**FOREST COVER TYPE PREDICTION**

|  |  |
| --- | --- |
| *M.BHUMIKA*  *SCSE, M.Tech Integrated Software Engineering, (Student) VIT University Chennai, Tamil Nadu, India*  *Manchikanti.bhumika2018@vitstudent.ac.in* | *B.PRIYANKA*  *SCSE, M.Tech Integrated Software Engineering, (Student) VIT University Chennai, Tamil Nadu, India*  *b.priyanka2018@vitstudent.ac.in* |
|  |  |

***Abstract*** *- The global problem of monitoring the situation of trees in nature, the role of trees in human life and the planet, their impact on nature; the main problems caused by deforestation can be monitored.Predict the forest cover type (the predominant kind of tree cover) from cartographic variables i.e., position, size, shape, value, color, orientation, and texture.. The data is in raw form and contains binary columns of data for qualitative independent variables such as wilderness areas includes elevation, slope, distance from road, hill shade etc… and soil type. Based on the variables we are going to predict the forest cover types i.e., the seven types of forest covers are Spruce/Fir, Lodge pole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, Krummholz. The main goal of the Project is to predict the cover types of seven different forests in four different wilderness areas i.e.,Rawah, Neota,Comanche Peak, Cache la Poudre. These areas represent forests with less disturbances caused by human, so that existing forest cover types are the results of ecological processes.*

***Keywords*** *– Deforestation, soiltype, Forest cover type, Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir, Krummholz*

1. **Introduction**

In this experiment we will be analyzing various machine learning algorithms on Forest Cover Type. We need to predict the cover type using the given data. We will also analyze various parameters of the applied algorithm and view the effect on the dataset. Parameter tuning using Grid search is done. The dataset is asking one to predict the type of forest cover from given cartographic variables. The actual forest cover was modeled into a 30 by 30-meter cell and was issued by USFS. The study includes four different types of wilderness areas and 7 different types namely Spruce, Lodge pole Pine, Ponderosa Pine, Cottonwood, Aspen, Douglas-fir, Krummholz..

1. **Objective**

* Identifying the best suitable model with highest accuracy and based on that particular model predict the forest cover type from cartographic variables i.e., position, size, shape, value, color, orientation, and texture.

1. **Literature Review**

|  |  |  |  |
| --- | --- | --- | --- |
| S.no | Author | Published | Title |
| 1. | Rahul R. Kishore | 2016 | Comparison of different accuracies of algorithmsfor prediction of forest cover type. |
| 2. | Hector Franco-Lopez | 2012 | Estimating and mapping of forest Covertype based on density and volume using the k-nearest neighbours method |
| 3. | Brett Meyer | 2015 | Forest Covertype prediction |
| 4. | * Vasyl Kiyko | 2020 | Based on Environment Characteristics predicting Forest cover type using Machine Learning Technologies. |
| 5. | Tun, Aye Mon | 2015 | Comparison of Data Mining Classification Techniques: A Case Study. |

[1]These paper is examining all of these classifiers coupled with feature selection in order to evaluate which one is best suited for forest cover type classification. Logistic regression, K-nearest neighbours, Decision tree, Ensemble learning.

[2] Identifying ancillary factors driving local response would be very helpful to improve the KNN estimation procedure. The quality of cover type classification results were compared with the quality obtained using parametric approaches.

[3] 70% classification was achieved, matching the results obtained by Blackard in his comparison of linear discriminant analysis and a back-propagation algorithm.

[4] This article briefly describes the global problem of monitoring the situation of trees in nature, the role of trees in human life and the planet, their impact on nature; the main problems caused by deforestation are highlighted. This work aims to develop a model for the forest cover type determination based on environmental characteristics and machine learning as the currently developing project part “Monitoring the condition of trees using drones’. The main aim of the project or study is to simplifyand automate the controlling and monitoring of trees under forests using the drones and machine learning to improve the situation of forest by providing better predictions.

