A Comparative Study on Decision Tree and Random Forest using Konstanz Information Miner (KNIME)

Dr. Akanksha Khanna[1,] Nibedita Dey[2]
[1]Assistant Professor, Department of Commerce, CHRIST (Deemed to be University), Bangalore-560029
[2]Post Graduate (M.Com) Research Scholar, CHRIST (Deemed to be University), Bangalore-560029
[1]akanksha.khanna@christuniversity.in
[2]nibedita.dey@mcom.christuniversity.in

Abstract

With vast amounts of data floating around everywhere, it is imperative to comprehend and draw meaningful insights from the same. With the proliferation of Internet and Information Technology, data has been increasing exponentially. The 5 Vs of data i.e. Value, volume, Velocity, variety and veracity will only make sense if we are able to examine the data and uncover the hidden, yet meaningful insights. With large data becoming a norm, a lot of data mining algorithms are available that help in data mining. We have tried to compare two classification algorithms, primarily Decision trees and Random forest. A total of 10 datasets have been taken from UCI Repository and Kaggle and with the help of Konstanz Information Miner (KNIME) workflows, a comparative performance has been made pertaining to the accuracy statistics of Random Forest and decision Tree. The results show that Random Forest gives better and accurate results for a dataset as compared to decision trees.

Key Words: *Data Mining, classification, Decision Tree, Random Forest, KNIME, Accuracy statistics, Confusion Matrix.*

1. Introduction

Data, as is rightly said, is the new fuel of the present era. All important decisions are backed by a huge mass of information constituting the Big Data. With the Data Mining exercise, the necessary elements pertaining to a data set can be extracted based on the user requirements. Data Mining is thus a computer assisted knowledge discovery process used to predict the next move of the data based on certain patterns or behavior and thereby take crucial data-driven decisions (Rani & Gupta, 2019).The method follows both supervised and unsupervised learning (Kaur & Kaur, 2017) Data Classification is a supervised learning approach which organizes data in a way enabling easy retrieval (Agarwal, Panday, & Tiwari, 2012). Decision Tree and Random Forests are among the popular tools widely used across several programming languages to predict the data behavior. Both the methods work by taking a certain portion of data as a training data set and apply the prediction to the remaining data set.

2. Review of Literature

KNIME or Konstanz Information Learner is a GUI based analytical platform that works mainly by drag and drop of nodes which enables in establishing a flow of relations between them. Being an open source platform, it assists in exploration of a huge spectrum of databases and their resulting prediction on the basis of analysis (Sharma & Bansal, 2015). Nodes are the basic programming units, each of them performing definite tasks like- Read, Explore, Transform, Analyze & Deploy. Each of these functions are in turn represented by specific colors. The ports (input & output) attached to each of the nodes are joined to form workflow. The working status of each node is understood by three small traffic lights attached right below them (KNIME, n.d.) The KNIME repository comprises of 1000 nodes (Sharma & Bansal, 2015).

Decision Tree is a supervised Machine Learning technique which is used for both Classification and Prediction of data sets (Mesarić & Šebalj, 2016),however it is more effective for classification purpose. The decision map shows the set of possible outcomes if a particular decision is taken (Lucid Chart, n.d.). The working mechanism of Decision Tree algorithms involves nodes like- Root Node, Decision Node and Leaf Node, where all of the nodes are structured in a tree- like diagram.

Root Node- where the entire dataset is assigned with a view to solve a particular problem.

Decision Node-Specifying test on attributes which are formed by subdividing the Root Node.

Leaf Node- The final nodes indicating the values of target attributes.

With Pruning Method, the unnecessary nodes of the decision tree can be removed, so that more refined result is obtained and the errors are reduced (Ali, Khan, Ahmad, & Maqsood, 2012). Decision Trees can be used to generate online catalogs for e-commerce businesses, additionally it can be used in deriving probabilistic business models which can be used for effective CRM (Patel & Rana, 2014).

ROC (Receiving operating characteristics) curve is a probability curve and AUC (Area under Curve) represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. It is one of the most imperative method for evaluation of a dataset with respect to classification performance and is also known as Area under receiving operating characteristics(AUROC) An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact, it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity whatsoever (Narkhede, 2018)
Advantages of Decision Trees-

i.) User Friendly- Results derived from this method is easier to understand. They are more intuitive and hence simple to comprehend.

ii.) They are always better if the aim is exploratory analysis as they facilitate the understanding of data relationships in a tree like structure. They do not require much of data preprocessing.

iii.) It is always applicable when a dataset has a “feature” that is really significant to come up with an appropriate decision

Disadvantages of Decision Trees-

i.) Prone to over- fitting. Setting a maximum depth solves over-fitting but also introduces bias and hence accuracy is compromised (Prajwala, 2015).

ii.) It is over sensitive regarding irrelevant attributes in the training dataset (Rokach & Maimon, 2013).

iii.)Difficult to partition attributes containing continuous data (Rokach & Maimon, 2013).

Ideally Decision trees are preferred over Random forest when explainability between the variables is prioritized over accuracy. A decision tree can be more clearly visualized this enhances the explainability of a model, especially to the non-technical users.

