

**PAPER REVIEW**

**COMPARISON BETWEEN GOOD AND BAD DEBT CREDIT RESULT USING WEKA J-48, RANDOM FOREST AND RANDOM TREE**

COURSE CODE : MANB1143

COURSE NAME : BUSINESS INTELLIGENCE

PROGRAMME : MSc (BIA)

PREPARED BY : AHMAD NIZAM BIN MOHD LAZIM - MAN171058

AHMAD SHAKIR BIN AHMAD ZULKIFLI - MAN171020

AMJAD ALASSEH - MAN171009

MUHAMMAD NIDZAM MASO’OD - MAN171043

SUBMITTED TO : DR AZIZUL AZIZAN

Comparison between good and bad debt credit result using Weka J-48, Random Forest and Random Tree

Ahmad Nizam Mohd Lazim, Ahmad Shakir Bin Ahmad Zulkifli, Amjad Alasseh , Muhammad Nidzam Mas’oud

Advanced Informatics School, Universiti Teknologi Malaysia, Jalan Semarak, Kuala Lumpur, Malaysia

nizam1973@graduate.utm.my, ahmadshakir@graduate.utm.my, alassehamjad@graduate.utm.my, nidzam@graduate.utm.my

Abstract:

Classification is widely used in banking industry to do risk assessment on loan credit risk. In this paper, we will be discussing on three common classifiers, which are Weka J-48, Random Forest and Random Tree. We will apply these three classifiers on the credit-g data set from OpenML site [1]. For this paper, we will be using WEKA for our classification machine learning purposes. This paper then will look into the accuracy result of each classifier and do the comparison between them. In the conclusion, we will give suggestion on which algorithm that we think is the best for loan credit risk assessment.

Keyword: Weka J-48, Random Forest, Random Tree, WEKA, banking

1. Introduction

Machine learning is becoming increasing popular nowadays. With the advances of computing power, once an exclusive way of doing prediction or forecasting can now be perform with ease from your laptop. So what is actually machine learning? Machine learning is method of forecasting future behaviours, outcomes and trends using existing data and data science technique in which computers can learn to do this without being explicitly programmed [2].

In banking sectors, machine learning can be apply in marketing, risk management, and customer acquisition and retention. For marketing purpose, machine learning is use to improve marketing techniques and advertisement targeting to potential customer. Examples of application for marketing are classifying respondent on previous advertisement campaigns, customer profiling and customer attrition trend. In risk management, machine learning is use to do loan default prediction, high-risk loan detection, classifying profile of high profitable loans and detecting credit card fraud. Machine learning can help to increase customer acquisition and retention by predicting the needs of customer’s base on the financial pattern and classifying the criteria of a loyal customer. Using this information, bank can build a good relationship with the customer.

There are three core types of machine learning: supervised learning, unsupervised learning and reinforced learning. Supervised learning is use to predict the outcome when the data are labelled. Unsupervised learning is use to find hidden structure in the data when the data are not labelled. The last type of machine learning is reinforcement learning which is use to learn series of actions in a decision process.

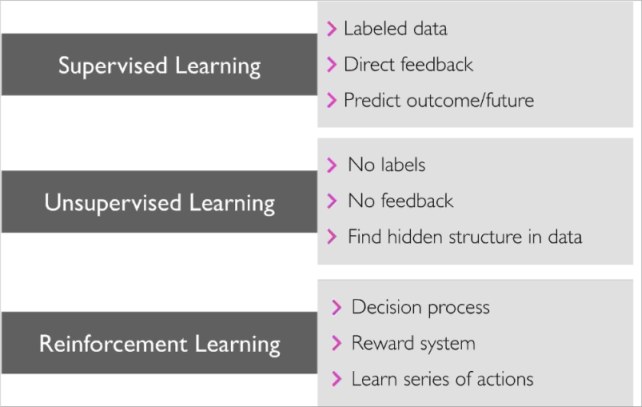


Figure 1. Machine Learning Type[3]

Classification is a subcategory of supervised learning where the goal is to predict the categorical class labels of new instances, based on past observations. Those class labels are discrete, unordered values that can be understood as the group memberships of the instances [3]. In this paper, we will apply three different classifier algorithm to our selected data and try to predict whether a loan credit risk assessment is good or bad. The algorithm that we will be using are Weka J-48, Random Forest and Random Tree. In the following chapter, we will explain details on these algorithm and comparison of result that can be observe when we run these three algorithm to our data.

