ORIGINAL ARTICLE

© UIJIR | ISSN (O) - 2582-6417
July 2020 | Vol. 1 Issue.2
www.uijir.com

BORUTA ALGORITHM IS SIGNIFICANT FOR LARGE FEATURE SELECTION OF STUDENT MARKS DATA OF POKHARA UNIVERSITY NEPAL

Author's name: YagyanathRimal
School of Engineering, Pokhara University
E-mail:rimal.yagya@gmail.com

Abstract *Boruta algorithm is the best tool for large research data reduction highly used dimension reduction. Although, there was a larger research gap between the data collection and its interoperation, analysis of the appropriate variable of the big database for research generalization. There were many researchers have unknowingly misled their research due to the large set of data before applying the calculation of model execution. Therefore this primary study tries to reduce its features before model deployment, which ultimately significant model up for large data reduction for time and resources execution of machine learning model design. Therefore, this primary reduced 40 independent variables has been reduced largely to 21 variables of Pokhara university student grade data of student grade prediction. So that machine learning model executes quicker, and converse quickly rather than using unimportant variables of large datasets. The machine learning model produces a similar output with 0.7895 accuracies of 95% confidence interval, the p-value is 0.000236 has a significant output of reducing using the boruda algorithm of machine learning design for SGPA prediction of student grades.*

Keywords *Boruta Algorithm, Principal Component Analysis, Semester Grade Point Average, school leaving certificate*

INTRODUCTION

In the modern world, the data science researcherhas to handle many different sets of data with more dependent on single or multiple independent variables. In many cases, data scientists must reduce the dimensionality of data through before main components deployment(F Provost, 2013). The decomposition of singular value items into multiple attributes leads to divergence of research generalization(L Sorber, 2013) therefore this study attempt to reduce the multidimensional data before the design of the model for the research conclusion (Fallucchi, 2009). Manytechniques areapplied to unsupervised forms of feature selection for the calculation of PCA, which uses variance in the data for finding the components (Li, 2019). These techniques do not take into account the entity values and the class or target values. Furthermore, there are some assumptions, such as normality of data sets, associated with these methods are require some kind of transformation before starting to apply model deployment (Arcos, 2016)(F Ye, 2014). On the other hand, Boruta finds all the significant features that are very relevant to the decision variable (Kursa MB, 2016).Boruta's algorithm is a fundamental wrapper for the forest's random classification algorithm to reduce its data size which tries to capture all the important and interesting features of the data set for a dependent result variable (Dutta, 2018) of a categorical type(F Muharemi, 2019). These processes first duplicate the dataset, mix the values in each column of all datasets, and finally compare the outputs for the selection of an appropriate model (SB Kotsiantis, 2007). However, creativity is the shadow characteristics (Mola, 2015) for the development of the model. So, the algorithm checks each of its real characteristics if they are more important in each iteration of the development of the model algorithm compares the Z scores of the mixed copies of the characteristics with original characteristics(R Hinterding, 1999).If the last one worked better than the first, the last will be selected for applying model development. This process produced using the prediction and confusion matrix will be easily assessed accurately(E Arisholm, 2010). High-dimensional data, in terms of the number of features, are becoming increasingly common these days in machine learning problems (Danasingh, 2015) will be significantly reduced. The useful information from these high data volumes is tested with confusion matrix and statistical forecasting techniques

© UIJIR | ISSN (O) - 2582-6417
July 2020 | Vol. 1 Issue.2
www.uijir.com

for reducing noise or redundant data were analyzed (Xiong, 2019). This is significant because training in all feature models is often required (Maya Gopal P.S, 2018). The feature set can be viewed by creating a box chart of varying importance for each potential feature selection. The sophisticated feature selection algorithms such as boruta, genetic algorithms or simulated(R Prasad, 2019) annealing techniques are well known with the very high computational cost when the data set is scaled up (Kosinski, 2017). Boruta's feature selection package is ideal for big data reduction (Kursa M. B., 2010). There have been great advantages in this method which works with both classification and regression problems in multiple datasets of variable relationships. Boruta's improvement measures the importance of the random variable in the forest algorithm(J Elith, 2008), which is a very popular method for selecting variables reduction, on characteristics relevant to the result variable (Lewinson, 2019).The selection algorithms follow a method optimal minimum in which they are based on a small subset of characteristics that produce a minimum error in a classifier that is chosen with reducing a random measure of random forest (Kordeczka, 2018)(VF Rodriguez-Galiano, 2012).