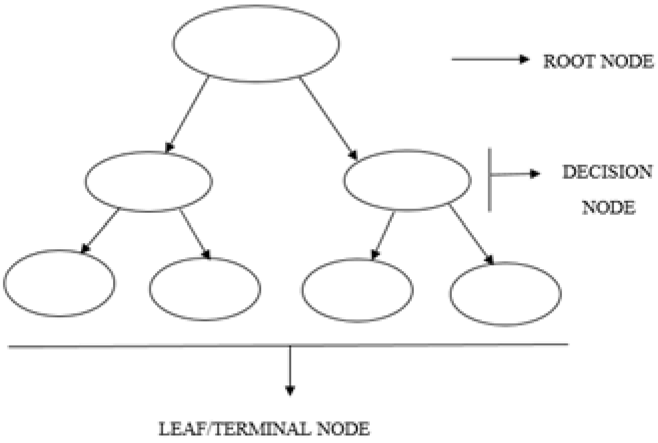


Figure 1: Decision Tree with Root Node, Decision Node and Terminal/Leaf Node

Random Forest is another supervised Machine Learning technique. As the name suggests it is a collection of decision trees resulting in ‘Forest’. While ‘Random’ suggests that each Decision Trees has an equal chance of getting selected as a sample (Ali et al., 2012) It uses Ensemble Learning classifier which uses decision tree algorithm in a randomized fashion. According to Briemen, Random Forest is more effective for large Databases, even when there are missing values, it maintains a fair degree of accuracy (Breiman, 2001).In the given Training dataset, the respective are jumbled even repeated in order to derive at Bootstrapped data. This Bootstrapped data will later be used in constructing several decision tree models.

Bootstrapped Data refers to the randomly chosen data records from the original dataset. The Variables from the Bootstrapped data then becomes the Root Node for the construction of Decision Trees, which are chosen randomly. By bootstrapping our dataset, a Random forest provides a more accurate and generalistic model.

Advantages of Random Forest

i.)Application of Bagging Method enhances the accuracy (Ali et al., 2012).

ii.) Overcomes the problem of overfitting of data. The testing performance of a Random forest does not decrease as the number of trees increases.

iii.) Construction of numerous Decision Tree models generates more accurate results (Prajwala, 2015).

Disadvantages of Random Forest- Certain shortcomings derived from the review (Kaggle, n.d.)are-

i) Select the "features" randomly to build the trees and sometimes ignores the most important "feature" to take a decision

ii) They require data preprocessing and assumptions of data distribution.

iii) If speed is critical and accuracy can be traded off, they aren't favorable.

Thus when the data is heavily biased, Random forest will help in overcoming the issues related to over-fitting. It is always preferred when accuracy is prioritized over explainability.

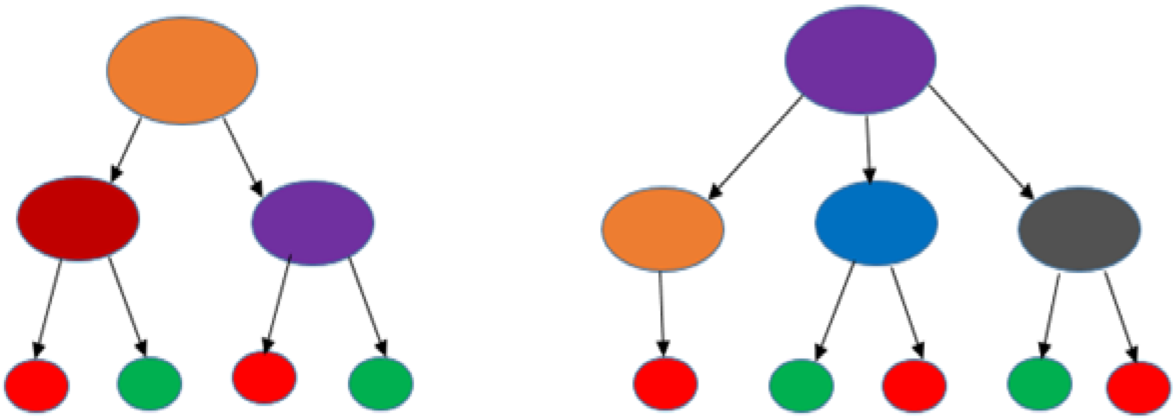


Figure 2: Random Forests

3. Objective

To do a comparative performance of Decision Tree and Random Forest as classification algorithms with respect to prediction in terms of accuracy statistics using KNIME.