1. About the dataset

The data set that we will be using in this paper was prepared by Professor Dr. Hans Hofmann from Institut f"ur Statistik und "Okonometrie Universit"at Hamburg. This dataset classifies people with good or bad credit risks based on a set of attributes. Below are the details of the data and its attributes

Table 1 : Dataset details

|  |  |
| --- | --- |
| Items | Details |
| Data Set Characteristic | Multivariate |
| Attribute Characteristics | Categorical, Integer |
| Associated Tasks | Classification |
| Number of Instances | 1000 |
| Number of Attributes | 21 |
| Missing Values? | N/A |
| Area | Financial |
| Date Donated | 1994-11-17 |

Table 2 : Attributes details

|  |  |  |  |
| --- | --- | --- | --- |
| Attributes | Description | Type | Values |
| checking\_status | Status of existing checking account, in Deutsche Mark | nominal | 4 unique values 0 missing |
| Duration | Duration in months | numeric | 33 unique values 0 missing |
| credit\_history | Credit history (credits taken, paid back duly, delays, critical accounts) | nominal | 5 unique values 0 missing |
| Purpose | Purpose of the credit (car, television,...) | nominal | 10 unique values 0 missing |
| credit\_amount | Credit amount | numeric | 921 unique values 0 missing |
| savings\_status | Status of savings account/bonds, in Deutsche Mark. | nominal | 5 unique values 0 missing |
| employment | Present employment, in number of years. | nominal | 5 unique values 0 missing |
| installment\_commitment | Installment rate in percentage of disposable income | numeric | 4 unique values 0 missing |
| personal\_status | Personal status (married, single,...) and sex | nominal | 4 unique values 0 missing |
| other\_parties | Other debtors / guarantors | nominal | 3 unique values 0 missing |
| residence\_since | Present residence since X years | numeric | 4 unique values 0 missing |
| property\_magnitude | Property (e.g. real estate) | nominal | 4 unique values 0 missing |
| Age | Age in years | numeric | 53 unique values 0 missing |
| other\_payment\_plans | Other installment plans (banks, stores) | nominal | 3 unique values 0 missing |
| Housing | Housing (rent, own,...) | nominal | 3 unique values 0 missing |
| existing\_credits | Number of existing credits at this bank | numeric | 4 unique values 0 missing |
| Job | Job | nominal | 4 unique values 0 missing |
| num\_dependents | Number of people being liable to provide maintenance for | numeric | 2 unique values 0 missing |
| own\_telephone | Telephone (yes,no) | nominal | 2 unique values 0 missing |
| foreign\_worker | Foreign worker (yes,no) | nominal | 2 unique values 0 missing |

1. Tools

We are using WEKA to do our data mining process. WEKA is a program that contain various of algorithms of machine learning to do data mining operations. The application writing in java langue and build in with all algorithms that can use it on your dataset directly or embedded in your system. WEKA has a lot of tools for data pre-processing, classification, regression, clustering, association rules, and visualization. And also, well fit developer who develop a new machine learning schemes. Weka is open source software issued under the GNU General Public License. [9]

1. Data pre-processing

We will apply 2 types of data pre-processing for this study:

1. Normalization.

Normalization is transforming the numeric attributes in the dataset. It will standardize the attributes to 0 to 1 / -1 to 1.

1. Discretization (set to 10 bin).

Transforming numeric attributes to nominal and set equal-width binning to 10 bins. For this, it will apply to all numeric attributes.

Figure 2 below shows the visualization in histogram for all 21 attributes in the dataset before pre-processing.