[5] To predict unknown class labels of new data based on the trained data of known class labels. In forest management, forest inventory will play the vital role and support as the core repository for forest data. For the availability of accurate forest data for forest inventory is a time-consuming, labour intensive and task with more cost. This research examines three commonly used data mining classification techniques: Support Vector Machines (SVM), Random Forest (RF) and k-Nearest Neighbours (k-NN), and compares their classification abilities of correct forest cover type by using forest cover type data.

1. **Proposed Methodology**

The proposed methodology of this research includes identifying the best suitable model with highest accuracy and based on that particular model predict the forest cover type from cartographic variables.

The used machine learning models are

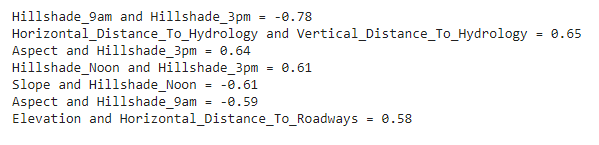
* **KNN Classifier**:The k-nearest neighbours (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both regression and classification problems. It is very simple and very easy to implement and understand the algorithm, but has a major drawback of becoming slows as the size of that data increases..
* **Random Forest Model**: It is an ensemble learning method for classification, Regression and other tasks that operate by constructing a multitude of decision Tree at training time.
* **Naïve Bayes Classifier:** Its a supervised machine learning algorithm i.e., based on Bayes theorem and used for solving classification problems. ...It is one of the most simplest and effective Classification algorithms that helps in building the model that can make fast predictions.
* **Gradient Boosting Using sklearn Model**: Gradient Gradient Boosting for classification. Gradient boosting algorithm builds a model in a stage-wise forward fashion; it allows for the minimization of arbitrary differentiable loss functions. In each stage there will be n classes, regression trees are fit on the negative gradient side of the binomial, trinomial or multinomial deviance loss function..

1. **Proposed System Procedure**
2. Data set import
3. Importing packages in Jupiter
4. Data interpretation and analysis (data types of the attributes, size etc..)
5. Correlation among the features
6. Data visualization based on wilderness area and soiltype
7. Data cleaning(Removing unwanted columns)
8. Data preparation(Test and Train data)
9. Applying methodologies
   * + - Random Forest
       - KNN Classifier
       - Decision Tree
10. Analysing and comparing all the methodologies and selecting the best one.

**Methodologies Used:-**

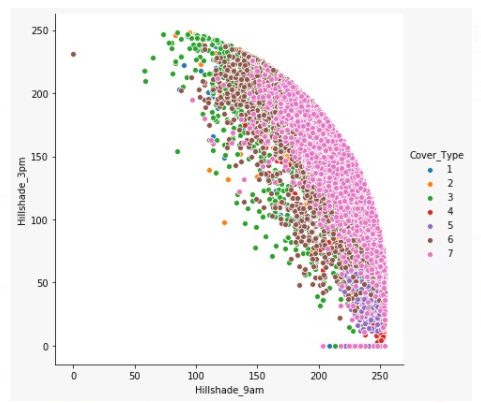
1. Random Forest
2. KNN Classifier
3. Decision Tree
4. **Implementation**

**Correlation among all Features:**

****

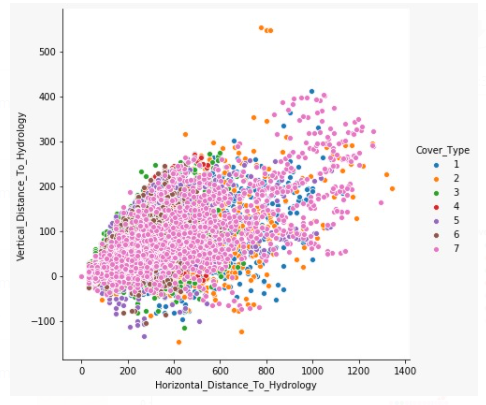
***Fig\_1*** *correlation between selected featrues*

**Fig\_1** The threshold value for correlation is set to **0.5** . If the correlation between features is greaterthan 0.5 to 1 then they have good correlation else bad correlation between features. The above *Fig\_1*depicts the correlation between features which have good correlation.