4. Methodology

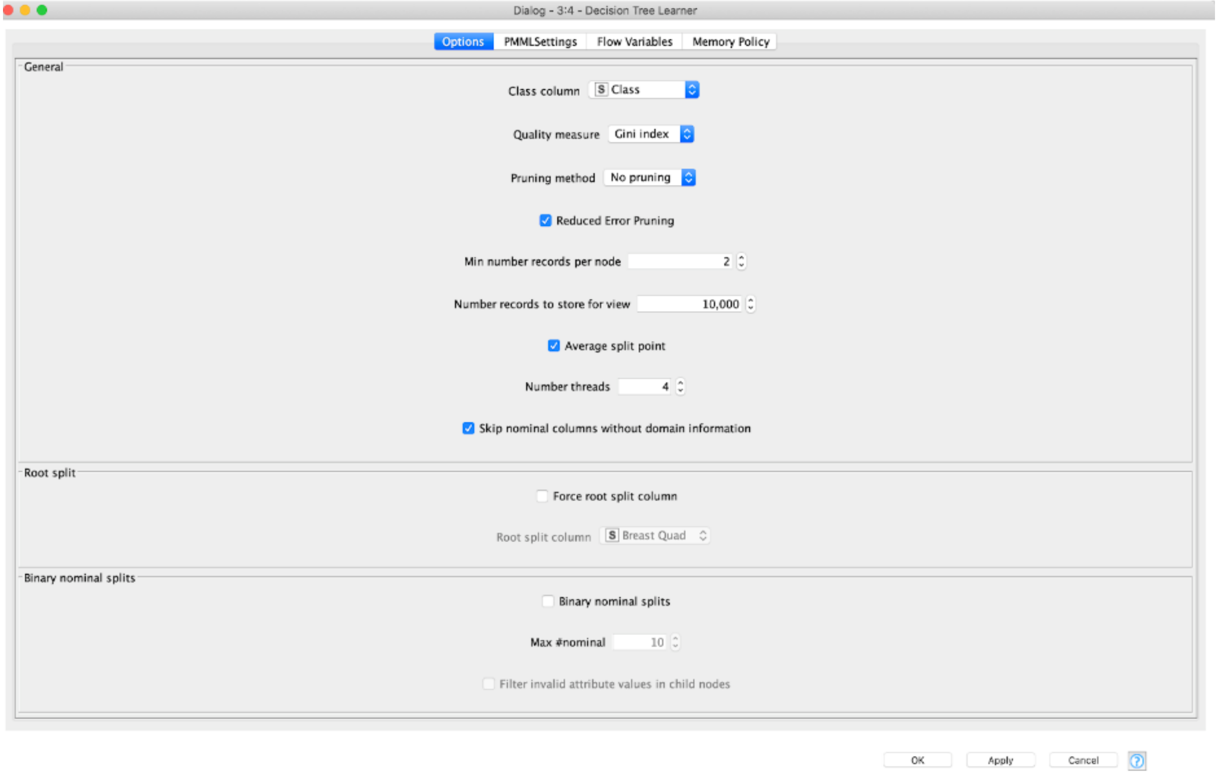
In order to understand the classification performance of Random Forest and Decision trees using KNIME, we have taken small and large datasets from UCI Machine Learning Repository and Kaggle. The objective of the classification performance of Random forest and Decision tree is to come up with a comparative performance analysis using Accuracy statistics, Confusion Matrix, Precision, Recall, F-measure, Cohen's Kappa Value and area under the ROC curve.

A total of 10 datasets with varying instances, type of datasets and attributes were taken for the study and the same are shown in Table 1.

The Configuration parameter settings for Decision Tree and Random forest are depicted in Fig3 and Fig 4 Respectively

Table 1- Datasets, Instances and attributes

SL. NO.	DATASET	NUMBER OF INSTANCES	NUMBER OF ATTRIBUTES	MISSING VALUES	TYPE OF DATASET	TYPE OF FILE
1.	Parkinson	197	23	No	Multivariate	Csv
2.	Breast Cancer	286	9	Yes	Multivariate	Xls
3.	Diabetic Retinopathy Detection	1151	20	No	Multivariate	Csv
4.	Echocardiogram	132	12	Yes	Multivariate	Csv
5.	Chronic Kidney	400	25	No	Multivariate	Csv
6.	Iris	150	4	No	Multivariate	Xls
7.	Wine	178	13	No	Multivariate	Xls
8.	Orthopedic Patients	310	7	No	Multivariate	Csv
9.	Mushroom	8124	23	Yes	Multivariate	Csv
10.	Glass	214	10	No	Multivariate	Csv