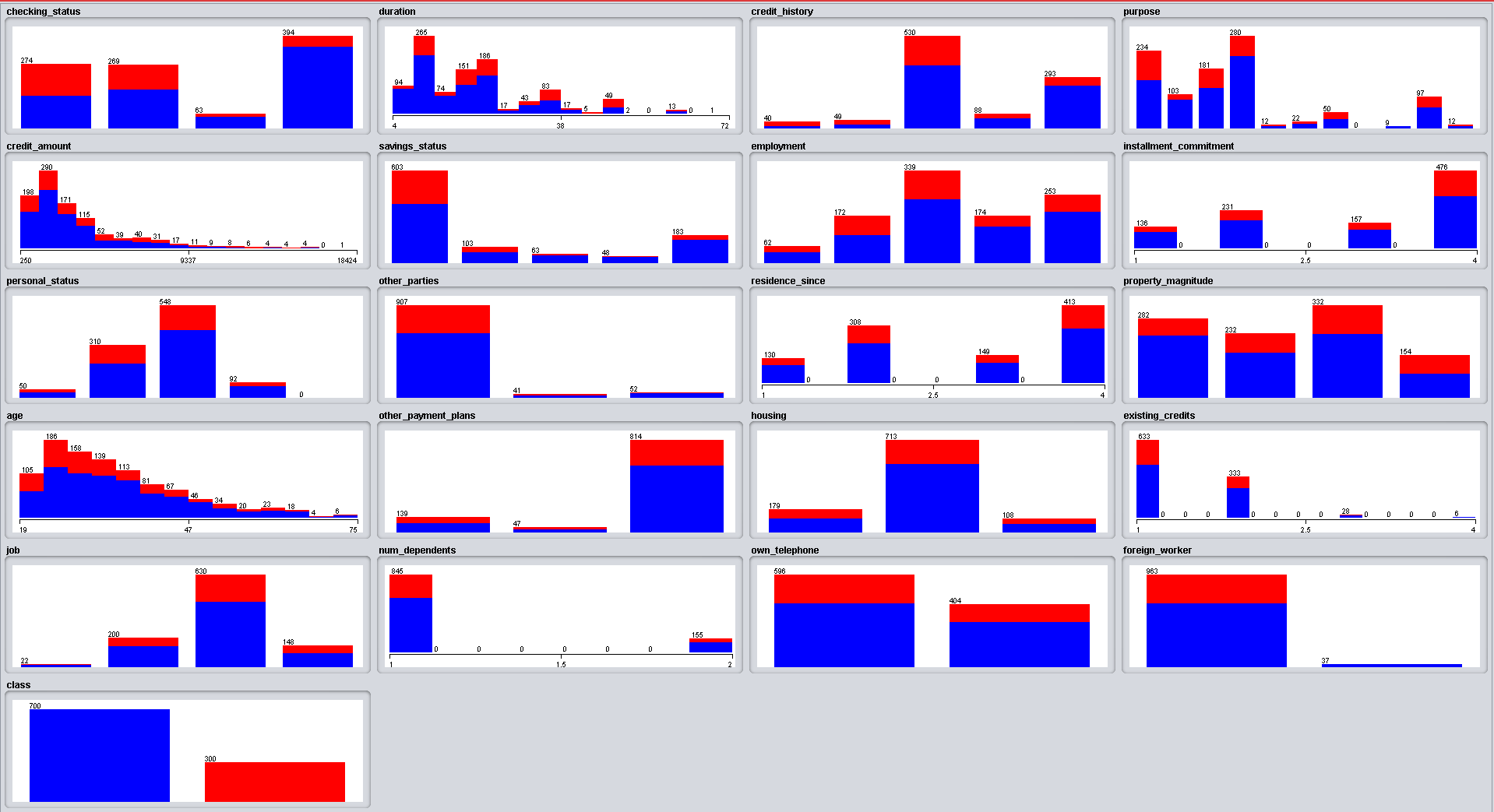


Figure 2. Visualization of the dataset before pre-processing.

Figure 3 below shows the visualization in histogram for all 21 attributes in the dataset after the pre-processing. We can see that there is changes happen on some of the attributes that have numeric type after the pre-processing.

This pre-processing will help in reducing the time taken to build the model and also the time taken to test the model. It also will increase the accuracy of the result. The correctly classified instance percentage also will increase and the degree of mistakes done by the algorithms also will decrease.