METHODOLOGY

Boruta algorithm for feature selection was largely applied for produces less accurate output due to manyreasons. Which a large number of features reducesdelimitates time constraints to the former model using some shadow attributes on the model which shuffled all values ultimately creates a model including having importance reduction of model developments(U Wilensky, 2015). After data collection of student's grade whose independent variables have a large association for the predication of SGPA(R Zwick, 2005). The data are imported into the machine learning model using r programming were analyzed as after loading boruta, mlbench, caret, random forest and dplyr packages the data sets of students marks were loaded using read.csv and stored in 39 independent and single dependent variables. The data frame could be converted into tidyr format using tbl_df function whose data structure could be viewed using str function as below with 110 records of 41 columns data of student of Pokhara university student records. All records except SGPA grade all variables in numeric of marks of students like SLC, plus two, Chemistry, Maths, geology, physics, bio, with their internal practical and theoretical marks, etc for the dependent for grade prediction, the grade is in the factor of students scored 4,3,2 and SGPA converted into A, B, C and fail grade respectively the data structure as below could be analyzed easily using r package.

$ SLC11:int 88 60 80 77 82 76 76 79 79 81 ...
$ Peplums:int 83 62 74 72 77 76 68 75 76 72 ...
……
$ Geology6:int 44 44 55 44 70 80 55 44 44 65 ...
$ SGPA: Factor w/ 4 levels "A Grade","B Grade", 4 4 3 4 2 2 3 4 4 4 ...

The data preparation process firstly removes not available i.e. NA and blank fields of all data set are omitted using na.omit(database). The independent variable SGPA with A, B, C, fail grade fields were stored using as. factor(S$SGPA) whose value could be easily converted and stored as numeric as. numeric(S$SGPA). Whose values were again viewed as str () function? The normalization process is always required before applying a model that could be done using mean max reduction of each value so that the data were ranged from 0 to 1 of all data because marks of all subjects ranging from 0 to number would be easily converted in each field as.data. frame (apply (S [, 1:40], 2, function(x) (x - min(x))/(max(x)-min(x)))) function. Whose data were again viewed as follows?

$ SLC11:num 0.9355 0.0323 0.6774 0.5806 0.7419 ...
……
$ Project4:num 1 1 1 1 0 1 1 1 1 1 ...

The new field on data sets with zero values was added and the factor values of the independent variables are added using Sr$SGPA=0 and Sr [,41] =p function. Whose again data structure was the view of all 110 obs of 41 variables as:

$ SLC11:num 0.9355 0.0323 0.6774 0.5806 0.7419 ...
$ Peplums:num 1 0.125 0.625 0.542 0.75 ...
$ PPPhysics:num 0.973 0.405 0.351 0.405 0.703 ...
……

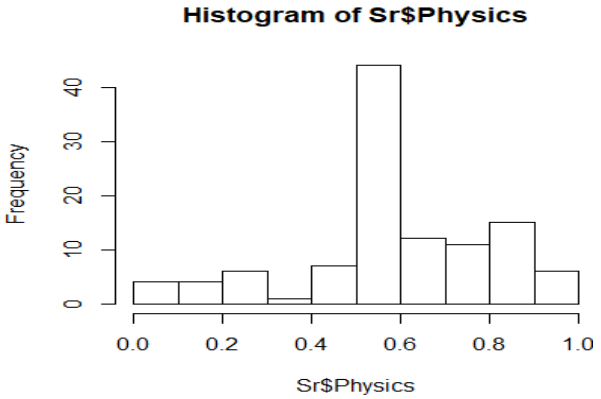
© UIJIR | ISSN (O) - 2582-6417
July 2020 | Vol. 1 Issue.2
www.uijir.com

$ Project4:num 1 1 1 1 0 1 1 1 1 1 ...
$ SGPA: Factor w/ 4 levels "A Grade","B Grade", 4 4 3 4 2 2 3 4 4 4 ...

The summary(Sr) function summarized the database with its minimum, quartile median mean and a max of each column values as below which implies the data with normalized ranged from 0 to 1.