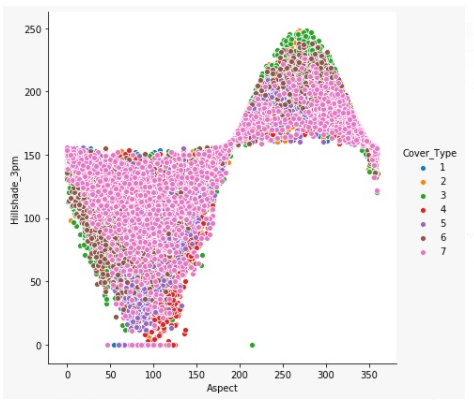
***Fig\_2:*** *Cover\_Type Based on hillshade\_3pm and hillshade\_9pm*

**Fig\_2** Based on hillshade\_3pm and hillshade\_9pm for each forest data, it predicts which type of Covertype it belongs to.



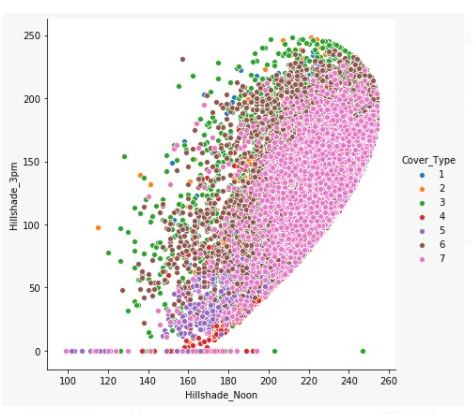
***Fig\_3:*** *Cover\_Type Based on Vertical\_Distance\_To\_Hydrology and Horizontal\_Distance\_To\_Hydrolozy*

**Fig\_3** Based on Vertical\_Distance\_To\_Hydrology and Horizontal\_Distance\_To\_Hydrolozyfor each forest data; it predicts which type of Covertype it belongs to.



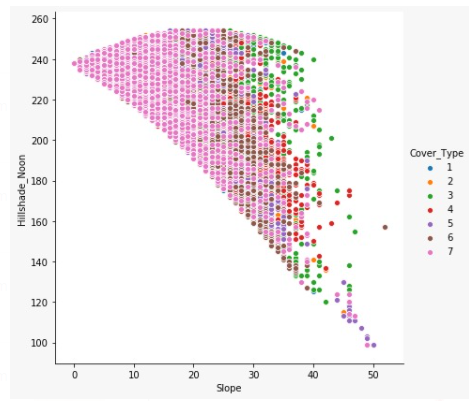
***Fig\_4:*** *Cover\_Type Based on Hillshade\_3pm and Aspect.*

**Fig\_4** Based on hillshade\_3pm and Aspect for each forest data, it predicts which type of Covertype it belongs to.



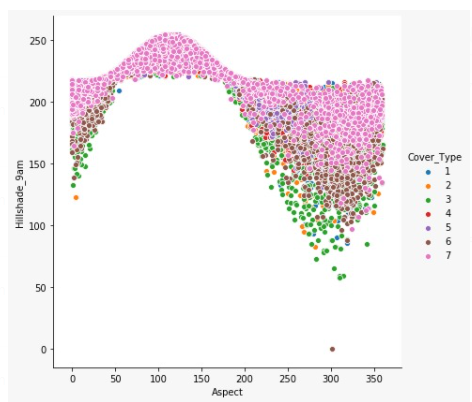
***Fig\_5:*** *Cover\_Type Based on Hillshade\_3pm and Hillshade\_Noon*

**Fig\_5** Based on hillshade\_3pm and Hillshade\_Noon for each forest data, it predicts which type of Covertype it belongs to.