General

Class column: S Class

Quality measure: Gini index

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 2

Number records to store for view: 10,000

☑ Average split point

Number threads: 4

☑ Skip nominal columns without domain information

Root split

□ Force root split column

Root split column: S Breast Quad

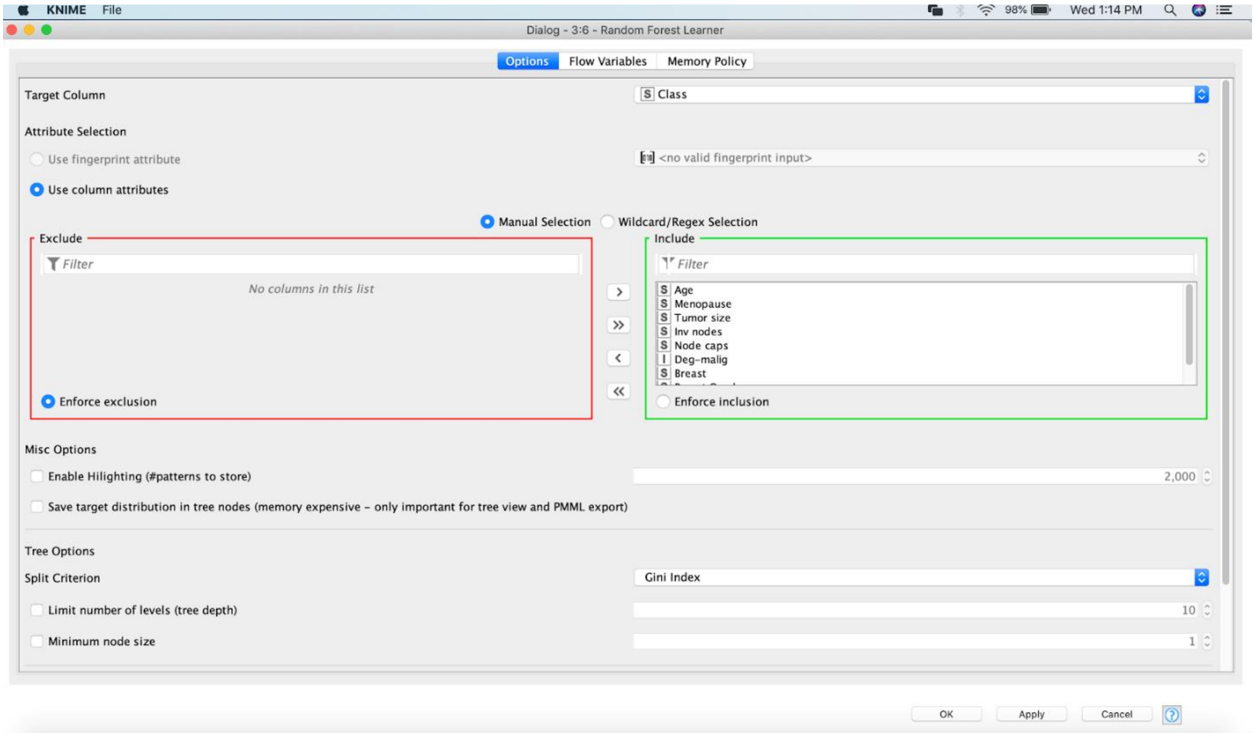
Binary nominal splits

□ Binary nominal splits

Max #nominal: 10

□ Filter invalid attribute values in child nodes

Figure 3- Configuration Parameter settings for Decision Tree



Target Column: S Class

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude

● Enforce exclusion

Include

S Age
S Menopause
S Tumor size
S Inv nodes
S Node caps
I Deg-malig
S Breast

○ Enforce inclusion

Misc Options

□ Enable Hilighting (#patterns to store) 2,000

□ Save target distribution in tree nodes (memory expensive – only important for tree view and PMML export)

Tree Options

Split Criterion: Gini Index

□ Limit number of levels (tree depth) 10

□ Minimum node size 1

Figure 4- Configuration Parameter settings for Random forest

In order to evaluate the performance of the algorithms, a detailed data methodology was deployed comprising processing the dataset, data sorting and implementation of classification algorithms for the purposes of training and testing. For the sake of uniformity, we have partitioned the data where 80% of the dataset has been used for training the algorithm while 20% was applied for evaluating the trained model. The sampling technique chosen is stratified sampling. The datasets were imported into the model by using an Excel reader, File reader or Table reader depending on the file format. The KNIME workflows in Fig 5 shows the data pipeline from the first stage(input) where the data is imported into the model either as an Excel reader or File reader, is then pre-processed for the purposes of data mining algorithms and finally the output comprises of Scorer, scatter plot, Excel writer or File writer and PMML writer.

5. Analysis and Interpretations

Table 2 shows the comparative performance of Random Forest and Decision tree. It depicts the correctly classified and incorrectly classified instances for Random forest and decision tree classifiers. The classification results for the 10 datasets reveal that irrespective of the number of instances and attributes for each data set, Random Forest has shown better and higher accuracy results as compared to Decision trees. Random forest ends up taking an average of various responses given by an individual model and hence it results in overcoming any limitations that might occur owing to over-fitting problems generated due to missing values. Also since the accuracy for Random Forest classification is higher for all the data sets, consequently it shows the Precision, Recall, F- measure and Cohen Kappa (k) values to be higher as compared to those of Decision trees.

Table 2: Comparative Performance of Random Forest and Decision Tree

SL.NO.	DATASET	RANDOM FOREST		DECISION TREE	
		Correctly Classified	Incorrectly Classified	Correctly Classified	Incorrectly Classified
1.	Parkinson	97.43%	2.56%	89.74%	10.25%
2.	Breast Cancer	79.31%	20.69%	71.93%	20.69%
3.	Diabetic Retinopathy Detection	70.13%	29.87%	63.63%	36.36%
4.	Echocardiogram	100%	0%	93.33%	6.67%
5.	Chronic Kidney	98.75%	1.25%	96.25%	3.75%
6.	Iris	93.33%	6.67%	90%	10%
7.	Wine	100%	0%	94.44%	5.56%
8.	Orthopedic Patients	82.30%	17.74%	82.25%	17.74%
9.	Mushroom	100%	0%	100%	0%
10.	Glass	79.07%	20.93%	60.46%	39.35%

On the basis of the confusion matrix generated by each predictive data mining algorithm, we have determined the accuracy, Cohen kappa, and Precision, recall and F- measure values. The same are shown in Table 3. The Accuracy is the percentage of the correct predictions made by the model with respect to the total number of predictions.