SLC11 PPplusPPPhysicsMMMathCCChemistryBBBio SLC1
Min.:0.00 Min. :0.00 Min. :0.00 Min. :0.00 Min. :0.00 Min. :0.00 Min. :0.00
1st Qu.:0.54 1st Qu.:0.51 1st Qu.:0.54 1st Qu.:0.54 1st Qu.:0.40 1st Qu.:0.68 1st Qu.:0.54
Median :0.64 Median :0.58 Median :0.54 Median :0.57 Median :0.56 Median :0.81 Median :0.64
Mean :0.63 Mean :0.60 Mean :0.57 Mean :0.63 Mean :0.55 Mean :0.67 Mean :0.63
3rd Qu.:0.80 3rd Qu.:0.75 3rd Qu.:0.70 3rd Qu.:0.76 3rd Qu.:0.78 3rd Qu.:0.87 3rd Qu.:0.80
Max. :1.00 Max. :1.00 Max. :1.00 Max. :1.00 Max. :1.00 Max. :1.00 Max. :1.00
PplusPPhysicsMMathCChemistryBBio SSLC Plus Strength3 Geology1
Geology2 Geology3 ProjectI Math3 Applied3 Material6 Fludi7
Strength6 Geology6 Project4 SGPA

The sample histogram could be easily calculated using hist(Sr$Physics)



The set.seed (111) command generally used to get the same output, if this command will run again and the names command will display all 41 variables the column name of database as.

[1] "SLC11""PPplus""PPPhysics""MMMath""CCChemistry""BBBio""SLC1""Pplus"
[9] "PPhysics""MMath""CChemistry" "BBio""SSLC""Plus""Physics""Math"
[17] "Chemistry""Bio""Math2""Applied2""Materials1""Materials2""Materials3""Fluid1"
[25] "Fluid2""Fluid3""Strength1""Strength2""Strength3""Geology1""Geology2""Geology3"
"ProjectI""Math3""Applied3""Material6""Fludi7""Strength6""Geology6""Project4" [41] "SGPA".

We can write the console output table to the file usingwrite.csv (Sr,file = "C:/Users/user/Desktop/publication/baruda/SSe.csv"). if the data in r library the data could load using data("Sonar") command.

The model boruta is designed with dependent variable SGPA with all other 40 variables of marks with trace 2 and max run with 400 and stored in the variable. boruta=Boruta (SGPA~., data=Sr, doTrace=2,maxRun=400). The model run produced the output with time expressing important and non-important attributes as below. A model run the attributes were analyzed and accepted and rejected attributes

1. run of importance source...
……
12. run of importance source...

After 12 iterations, +1.3 secs: confirmed 10 attributes: Applied3, Fludi7, Fluid2, Geology6, Material6, and 5 more; rejected 13 attributes: Fluid1, Fluid3, Materials2, Math, MMath and 8 more; still have 17 attributes left.

13. run of importance source...
……
17. run of importance source...

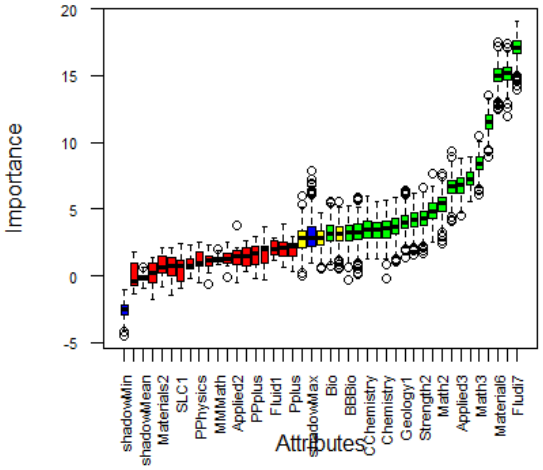
After 17 iterations, +1.6 secs: rejected 1 attribute: PPPhysics; still have 16 attributes left.

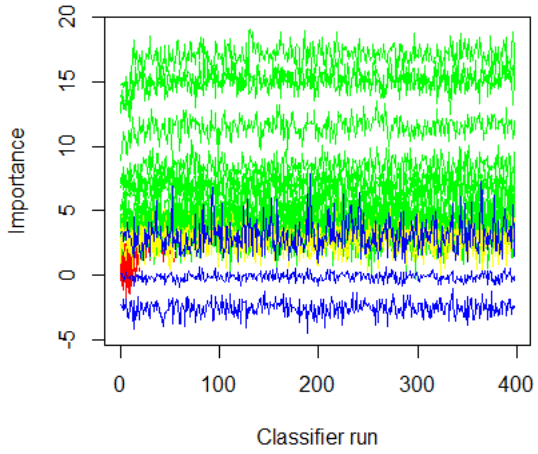
20. run of importance source...

