural Network: Understanding the Basic Concepts without Mathematics. Dementia and Neurocognitive Disorders, 17(3), 83-89. https://doi.org/10.12779/dnd.2018.17.3.83
- Ramadhan, Muhammad & Sitanggang, Imas & NASUTION, Fahrendi & GHIFARI, Abdullah. (2017). Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency. DEStech Transactions on Computer Science and Engineering. 10.12783/dtcse/cece2017/14611.