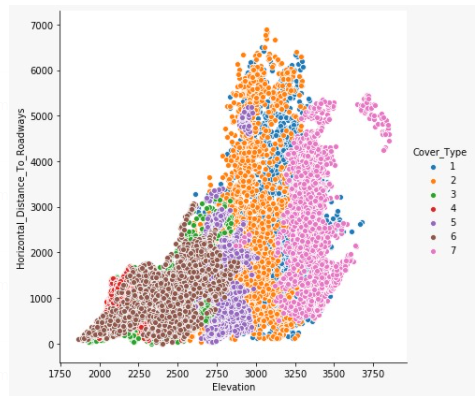
***Fig\_6:*** *Cover\_Type Based on Hillshade\_Noon and Slope*

**Fig\_6** Based on Hillshade\_Noon and Slope for each forest data, it predicts which type of Covertype it belongs to.



***Fig\_7:*** *Cover\_Type Based on Hillshade\_9pm and Aspect.*

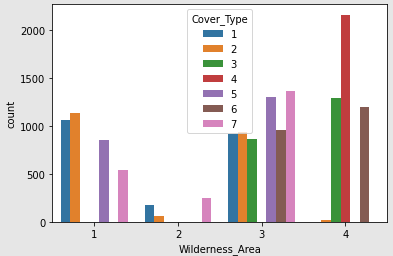
**Fig\_7** Based on Hillshade\_9pm and Aspect for each forest data, it predicts which type of Covertype it belongs to.



***Fig\_8:*** *Cover\_Type Based on Horizontal\_Distance\_To\_Roadways and Elevation*

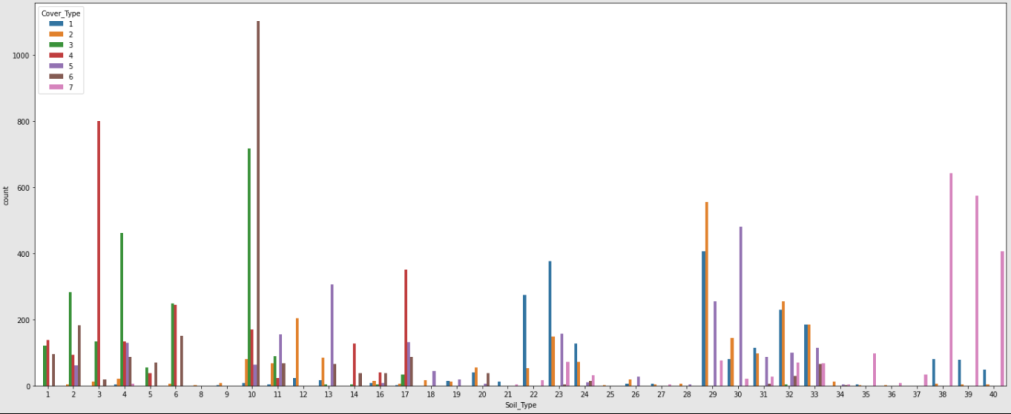
**Fig\_8** Based on Horizontal\_Distance\_To\_Roadways and Elevation for each forest data, it predicts which type of Covertype it belongs to.