After 20 iterations, +1.9 secs: confirmed 3 attributes: Geology1, Materials1, ProjectI; rejected 2

© UIJIR | ISSN (O) - 2582-6417
July 2020 | Vol. 1 Issue.2
www.uijir.com

attributes: Applied2, PPplus; still have 11 attributes left.
21. run of importance source...
……
57. run of importance source...
After 57 iterations, +4.5 secs: rejected 1 attribute: Plus; still have 10 attributes left.
……..
115. run of importance source...
……..
308. run of importance source...
After 308 iterations, +23 secs: confirmed 1 attribute: BBio; still have 3 attributes left.
309. run of importance source...
……..
399. run of importance source...
The final model produced using print(boruta) command display boruta performed 399 iterations in 29.92149 secs with 20 attributes confirmed as important are Applied3, BBBio, BBio, Bio, CCChemistry and 15 more and 17 attributes confirmed as unimportant are Applied2, Fluid1, Fluid3, Materials2, Math and 12 more and 3 tentative attributes left were left are Geology2, Geology3, Strength3.
The plot (Boruta, las=2, cex. axis=0.7) command demonstrates box plot of important in color green, the yellow box plots are in tentative attributes, and red colors were not important attributes similarly to the blue box plots are suspect shadow attributes.





The history plot describes the line chart describes important as green them were non-important the yellow was tentative and blue were declared as non-important attributes using plotImpHistory(boruta) function.
>bor=TentativeRoughFix(boruta) command declared for tentative rough fixed clearly expressed important or non-importantfinal model with including suspect attributes too. Therefore, the 20,17 and 3 attributes become 21 conformed and 19 rejected could be finalized using print(bor) command confirmed and rejected the output demonstrate as boruta performed 399 iterations in 29.92149 secs.Tentativerough fixed over the last 399 iterationsdeclared 21 attributes confirmed important as Applied3, BBBio, BBio, Bio, CCChemistry and 16 more;and 19 attributes confirmed unimportant as Applied2, Fluid1, Fluid3, Geology3, Materials2 and 14 more. Whose statists could have printedusingattStats(boruta) command produces the following output. Which describes accepted, rejected, and tentative with its statists values as below so that decision could have made easily.
meanImpmedianImpminImpmaxImpnormHits decision
SLC11 0.90633173 0.7725492 -0.15335152 2.283735 0.000000000 Rejected
……
CCChemistry 3.41333322 3.4886646 1.23593639 5.611042 0.644110276 Confirmed
BBBio 3.23313223 3.2670364 -0.29251056 5.301308 0.593984962 Confirmed
……
BBio 3.24511333 3.2908627 0.16667723 5.847855 0.596491228 Confirmed
SSLC 0.07366821 -0.4382772 -1.31001689 1.791377 0.000000000 Rejected

© UIJIR | ISSN (O) - 2582-6417
July 2020 | Vol. 1 Issue.2
www.uijir.com

Plus 2.11558915 2.1086140 0.64110570 3.8718800.037593985 Rejected
……
Strength3 2.75314302 2.8135053 0.05396369 5.964634 0.431077694 Tentative
Geology1 4.02361983 4.0379165 1.37082066 6.398025 0.764411028 Confirmed
Geology2 2.79822026 2.8636872 0.54999709 4.753601 0.493734336 Tentative
Geology3 3.12766595 3.1959905 0.49267083 5.607372 0.576441103 Tentative
……
Project4 0.17603587 0.2606932 -1.73346659 1.415440 0.000000000 Rejected
datapartitionis for increasingaccuracy of model testing startswith set. seed (222) and dividing 60% from training data and 40% for testing data sets as ind=sample (2, nrow(S), replace=T, prob=c (.60,.40)). The training will train=S[ind==1,] andtesting test=S[ind==2,]. The getNonRejectedFormula(boruta) formula includes all conformed and suspected values of (21+3=24) attributes as SGPA ~ CCChemistry + BBBio + CChemistry + BBio + Chemistry + Bio + Math2 + Materials1 + Materials3 + Fluid2 + Strength1 + Strength2 + Strength3 + Geology1 + Geology2 + Geology3 + ProjectI + Math3 + Applied3 + Material6 + Fludi7 + Strength6 + Geology6.
The getConfirmedFormula(boruta) finalized 21 attributes with adding suspect as SGPA ~ CCChemistry + BBBio + CChemistry + BBio + Chemistry + Bio + Math2 + Materials1 + Materials3 + Fluid2 + Strength1 + Strength2 + Geology1 + ProjectI + Math3 + Applied3 + Material6 + Fludi7 + Strength6 + Geology6.
After having set.seed (333) the 41 models will have trained with a train set of data rf41=randomForest (SGPA~., data=train). Produces random forest classification numberof 500 trees, no of variables tried at each split 6, and OOB estimate of error rate: 13.89% and produces confusion matrix.