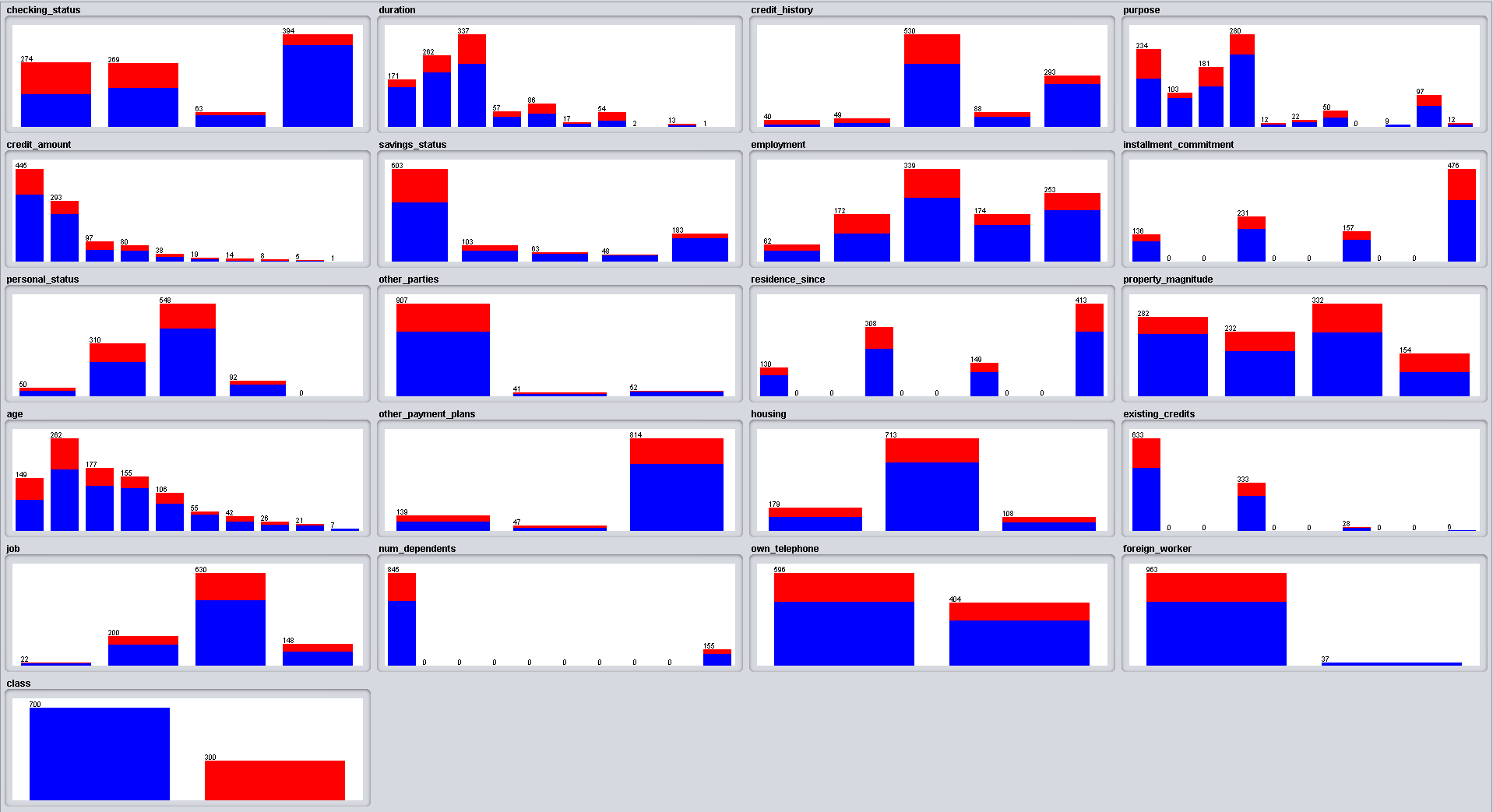


Figure 3. Visualization of the dataset after pre-processing.