On the basis of the following 4 values generated by Confusion Matrix, i.e. True Positive (TP)- the correctly predicted positive sample; True Negative (TN)- the correctly predicted negative samples; False Positive(FP) and False Negative (FN)- the incorrectly predicted positive and negative samples respectively, the performance measures were computed.

Precision is the number of correctly classified samples out of the total samples classified in that particular class. The same is mathematically computed as TP/TP+FP

Recall is the number of correctly classified samples out of the total samples that are truly in that particular class. The same is mathematically computed as TP/TP+FN

F-measure is the harmonic mean of recall and precision and is computed as (2*Precision*Recall)/Precision + Recall

As visible in Table 3, higher is the accuracy, higher are the Precision, Recall and F-measure values. Whereas lower is the accuracy, lower are the Precision, Recall and F-measure values. Nearly equal values of Precision, Recall and F- measure indicate micro-averages of the same and the fact that the models are somehow balanced. By balanced, it indicated that it has the ability to correctly classify positive samples is the same as its ability to correctly classify negative samples (Stack Exchange, n.d.).

Cohen's Kappa is a measure that takes class distribution into account and can be used in conjunction with accuracy. It handles both multiclass and imbalanced class problems.

Cohen Kappa(k) usually ranges between less than or equal to 1. Values of 0 or less, indicate that the classifier is useless. There isn't a standardized way to interpret its values. Landis and Koch (1977) provide a way to characterize values. According to their scheme a value less than 0 implies no agreement, 0-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial and 0.91-1 as almost perfect.(The Data Scientist, 2016)

Table 3: Precision, Recall, F- measure and Cohen Kappa for Random Forest and Decision Tree

SL.NO	DATASET	RANDOM FOREST				DECISION TREE			
		Precision	Recall	F-measure	Cohen Kappa(k)	Precision	Recall	F-measure	Cohen Kappa(k)
1.	Parkinson	0.974	0.974	0.973	0.93	0.897	0.897	0.897	0.748
2.	Breast Cancer	0.793	0.793	0.792	0.443	0.719	0.719	0.719	0.33
3.	Diabetic Retinopathy Detection	0.70	0.70	0.70	0.402	0.636	0.636	0.635	0.278
4.	Echocardiogram	1	1	1	1	0.933	0.933	0.932	0.842
5.	Chronic Kidney	0.987	0.987	0.987	0.973	0.962	0.962	0.961	0.919
6.	Iris	0.933	0.933	0.933	0.9	0.9	0.9	0.9	0.85
7.	Wine	1	1	1	1	0.94	0.94	0.93	0.916
8.	Orthopedic Patients	0.823	0.823	0.822	0.566	0.822	0.822	0.821	0.577
9.	Mushroom	1	1	1	1	1	1	1	1
10.	Glass	1	1	1	1	0.604	.604	.604	0.461