**Data Visualization**



***Fig\_9:*** *Data visualization based on wilderness area*

**Fig\_9** Based on four different wilderness areas it predicts which type of Covertype it is.

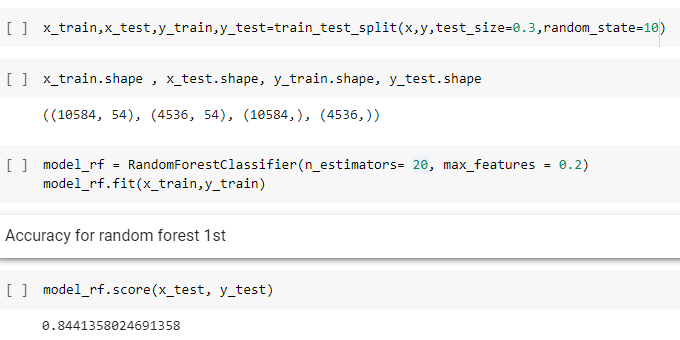


***Fig\_10:*** *Data visualization based on Soil\_type*

**Fig\_10** Based on soil types in different areas predicts which type of Covertype it is.

1. **Random Forest Methodology**

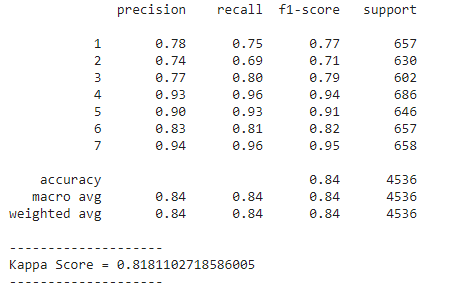
**Accuracy for sample1:**



***Fig\_11:*** *Accuracy for 20 percent Sample Data*

**Fig\_11** Based on Random Forest the accuracy found for 20 random estimators.

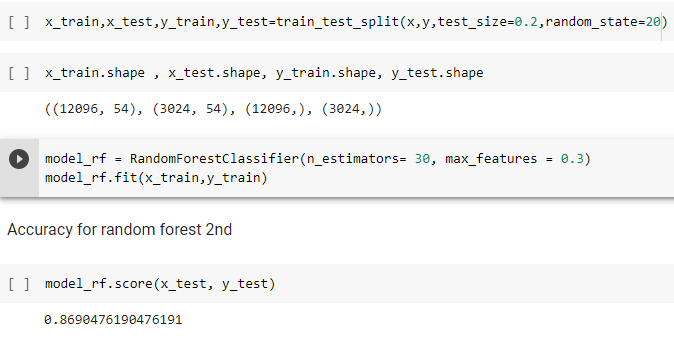
**Precision for sample 1:**



***Fig\_12:*** *Precision, Recall, F1-Score, Support*

**Fig\_12** Precision, Recall, F1-Score, Supportfor 7 Cover\_Types and average of Precision, Recall, F1-Score, Support.

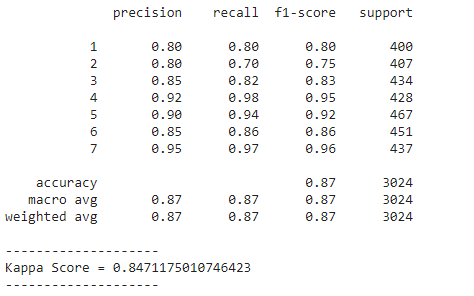
**Accuracy for Sample 2:**



***Fig\_13:*** *Accuracy for 30 percent Sample Data*

**Fig\_13** Based on Random Forest the accuracy found for 30 random estimators

**Precision for Sample2:**



***Fig\_14:*** *Precision, Recall, F1-Score, Support*

**Fig\_14** Precision, Recall, F1-Score, Supportfor 7 Cover\_Types and average of Precision, Recall, F1-Score, Support.

1. **KNN Classifier**

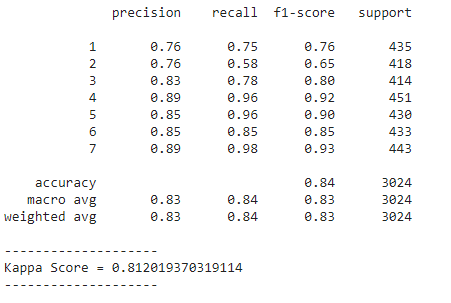
**Accuracy for Sample1**

****

***Fig\_15:*** *Accuracy for 30 percent Sample Data*

**Fig\_15** Based on KNN Classifier the accuracy found for 20 random estimators

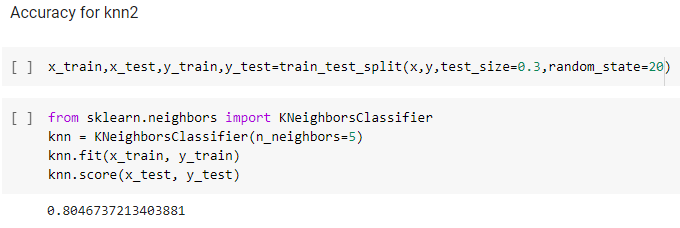