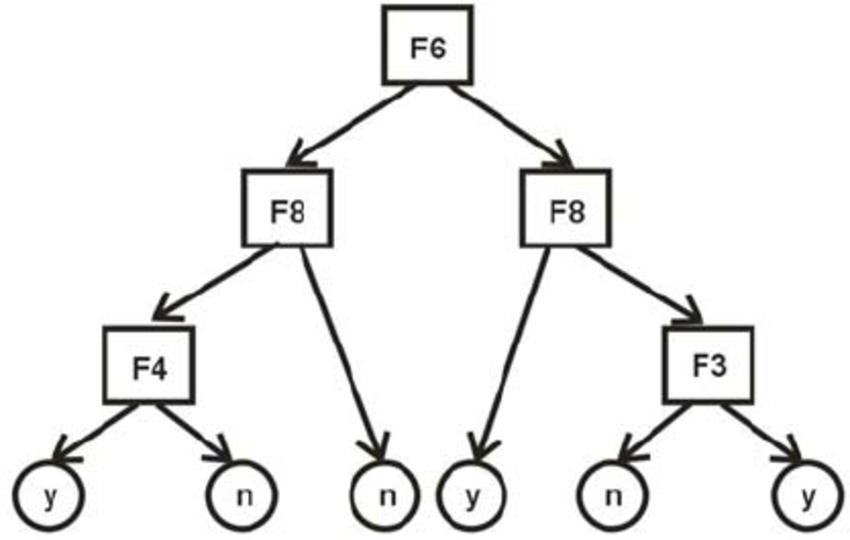
1. Weka J-48  
   

Figure 4. Weka J-48 Classifier

WEKA did implementations for a lot of classification and prediction algorithms. All these basic ideas behind using all of these are similar.

This calculation is helping in the insightful machine-learning model. The dependent variable for this figuring is another case in light of the diverse quality estimation of open data. The internal hubs of a choice tree mean the contrasting qualities, the branches between the centre points reveal to us the conceivable respect that these properties can have in the watched tests, while the terminal centre points reveal to us the last regard of the dependent variable [4].

The quality that will be foreseen is known as the dependent variable, since its regard depends on, or is picked by, the estimations of the different traits. Interchange characteristics, which help in predicting the estimation of the dependent variable, in the data set they have the name independent variables.

the classifier takes after the going with direct calculation. Remembering the ultimate objective to order something else, it first needs to settle on the tree in perspective of the trademark estimations of the open getting ready data. Thusly, at whatever point it encounters a plan of things it recognizes the trademark that isolates the distinctive illustrations by and large clearly. This part can uncover to us most about the information events with the objective that we can arrange them the best is said to have the most raised information get. By and by, among the possible estimations of this component, if there is any an impetus for which there is no ambiguity, that is, for which information cases falling inside its order have a comparative motivating force for the goal variable, by then we end that branch and apportion to it the target regard that we have gained. [5]

For substitute cases, we by then look for another characteristic that gives us the most dumbfounding information get. In this way, we continue along these lines until the point that we either get an unmistakable choice of what mix of characteristics gives us a particular target regard, or we miss the mark on properties. On the off chance that we miss the mark on properties, or in case we can't get an unambiguous result from the open information, we permit this branch a target regard that the vast majority of the things under this branch have.

Since we have the tree, we take the demand of attribute assurance as we have gotten for the tree. By checking all the individual qualities and they regard with those found in the decision tree appear, we can allow or suspect the target estimation of this new event. The above depiction will be all the clearer and less requesting to fathom with the help of a delineation.

1. Random Forest Tree:

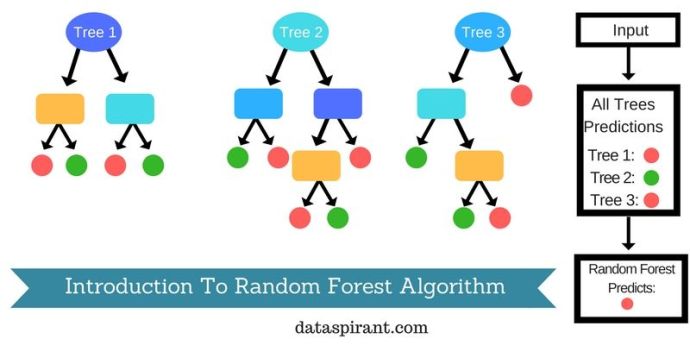


Figure 5. Random Forest Tree Classifier

Random forest algorithm is a supervised classification algorithm. The advantages of this algorithm [6]

1. it will take care of all the missing values and gives options to replace them
2. it will work on multi-level of trees it can work with unlimited count of them
3. it can create multi classifier and groups

Random Forest pseudocode:

first step we select randomly number of features from the total features the selected features we called ‘k’ and the total ‘m’.