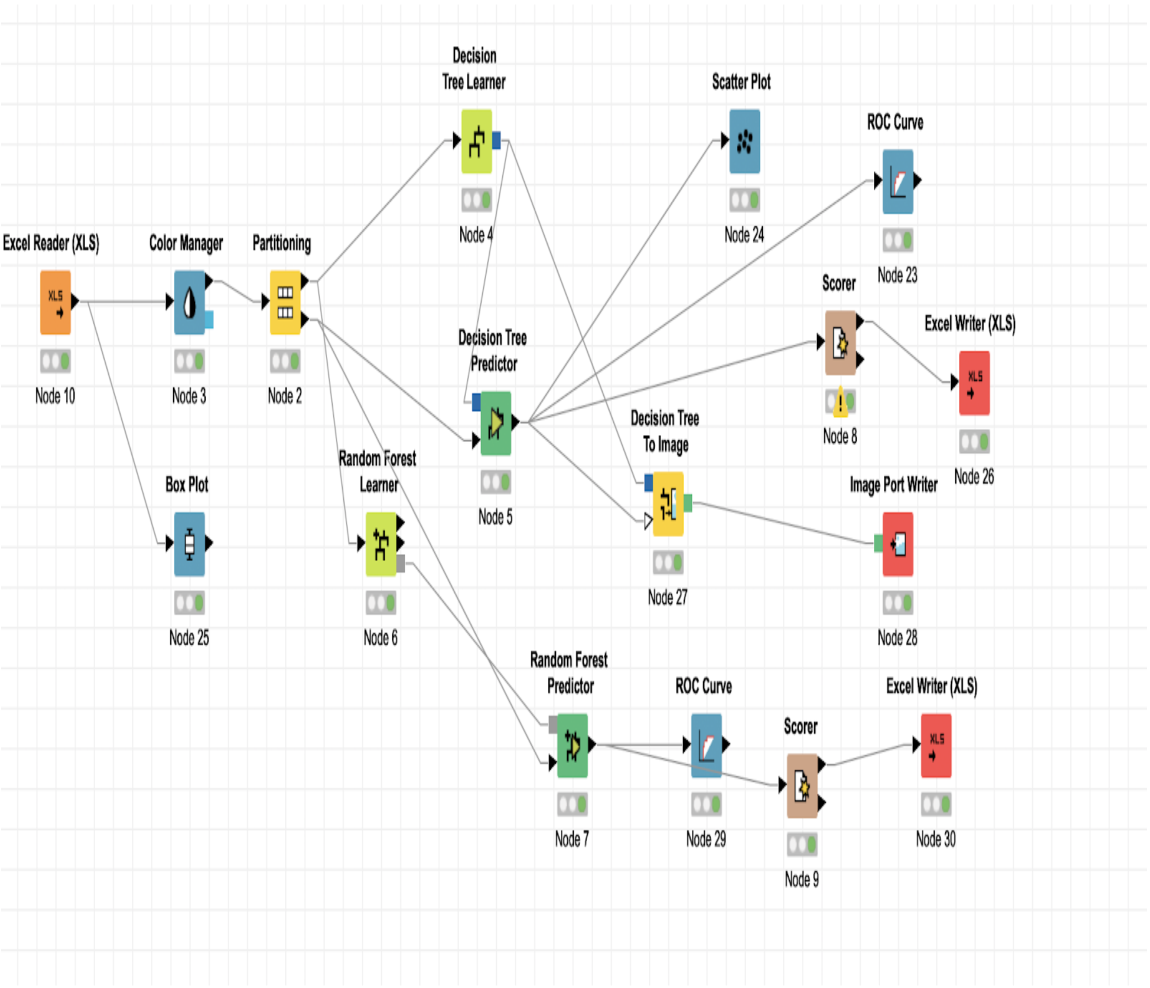


Figure 5- KNIME workflow showing the data pipeline from Input to output

The KNIME workflows in Fig 5 shows the data pipeline from the first stage(input) where the data is imported into the model either as an Excel reader or File reader, is then pre-processed for the purposes of data mining algorithms and finally the output comprises of Scorer, scatter plot, Excel writer or File writer and PMML writer.

The figures are for the Breast Cancer dataset, a multivariate data comprising 286 instances and 9 attributes with missing values. It has two classes which includes No recurrence events and recurrence events and the association task is that of classification.
Fig 6 shows the Decision tree view that we get for the Breast Cancer Dataset. The same is generated after the configuration settings of the Decision Tree Learner node.
Figure 7 and Fig 8 show the scatter plot view and Confusion matrix respectively. A scatter plot is one of the most significant data visualization technique available on the KNIME platform and helps to interactively visualize the relationship between the two columns in the data set. Also the visualization is enhanced with the help of color manager node that helps in assigning colors. In this dataset, Green color represents non- recurrence events while red color represents recurrence events.
A Confusion matrix is a concise representation of the performance of the model and provides us the values pertaining to True Positive, true negatives, false positives and false negatives. In the breast cancer dataset that we have shown here, Random Forest shows an accuracy of 79.31% while for decision tree we have that of 71.93% thereby indication a better accuracy shown by Random Forest as a classification algorithm.

