

# Predicting probable bankruptcy based on various data attributes.

Hassan M. Mahmudul ID: 15-28405-1 Serial No: 01 Course: Data Warehousing and Data Mining [A]
Fall 2019-2020
m.mahmud77@outlook.com

**Abstract**— Data Mining, discovers and extracts useful patterns from large set of data to find observable patterns. This paper been worked with supervised data [1]. All the data are considered for training purpose, and it is used in the five-classification algorithm, this paper presents the analyze and accuracy of probable bankruptcy based on seven different nominal attribute using data mining algorithms for various decision tree approaches using WEKA. Five classification algorithms such as J48, Random Tree (RT), NaiveBayes, REPTree and Random Forest (RF) are used to measure the accuracy. Data mining tool WEKA (Waikato Environment for Knowledge Analysis) has used for performing such five classification algorithms, RandomForest algorithm outperforms other algorithms by yielding an accuracy of 100%.

**Index Terms**—Data mining, bankruptcy, weka, random tree, J48, RandomForest.

## I. INTRODUCTION

Currently Bangladesh has Tk 99,370 crore of defaulted amount, according to data from the Bangladesh Bank [2]. It has grown above the alarming rate. The collected dataset is the actual dataset of banks from India. If we try to collect similar data and use such decisions, then it would be easy to predict any bankruptcy for any the current loan amounts. Weka tool has been used to generate decision trees based on different algorithms. Decision tree generated results and decision tree been added on this paper as well.

## II. RELATED BACKGROUND

Supervised learning is the machine learning task of learning function that's maps on input to output based on example input to output pairs. The calculations can apply straightforwardly to a dataset. Weka is an open-source information mining apparatus it bolsters information-mining calculations, packing, and boosting. Enlistment is the gaining from class named preparing tuples. In choice tree hubs speak to the information esteems, the edges will highlight all the potential moves, therefore from hub to leaf through the edge it's giving the objective qualities from which we can make order to anticipate. Some supervised

learning algorithm required the user to determine certain control parameters.

## III. PROPOSED METHODOLOGY

The dataset contains 7 attributes including 250 number of instances containing information on industry risk, management risk, financial Flexibility, credibility, competitiveness, operating Risk. Some sample data is given below, also used data set's url is also attached [3]. Based on the qualitative data parameters from experts, attempt is done to predict possible bankruptcy. Below algorithms are applied to classify correct instances

- RandomTree
- RandomForest
- J48
- NaiveBayes
- REPTree

## IV. RESULT AND ANALYSIS

### 1. RandomTree

Choose: REPTree-M 2-V0.001-N3-S1-L-1-10.0

test options

☐ Use training set  
☐ Supplied test set  
☒ Cross-validation Folds: 10  
☐ Percentage split % 66  
More options...

(Nom) Class

Start Stop

result list (right click for options)

- 01:09:51 - trees RandomTree
- 01:11:32 - trees RandomForest
- 01:11:57 - trees J48
- 01:12:17 - bayes NaiveBayes
- 01:12:41 - trees REPTree

Classifier output

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

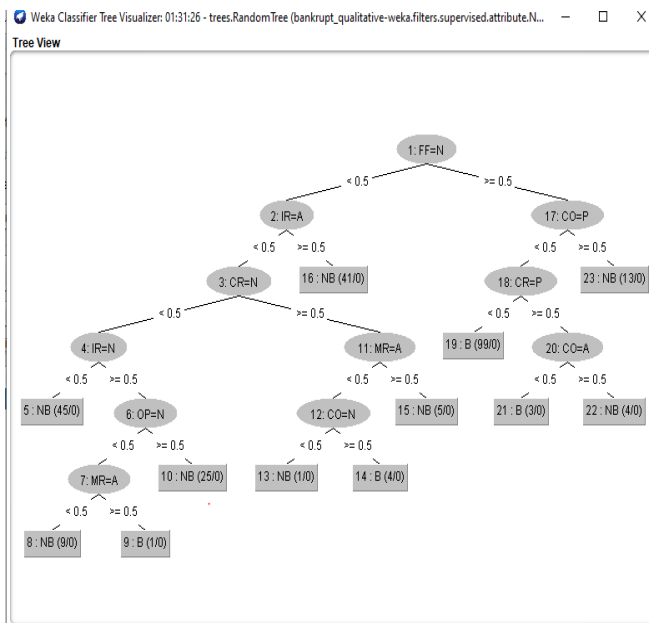
Correctly Classified Instances	249	99.6 %
Incorrectly Classified Instances	1	0.4 %
Kappa statistic	0.9918	
Mean absolute error	0.004	
Root mean squared error	0.0632	
Relative absolute error	0.0167 %	
Root relative squared error	12.7803 %	
Total Number of Instances	250	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
Weighted Avg.	0.993	0.000	1.000	0.993	0.996	0.992	0.997	0.991	B
	0.996	0.003	0.996	0.996	0.996	0.992	0.997	0.994	NB