	A	B	C	Fail	Error
A Grade	0	1	0	0	1.00000000
B Grade	0	24	0	2	0.07692308
C Grade	0	0	0	6	1.00000000
Fail	0	1	0	38	0.02564103

From the above miss classification, the Grade of A and C were not missing classification the grade B and Fail grade with miss classification 2 and 8, 62 data out of 72 data were accurately classified of training data.The final prediction is calculated p=predict (rf41, test) as
1 2 3 4 5 6 7 as Fail B Grade Fail FailFailFail B Grade
The confusion matrix (p, test$SGPA)
Confusion Matrix and Statistics
Reference: Prediction A Grade B Grade C Grade Fail

A Grade	0	0	0	0
B Grade	1	18	0	0
C Grade	0	0	0	0
Fail	0	1	6	12

Overall statistics, accuracy is 0.7895 percent with 95% confidence intervalis (0.6268, 0.9045), no information rate is0.5, P-Value is0.000236, Kappa is 0.6444 of McNamara's Test P-Value is NA.

	Class: A Grade	Class: B Grade	Class: C Grade	Class: Fail
Sensitivity	0.00000	0.9474	0.0000	1.0000
Specificity	1.00000	0.9474	1.0000	0.7308
PosPred Value	NaN	0.9474	NaN	0.6316
NegPred Value	0.97368	0.9474	0.8421	1.0000
Prevalence	0.02632	0.5000	0.1579	0.3158
Detection Rate	0.00000	0.4737	0.0000	0.3158
Detection Prevalen	0.00000	0.5000	0.0000	0.5000
Balanced Accuracy	0.50000	0.9474	0.5000	0.8654

The model with 21 +3 attributes again designed random forest with training data sets.
> rf24=randomForest (SGPA ~ CCChemistry + BBBio + CChemistry + BBio + Chemistry + Bio + Math2 + Materials1 + Materials3 + Fluid2 + Strength1 + Strength2 + Strength3 + Geology1 + Geology2 + Geology3+ ProjectI + Math3 + Applied3 + Material6 + Fludi7 + Strength6+ Geology6,

© UIJIR | ISSN (O) - 2582-6417
July 2020 | Vol. 1 Issue.2
www.uijir.com

data=train)
> rf21=randomForest (SGPA ~ CCChemistry + BBBio + CChemistry + BBio + Chemistry + Bio + Math2 + Materials1 + Materials3 + Fluid2 + Strength1 + Strength2 + Geology1 + ProjectI + Math3 + Applied3 + Material6 + Fludi7 + Strength6 + Geology6, data=train). And its prediction will be p=predict (rf24, test) calculated and confusionmatrix (p, test$SGPA) confusion matrix and statistics describes as

Prediction	A	B	C	Fail
A Grade	0	0	0	0
B Grade	1	18	1	0
C Grade	0	0	0	0
Fail	0	1	5	12

Overall model statisticsdescribes its accuracy is 0.7895 with 95% confidence interval is (0.6268, 0.9045), no information rate is0, P-Value is0.000236, Kappa is 0.6415,Mcnemar's Test P-Value is NA.Statistics by Class:

	A	B	C	Fail
Sensitivity	0.00000	0.9474	0.0000	1.0000
Specificity	1.00000	0.8947	1.0000	0.7692
PosPred Value	NaN	0.9000	NaN	0.6667
NegPred Value	0.97368	0.9444	0.8421	1.0000
Prevalence	0.02632	0.5000	0.1579	0.3158
Detection Rate	0.00000	0.4737	0.0000	0.3158
Detection Prevale	0.00000	0.5263	0.0000	0.4737
Balanced Accuracy	0.50000	0.9211	0.5000	0.8846

From the above, both 41 and 24 models both produce similar accuracy therefore we rather selecting all 41 variables researcher select 24 importance attributes for model design. Similarly, the model with 21 single variables without including suspected attributes will be further predicated as p=predict (rf21, test) and whose confusion matrix (p, test$SGPA) produces the output.