**Precision for Sample1**

****

***Fig\_16:*** *Precision, Recall, F1-Score, Support*

**Fig\_16** Precision, Recall, F1-Score, Supportfor 7 Cover\_Types and average of Precision, Recall, F1-Score, Support.

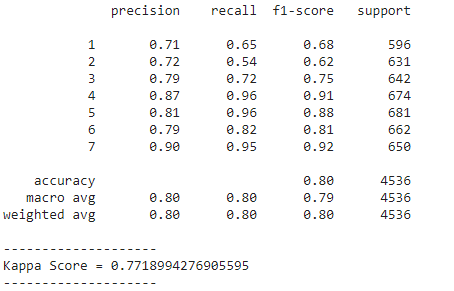
**Accuracy for Sample2:**

****

***Fig\_17:*** *Accuracy for 30 percent Sample Data*

**Fig\_17** Based on KNN Classifier the accuracy found for 30 random estimators

**Precision for sample2:**

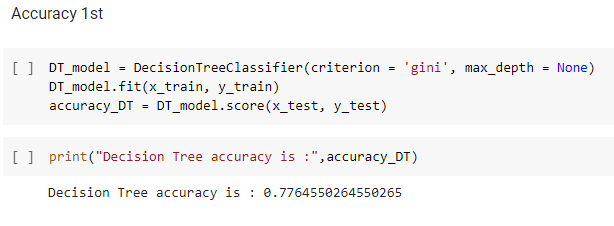
****

***Fig\_18:*** *Precision, Recall, F1-Score, Support*

**Fig\_18** Precision, Recall, F1-Score, Supportfor 7 Cover\_Types and average of Precision, Recall, F1-Score, Support.

1. **Decision tree:**

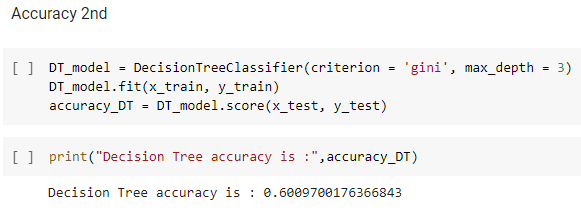
**Accuracy for sample1:**



***Fig\_19:*** *Accuracy*

**Fig\_19** Accuracy found based on Decision tree

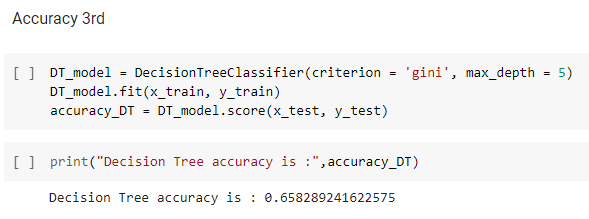
**Accuracy for sample2:**



***Fig\_20:*** *Accuracy*

**Fig\_20** Accuracy found based on Decision tree

**Accuracy for sample3:**



***Fig\_21:*** *Accuracy*

**Fig\_21** Accuracy found based on Decision tree

1. **Results**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | recall | F1 Score | Support | Kappa |
| Random Forest Sample1 | 84.41% | 84% | 84% | 84% | 4536 | 81.81% |
| **Random Forest Sample2** | **86.90%** | **87%** | **87%** | **87%** | **3024** | **84.71%** |
| KNN Classifier  Sample1 | 83.89% | 83% | 84% | 83% | 3024 | 81.20% |
| KNN Classifier  Sample2 | 80.46% | 80% | 80% | 80% | 4536 | 77.18% |
| Decision Tree  Sample1 | 77.64 | 77% | 77% | 77% | 2987 | 73.13% |
| Decision Tree  Sample2 | 60.09 | 61% | 61% | 61% | 2569 | 57.87% |
| Decision Tree  Sample3 | 65.82 | 66% | 66% | 66% | 2763 | 62.13% |