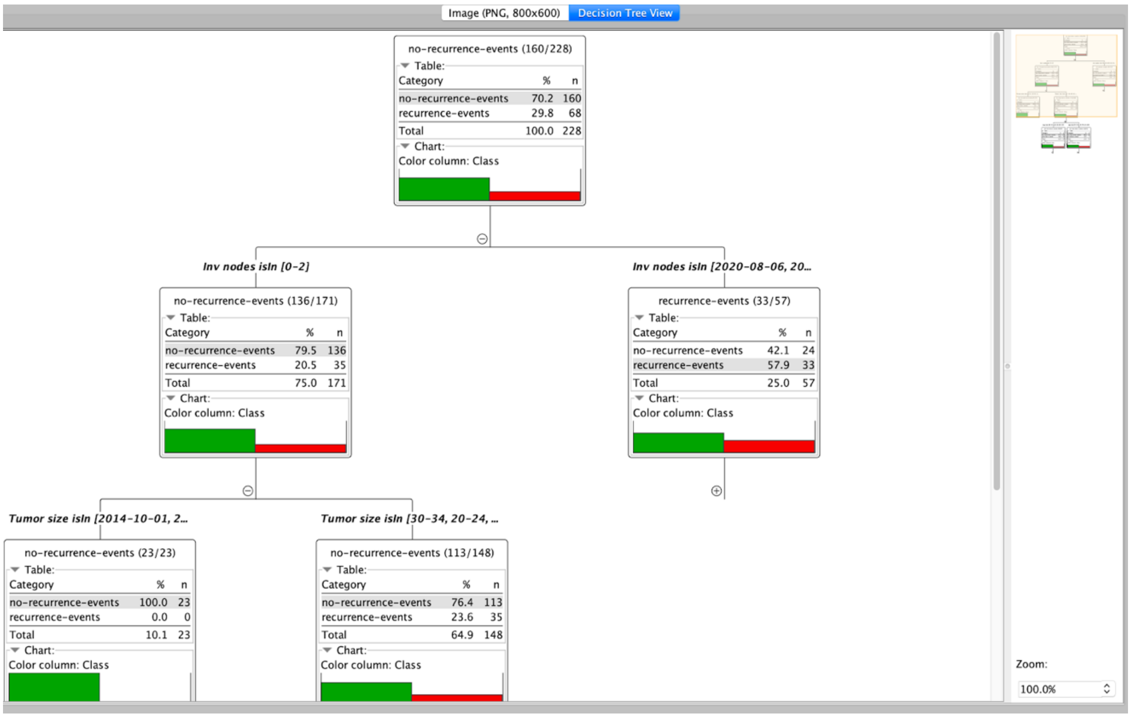


Image (PNG, 800x600)
Decision Tree View
Inv nodes isIn [0-2]
Inv nodes isIn [2020-08-06, 20...
Tumor size isIn [2014-10-01, 2...
Tumor size isIn [30-34, 20-24, ...
Chart:
Color column: Class
Zoom:
100.0%

no-recurrence-events (160/228)

Category	%	n
no-recurrence-events	70.2	160
recurrence-events	29.8	68
Total	100.0	228

no-recurrence-events (136/171)

Category	%	n
no-recurrence-events	79.5	136
recurrence-events	20.5	35
Total	75.0	171

recurrence-events (33/57)

Category	%	n
no-recurrence-events	42.1	24
recurrence-events	57.9	33
Total	25.0	57

no-recurrence-events (23/23)

Category	%	n
no-recurrence-events	100.0	23
recurrence-events	0.0	0
Total	10.1	23

no-recurrence-events (113/148)

Category	%	n
no-recurrence-events	76.4	113
recurrence-events	23.6	35
Total	64.9	148

Figure 6- Decision Tree View

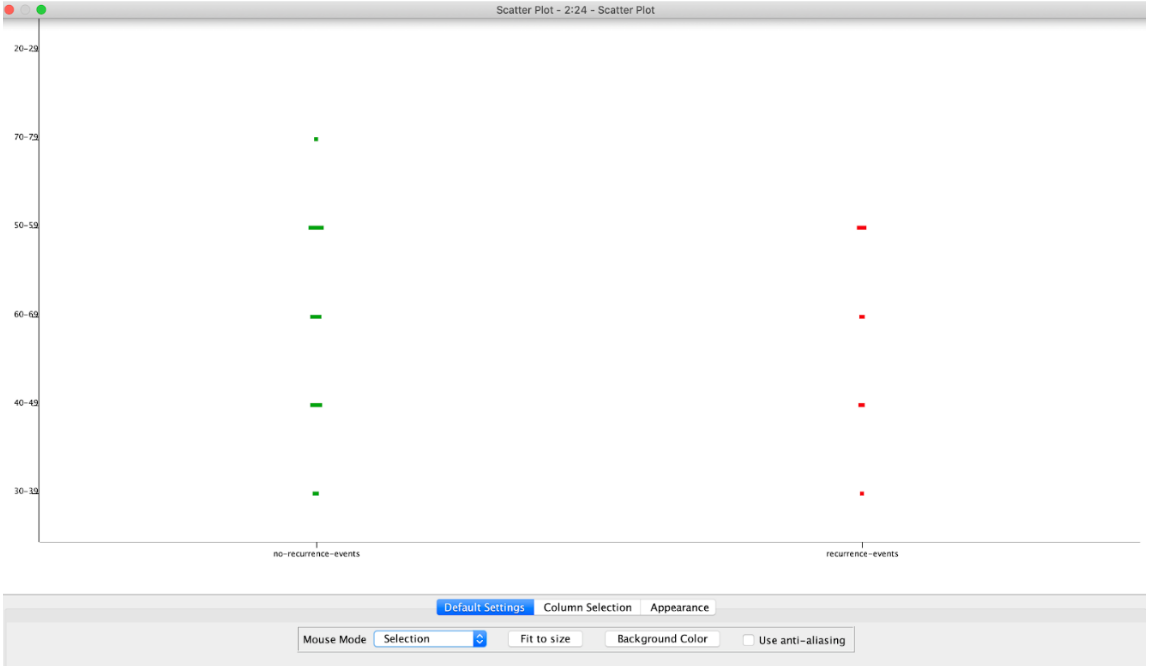
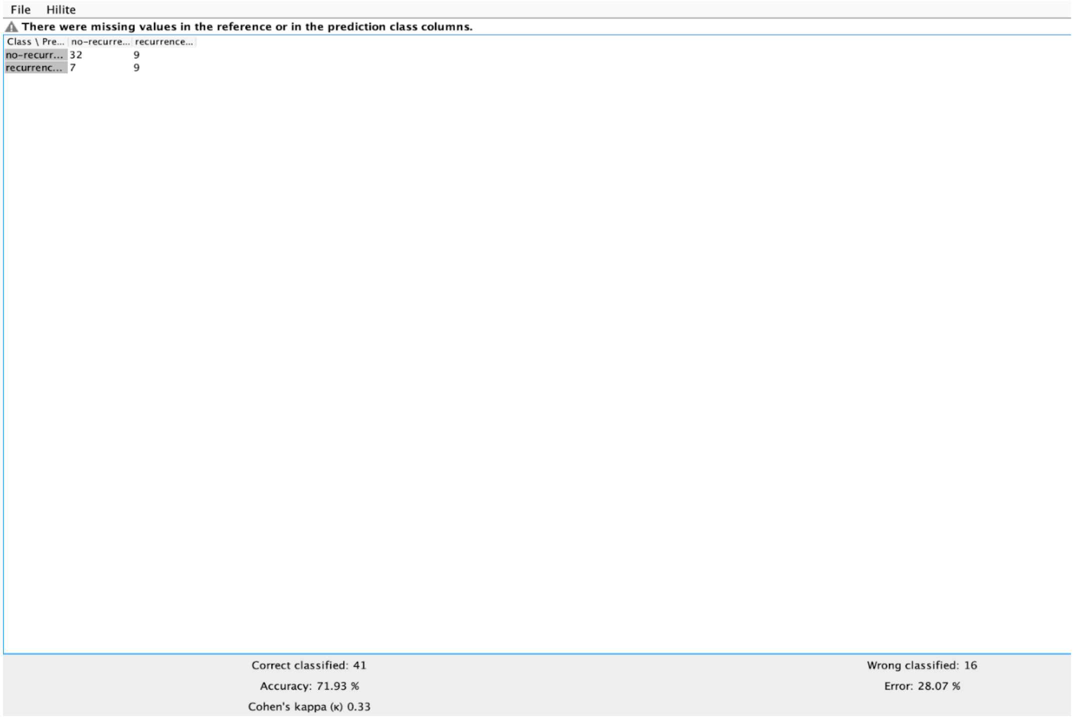