=== Confusion Matrix ===

a	b	<-- classified as	
107	0	a = B	
1	142	b = NB	



Choose REPTree-M2-V0.001-N3-S1-L-1-1-0.0

Test options

Use training set  
Supplied test set  
Cross-validation Folds 10  
Percentage split % 66

Classifier output

Time taken to build model: 0.02 seconds

Stratified cross-validation Summary

Correctly Classified Instances	249	99.2 %	
Incorrectly Classified Instances	2	0.8 %	
Kappa statistic	0.9837		
Mean absolute error	0.008		
Root mean squared error	0.0894		
Relative absolute error	1.6333 %		
Root relative squared error	18.0741 %		
Total Number of Instances	250		

Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
B	1.000	0.014	0.982	1.000	0.991	0.984	0.993	0.982	B
NB	0.986	0.000	1.000	0.986	0.993	0.984	0.993	0.994	NB
Weighted Avg.	0.992	0.006	0.992	0.992	0.992	0.984	0.993	0.989	

Confusion Matrix

	a	b	← classified as
107	0	1	a = B
2	141	1	b = NB

## Analysis

Correctly Classified instance = 249, 99.6%

Incorrectly Classified instance = 1, 0.4%

## 2. RandomForest

Classifier

Choose REPTree-M2-V0.001-N3-S1-L-1-1-0.0

Test options

Use training set  
Supplied test set  
Cross-validation Folds 10  
Percentage split % 66

Classifier output

Time taken to build model: 0.21 seconds

Stratified cross-validation Summary

Correctly Classified Instances	250	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0.0103	
Root mean squared error	0.0456	
Relative absolute error	2.0908 %	
Root relative squared error	9.2204 %	
Total Number of Instances	250	

Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
B	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	B
NB	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	NB
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	

Confusion Matrix

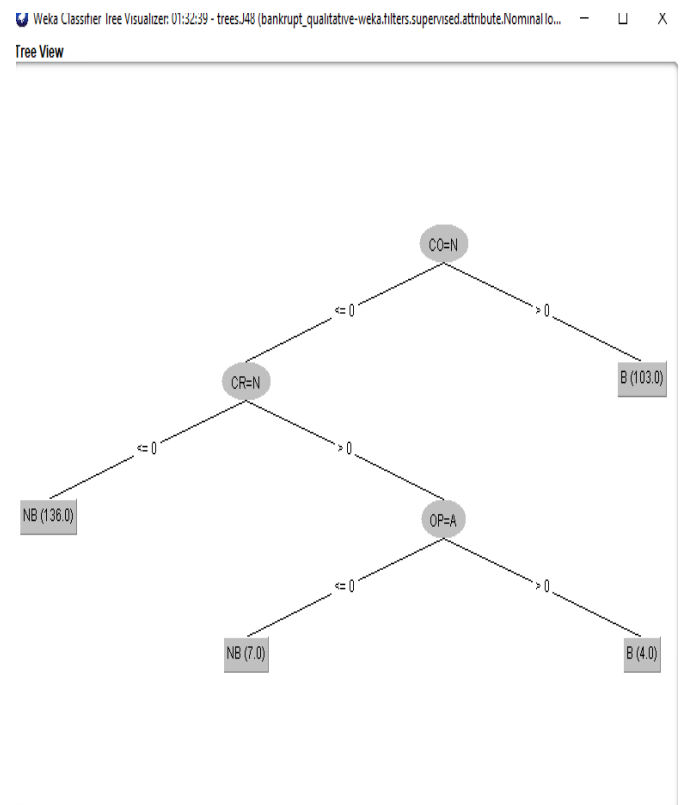
	a	b	← classified as
107	0	1	a = B
0	143	1	b = NB