Then after that using best split point we calculate the nodes and we will call it ‘d’

and we split the rest of the nodes after that

we keep repeating the previous steps until we reach to the desired number of nodes ‘i’

with repeating all previous steps with the multi trees we will get our results

The begin of irregular backwoods calculation starts with indiscriminately picking "k" incorporates out of total "m" features. In the photo, you can watch that we are discretionarily taking features and recognitions.

In the accompanying stage, we are using the irregular picked "k" features to find the root hub by using the best split approach.

The accompanying stage, we will find out the following hub using a comparable best split approach. Will the underlying 3 stages until the point when the moment that we outline the tree with a root center point and having the goal as the leaf hub.

At last, we re-try 1 to 4 stages to make "n" subjectively made trees. This erratically made tree shapes the random forest [7].

1. **Random Trees algorithm:**

the calculation what we using with this algorithm is the general name for various sorts of trees calculation to get the outcomes and help with basic leadership. Irregular Trees are great when you simply require a "discovery" which can anticipate your needy variable as precisely as could reasonably be expected

Necessities. you require no less than one Data fields and one Target field. Target and data fields can be constant or obvious. Fields that are set to either Both or None are neglected. Fields that are used as a piece of the model must have their sorts totally instantiated, and any ordinal handle that is used as a piece of the model must have the numeric limit (not string). If imperative, the Rename centre can be used to transform them.

Qualities. They are overwhelming when you are overseeing immense educational accumulations and amounts of fields. In light of the use of sacking and field investigating, they are essentially less slanted to overfitting and in this way, the results that are found in testing will most likely be repeated when you use new data [8].

1. Result

For the classification, three algorithms were used. The results of these algorithms are then compared to find out the best fit for the dataset. The algorithms used are:

1. Weka J-48
2. Random Forest Tree
3. Random Tree

The algorithms mentioned above are selected for implementation to find out the most accurate algorithm for loan credit risk assessment that which is to compare between good and bad debt. The dataset consist of instances. These instances were split into 66% of training data and 34% of testing data.