Figure 7- Scatter Plot View



⚠ There were missing values in the reference or in the prediction class columns.

Class \ Pre...	no-recurre...	recurrence...
no-recurr...	32	9
recurrenc...	7	9

Correct classified: 41
Accuracy: 71.93 %
Cohen's kappa (κ) 0.33

Wrong classified: 16
Error: 28.07 %

Figure 8 - Confusion Matrix view

6. Conclusion

Thus from the results of the classification performance for the 10 datasets taken for the study, we may conclude that the comparative performance for random forest is better as compared to decision tree and it yields more accurate and precise results. Hence the same is highly recommended as a performance evaluation model. However further studies are required that helps in understanding the way performance of the model can be enhanced and improved.

References

[1] Agarwal, S., Panday, G. N., & Tiwari, M. D. (2012). Data Mining in Education: Data Classification and Decision Tree Approach. *International Journal of E-Education, e-Business, e-Management and e-Learning*, *2*(2), 140–144. https://doi.org/10.7763/ijeeee.2012.v2.97.

[2] Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, *9*(5), 272–278.

[3] Breiman, L. (2001). Random Forest. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1017/CBO9781107415324.004.

[4] Kaggle. (n.d.). When Would you prefer Decision Tree?.

[5] Kaur, S., & Kaur, H. (2017). Review of Decision Tree Data mining Algorithms : CART and C4 . 5. *International Journal of Advanced Research in Computer Science*, *8*(4), 436–439.

[6] KNIME. (n.d.). Cheat Sheet : Control and Orchestration with KNIME Analytics Platform. Retrieved from https://www.knime.com/sites/default/files/07252019_CheatSheet_Advanced_A4_Control_Orchestration.pdf.

[7] Lucid Chart. (n.d.). *What is a Decision Tree Diagram*. Retrieved from https://www.lucidchart.com/pages/decision-tree.

[8] Mesarić, J., & Šebalj, D. (2016). Decision trees for predicting the academic success of students. *Croatian Operational Research Review*, *7*(2), 367–388. https://doi.org/10.17535/crorr.2016.0025.

[9] Narkhede, S. (2018). Understanding AUC - ROC Curve. Retrieved from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.

[10] Patel, B. R., & Rana, K. K. (2014). A Survey on Decision Tree Algorithm For Classification. *International Journal of Engineering Development and Research*, *2*(1), 1–5.

[11] Prajwala, T. (2015). A Comparative Study on Decision Tree and Random Forest Using R Tool. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(1), 196–199. https://doi.org/10.17148/ijarcce.2015.4142.

[12] Rani, H., & Gupta, G. (2019). Prediction Analysis Techniques of Data Mining: A Review. *International Journal of Computer Science and Mobile Computing*, *8*(5), 15–22. https://doi.org/10.2139/ssrn.3350303.

[13] Rokach, L., & Maimon, O. (2013). Decision Tress. *DATA MINING AND KNOWLEDGE DISCOVERY HANDBOO*. https://doi.org/10.4018/978-1-60960-557-5.ch006.

[14] Sharma, N., & Bansal, K. (2015). Performance of Data Mining Tools: A Case Study Based on Classification Algorithms and Datasets. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*(6), 363–370.

[15] Stack Exchange. (n.d.). Cross Validated. Retrieved from https://stats.stackexchange.com/questions/99694/what-does-it-imply-if-accuracy-and-recall-are-the-same/99697.

[16] The Data Scientist. (2016). Performance Measures: Cohen's Kappa statistic. Retrieved from https://thedatascientist.com/performance-measures-cohens-kappa-statistic/.