## Analysis

Correctly Classified instance = 250, 100%

Incorrectly Classified instance = 0, 0%

## 3. J48

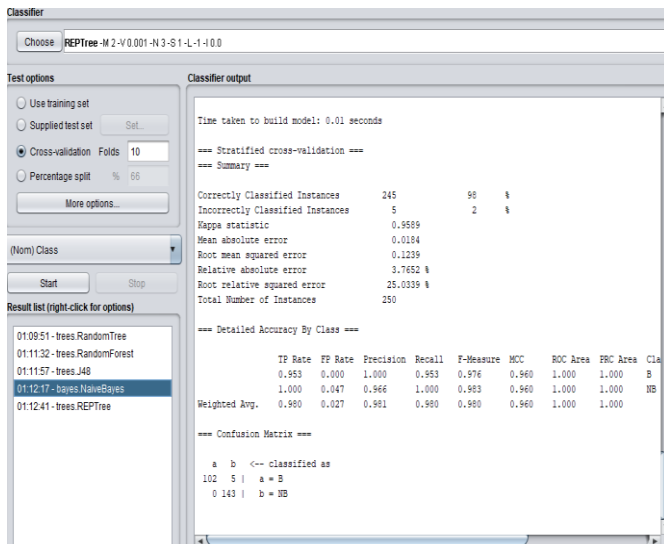


## Analysis

Correctly Classified instance = 248, 99.2%

Incorrectly Classified instance = 2, 0.8%

## 4. NaiveBayes

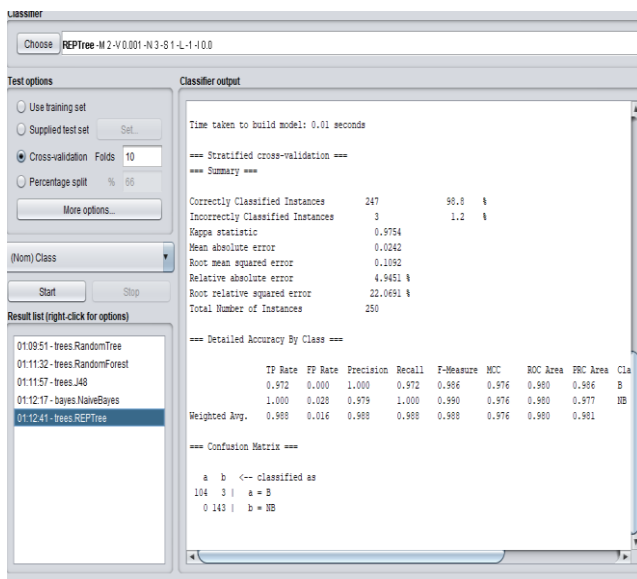


## Analysis

Correctly Classified instance = 245, 98%

Incorrectly Classified instance = 5, 2%

## 5. REPTree



## Analysis

Correctly Classified instance = 247, 98.8%

Incorrectly Classified instance = 3, 1.2%

## Result Analysis: (Correctly classified instances)

- RandomTree: 99.6%
- RandomForest: 100%
- J48: 99.2%
- NaiveBayes: 98%
- REPTree: 98.8%

## V. CONCLUSION

As we examined the tree calculations, we can reach resolution that for credit informational index RandomForest tree is most appropriate for basic leadership as it is giving 100 percent of accuracy. In future, the proposed technique will be stretched out to other informational indexes from the regions like banking, and financial exchange and so on.

## REFERENCES

- [1] <https://blogs.nvidia.com/blog/2018/08/02/supervised-supervised-learning/>
- [2] <https://www.thedailystar.net/business/banking/bangladesh-bank-moves-amend-bankruptcy-act-1997-1698106>.
- [3] [https://archive.ics.uci.edu/ml/datasets/qualitative\\_bankruptcy](https://archive.ics.uci.edu/ml/datasets/qualitative_bankruptcy).