**Weka J-48**

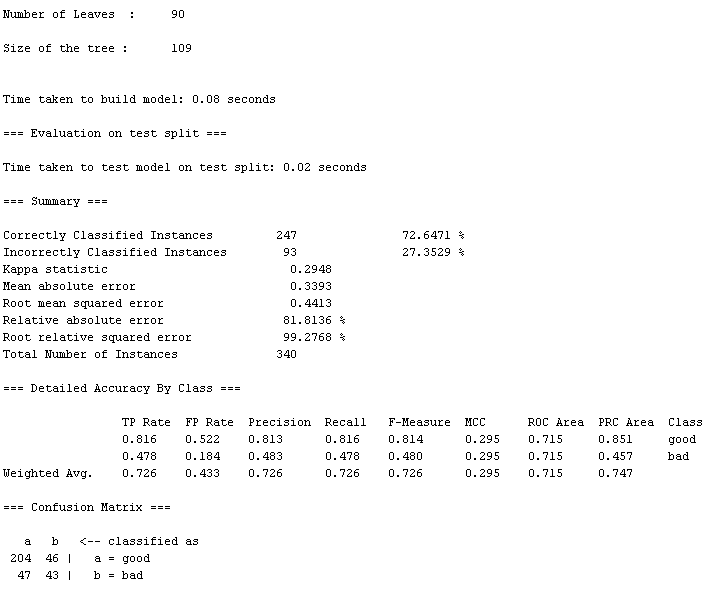


Figure 6. Weka J-48 result

**Random Forest Tree**

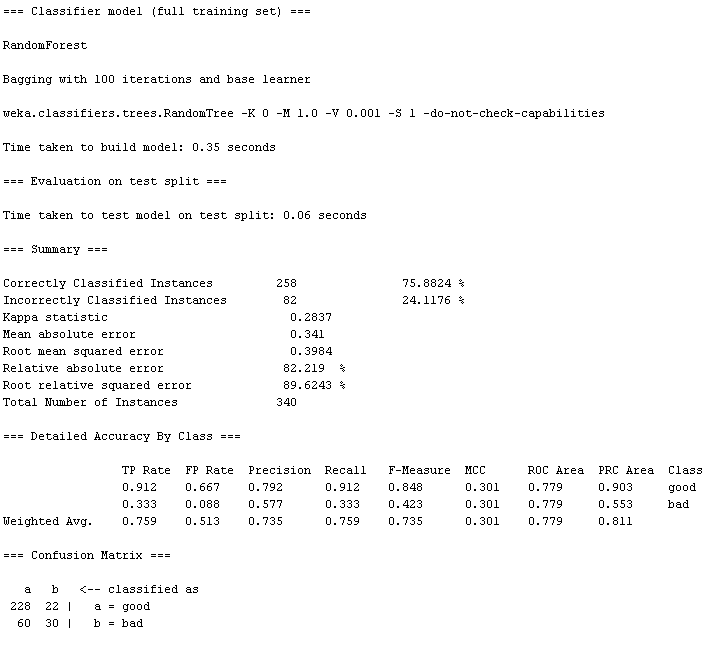


Figure 7. Random Forest Tree result

**Random Tree**

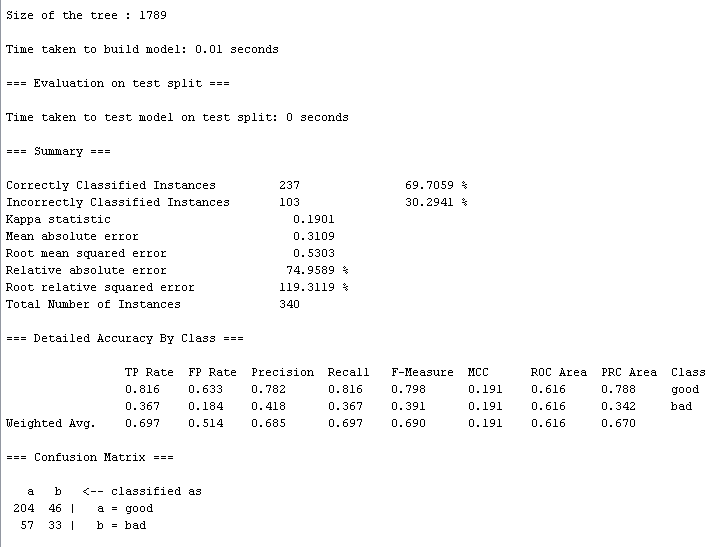


Figure 8. Random Tree result

**Comparison between three algorithms:**

We will compare the three algorithms to find out the most accurate algorithm for loan credit risk assessment that which is to classify between good and bad debt. We will compare using the confusion matrix, accuracy measures, error measures, and accuracy rate.

**Confusion Matrix**

Confusion matrix are obtained to find out the accuracy of the data and how many instances belong to which class. It also shows how many instances are correctly identified and which class they belong to. In this study, we have two classes hence we have confusion matrix of 2 by 2. The confusion matrix for each algorithm has been given below.

**Table 3: Comparison of confusion matrix between three algorithms.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Confusion Matrix** | | |
| Weka J-48 |  | a (Good Debt) | b (Bad Debt) |
| a (Good Debt) | 204 | 46 |
| b (Bad Debt) | 47 | 43 |
| Random Forest Tree |  | a (Good Debt) | b (Bad Debt) |
| a (Good Debt) | 228 | 22 |
| b (Bad Debt) | 60 | 30 |
| Random Tree |  | a (Good Debt) | b (Bad Debt) |
| a (Good Debt) | 204 | 46 |
| b (Bad Debt) | 57 | 33 |

Based on the Table 3 above, we can conclude that Random Forest tree algorithm shows the highest accuracy in identifying in which class the instance belongs. With 228 instances belong to the good debt class and 30 instances belong to the bad class, it left only 82 instances unidentified which is the lowest from all three algorithms.

**Accuracy Measures**

Accuracy measures table will be used to determine how accurate the model.