Prediction	A	B	C	Fail
A Grade	0	0	0	0
B Grade	1	18	0	0
C Grade	0	0	0	0
Fail	0	1	6	12

Overall statistics of accuracyis 0.7895 with 95% confidence interval is (0.6268, 0.9045), noinformation rate is 0.5 with p-value is 0.000236, Kappa is 0.6444, Mcnemar's Test P-Value is NA.

	A	B	C	Fail
Sensitivity	0.00000	0.9474	0.0000	1.0000
Specificity	1.00000	0.9474	1.0000	0.7308
PosPred Value	NaN	0.9474	NaN	0.6316
NegPred Value	0.97368	0.9474	0.8421	1.0000
Prevalence	0.02632	0.5000	0.1579	0.3158
Detection Rate	0.00000	0.4737	0.0000	0.3158
Detection Prevalen	0.00000	0.5000	0.0000	0.5000
Balanced Accuracy	0.50000	0.9474	0.5000	0.8654

From the above model accuracy of the final 21 attributes model could produce a similar output of 24 attributes and 41 attributes model could replace for features selection for machine learning design so that time and model implementation largely reduced using appropriate selection.

CONCLUSION

During data preparation, many researcherssuccessfully usefiltered out the most important features onthe dataset roughly reduces a few lines of code. This sometimes leads to misleading data analysis and interoperation could be easily reduced using Boruta algorithm for feature selection using r programming.This process not only reduced the noise data but also really beneficial for any classifier to assign a label to research data. The model development will be

© UIJIR | ISSN (O) - 2582-6417
July 2020 | Vol. 1 Issue.2
www.uijir.com

training and testing modelincreasefeatures selection will improve the model's performance of using Boruta algorithm is very simple: there aren't many parameters to tune. One has to remember that the data set has to be complete (no NA's). this article tries to explain each line ofdesign of large student data sets of Pokhara university student marks of 41 dependent variables largely reduced 21 before model with having similar accuracy and importance of variables could achieve.

REFERENCES

1. Arcos, D. F. (2016). When is data transaformation needed when dealing with community composition data.
2. Danasingh, A. A. (2015). Literature review on feature selection methods for high-Dimensional data.
3. Dutta, D. (2018). How to perform feature selection (i.e. pick important variables) using Boruta package in R,https://id.analyticsvidhya.com
4. E Arisholm, L. B. (2010). A systematic and comprehensive investigation of methods to build and evaluate fault prediction models.
5. F Muharemi, D. L. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set.
6. F Provost, T. F. (2013). Data Science for Business. What you need to know about data mining and data-analytic thinking.
7. F Ye, D. L. (2014). Comparing three commonly used crash severity models on sample size requirements: multinational logit, ordered probit and mixed logit models.
8. Fallucchi, F. (2009). Singular value Decomposition for Feature selection in Taxonomy Learning.
9. J Elith, J. L. (2008). A working guide to boosted regression trees.
10. Kordeczka, A. (2018). Boruta – modern dimension reduction algorithm.
11. Kosinski, M. (2017). Feature selection with the Boruta Algorithm.
12. Kursa, M. B. (2010). Feature Selection with the Boruta Package. Statistical journal.
13. Kursa, M.B. (2016). Interdisciplinary Centre for Mathematical and Computational Modelling University of Warsaw.
14. L. Sorber, M.V. (2013). Decomposition, Decomposition in Rank- terms and a new generalization.
15. Lewinson, E. (2019). Explaining feature Inportance by example of a Random Forest.
16. Li, L. (2019). Proncipal Component Analysis for Dimensionality Reduction.
17. Maya Gopal P.S, 2, R. (2018). Feature Selection for yield prediction Using BorutaAlgorith. Internation Journal of Pune and applied mathematics.
18. Mola, D. (2015). BorutaPy – an all relevant feature selection method.
19. R. Hinterding, Z. M. (1999). Parameter control in evolutionary algorithms.
20. R Prasad, R. D. (2019). Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta- random forest hybridizer algorithm approach.
21. R Zwick, J. S. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language.
22. SB Kotsiantis, I. Z. (2007). Supervised machine learning: A review of classification techniques.
23. U Wilensky, W. R. (2015). An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with netLogo.
24. VF Rodriguez-Galiano, B. G. (2012). An assessment of the effectiveness of a random forest classifier for land- cover classifies.

© UIJIR | ISSN (O) - 2582-6417
July 2020 | Vol. 1 Issue.2
www.uijir.com

25. Xiong, H. (2019). Enhancing data analysis with noise removal.