***Table\_1:*** *Comparison between different methodologies.*

**Table\_1** From observation of above ***Table\_1***it predicts **Random Forest Methodology** for sample2 (random 30 at a time) is the best methodology for this data.

1. **Conclusion**

|  |  |
| --- | --- |
| **Algorithms Implemented** | **Accuracy** |
| **Random Forest Algorithm** | **86.90%** |
| KNN Classifier | 83.89% |
| Decision Tree Classifier | 77.64% |

***Table\_2:*** *Accuracies for different Algorithms*

**Table\_2**Based on the above table it concludes Random Forest is the best model for the data so based on the random forest predicted the forest cover types.

Random Forest Algorithm performed marginally better than the others for prediction of forest cover type prediction.So in future, if anyone want to predict any forest cover types who belongs to that area, we suggest to use random forest because it gives less risk more corrective results. So that no any future risks will happen.

1. **Acknowledgement**

We would like to express our special thanks to our teacher who gave us the golden opportunity to do this wonderful research on the topic “Forest cover type prediction”, which also helped us doing a lot of research and we come to discover about so many new things. It helped us to increase our knowledge and skills.

1. **REFERENCES**

[1] H Gholami, A Mohamadifar, A Sorooshian… - Atmospheric Pollution - Elsevier \_Machine-learning algorithms for predicting land susceptibility to dust emissions…, 2020

[2] C.-C. Chang and C.-J. Lin, "LIBSVM - a library for support vector machines," in http://www.csie.ntu.edu.tw/~cjlin/libsvm/, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan, 2017.

[3] Q Zhao, S Yu, F Zhao, L Tian, Z Zhao - Forest Ecology and Management, 2019 - Elsevier\_Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments.

[4] R. Collobert, S. Bengio, and Y. Bengio, "A parallel mixtures of SVMs for very large scale problems," Neural Computation, vol. 14, pp. 1105-1114, 2020.

[5] Bahzad Taha Jijo1\*, Adnan Mohsin Abdulazeez2 1 IT Department, Technical College of Informatics Akre, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq, 2021.

[6] [C Voight](https://scholar.google.com/citations?user=MT4ImuUAAAAJ&hl=en&oi=sra), [K Hernandez-Aguilar](https://scholar.google.com/citations?user=clQvI1sAAAAJ&hl=en&oi=sra), [C Garcia](https://scholar.google.com/citations?user=hcJOpTcAAAAJ&hl=en&oi=sra), S Gutierrez - Remote Sensing, 2019.

[7] BT Pham, A Jaafari, M Avand, N Al-Ansari, T Dinh - Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction Symmetry, 2020.

[8] G.-B.Huang, L. Chen, and C.-K.Siew, "Universal approximation using incremental feedforward networks with arbitrary input weights," in Technical Report ICIS/46/2003, (School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore), Oct. 2013.

### [9] DA Petrusevich - IOP Conference Series: Earth and …, 2021- [Models for dominating forest cover type prediction](https://iopscience.iop.org/article/10.1088/1755-1315/677/5/052119/meta)

[10] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in International Conference on Machine Learning, pp. 148-156, 2016.

[11] ATN Dang, S Nandy, R Srinet, NV Luong, S Ghosh… - Ecological …, 2019 - Elsevier\_Forest aboveground biomass estimation using machine learning regression algorithm