**Table 4: Comparison of accuracy measures between three algorithms.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **TP** | **FP** | **Precision** | **Recall** | **Class** |
| Weka J-48 | 0.816 | 0.522 | 0.813 | 0.816 | Good debt |
| 0.478 | 0.184 | 0.483 | 0.478 | Bad debt |
| Random Forest Tree | 0.912 | 0.667 | 0.792 | 0.912 | Good debt |
| 0.333 | 0.088 | 0.577 | 0.333 | Bad debt |
| Random Tree | 0.816 | 0.633 | 0.782 | 0.816 | Good debt |
| 0.367 | 0.184 | 0.418 | 0.367 | Bad debt |

From the Table 4 above, we can conclude that Random forest tree has the highest accuracy among the three algorithms.

**Error Measures**

Error measure is to shows the degree of mistakes done by each algorithm. The smaller the error, the more accurate it will be.

**Table 5: Comparison of error measures between three algorithms.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Evaluation Criteria** | **Classifiers** | | |
| **Weka J-48** | **Random Forest Tree** | **Random Tree** |
| Kappa Statistics | 0.2948 | 0.2837 | 0.1901 |
| Mean absolute error | 0.3393 | 0.3410 | 0.3109 |
| Root mean squared error | 0.4413 | 0.3984 | 0.5303 |
| Relative absolute error | 81.8136% | 82.2190% | 74.9589% |
| Root relative squared error | 99.2768% | 89.6243% | 119.3119% |

From the Table 5 above, we can conclude that Random tree algorithm has the highest accuracy follow by Random Forest tree algorithm, while Weka J-48 algorithm has the lowest accuracy among three algorithms.

**Accuracy Rate**

This table shows the overall result of which algorithm is the best and achieved the highest accuracy to classify good and bad debt for the credit loan assessment from the dataset.

**Table 6: Comparison of accuracy rate between three algorithms.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Evaluation Criteria** | **Classifiers** | | |
| **Weka J-48** | **Random Forest Tree** | **Random Tree** |
| Time to build model (in seconds) | 0.08 | 0.35 | 0.01 |
| Correctly classified instance | 247 | 258 | 237 |
| Incorrectly classified instance | 93 | 82 | 103 |
| Accuracy (%) | 72.6471% | 75.8824% | 69.7059% |

Based on Table 6 above, we can conclude that among the three algorithms, Random Forest tree algorithm has the highest accuracy at 75.8824%. So, from all the comparison that we have made, we can say that Random Forest tree is the best algorithm to be used in classifying good and bad debt for the credit loan assessment.

1. Conclusion

In the banking or financial sectors, it is very crucial for the organization to take strict measures on the information that they have and what decisions they are going to produce and minimizing the risk factors. Loan credit risk assessment information is one of the key parameter or risk factors used by banking and financial industries before acquiring new customers. As far as we understood, banking business is about making as much money as possible while without neglecting the customers.

In this paper, with the data set that we have collected and carefully analysed using three different classifier methods namely Weka J-48, Random Forest, and Random Tree and using Weka Tool as machine learning algorithms for data mining, we can have come to conclusion that Random Forest tree is the best algorithm to be used in comparing good and bad debt for the credit loan assessment. This has been proven with the results based on the assessment using confusion matrix, accuracy measures, error measures, and accuracy rate test tabulated in the previous section above.

1. References

[1] ‘UCI Machine Learning Repository’, http://archive.ics.uci.edu/ml, accessed 3 January 2018

[2] ‘Introduction to Machine Learning in the Azure cloud’, <https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-machine-learning>, accessed 5 January 2018

[3] ‘3 Different Types of Machine Learning’, <https://www.kdnuggets.com/2017/11/3-different-types-machine-learning.html>, accessed 5 January 2018

[4] ‘Chapter 5- Classification Method’, http://www.d.umn.edu/~padhy005/Chapter5.html, accessed 7 January 2018

[5] ‘Class J48’http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html, accessed 8 January 2018

[6] ‘How The Random Forest Algorithm Works In Machine Learning’, https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/, accessed 10 January 2018

[7] ‘Wikipedia- Random Forest’, https://en.wikipedia.org/wiki/Random\_forest, accessed 11 January 2018

[8] ‘Random Trees algorithm in SPSS Modeler’, https://developer.ibm.com/predictiveanalytics/2015/10/11/random-trees-algorithm-in-spss-modeler-17-1/, accessed 11 January 2018

[9] ‘Data mining softwre in java’ https://www.cs.waikato.ac.nz/ml/weka/index.html