A Comparative Study of Statistical and Data Mining Algorithms for Prediction Performance

Amjad Harb and Rashid Jayousi

Al-Quds University

Jerusalem, Palestine

amjad@pcbs.gov.ps and rjayousi@science.alquds.edu

***Abstract*-The aim of this study is to perform a comparison experiment between statistical and data mining modelling techniques. These techniques are statistical Logistic Regression, data mining Decision Tree and data mining Neural Network. The comparison will evaluate the performance of these prediction techniques in terms of measuring the overall prediction accuracy percentage agreement for each technique. The ratio of the binary values of the dependent variable in the training dataset and the population is used on the three techniques to find the effect of this ratio on the prediction performance. For a given data set, the results shows that the performance of the three techniques is comparable in general with small outperformance for the Neural Network. An affecting factor that makes the prediction accuracy varied is the dependent variable values distribution (distribution of "0"s and "1"s). It is seen that, for all of the three techniques, the overall prediction accuracy percentage agreement is high when the ratio of "0"s and "1"s is 3:1, whereas for the ratios 2:1 and 1:1 the performance is lower.**

***Keywords*- Data Mining, Classification, Prediction Model, Statistical Logistic Regression, Neural Network, Decision Tree.**

1. Introduction

An important and challenging area of research is information management. Historical data was analyzed using several ways for hidden knowledge extraction that can help in decision making. This is called Knowledge Discovery or Data Mining. The popular goal from data mining is prediction and the popular data mining technique used for prediction is classification. Classification can be accomplished statistically or by data mining methods. [1]

Prediction techniques performance comparison issues is an interesting topic for many researchers. A comparative study by Lahiri R. [1] compared the performance of three statistical and data mining techniques on Motor Vehicle Traffic Crash dataset, resulted that the data information content and dependent attribute distribution is the most affecting factor in prediction performance. Delen D. et al. [2] targeted data mining methods comparison as a second objective in the study, while the main objective was to build the most accurate prediction model in a critical field, breast cancer survivability. In the same area, Artificial Intelligence in Medicine, Bellaachia A. et al. [3] continued the work of [2] and improved the research tools especially the dataset. An important application area that exploited data mining techniques heavily was the network security. Panda M. et al. [4] also performed a comparative study to identify the best data mining technique in predicting network attacks and intrusion detection. Also the data contents and characteristics revealed as an affecting factor on the data mining and prediction algorithms performance.

In this research we will continue on the work of Lahiri R. [1] to perform a comparison on the same data mining techniques: Logistic Regression, Neural Network and Decision Tree, but with more accurate data content and quality. Also we will work on Lahiri's future work recommendation by determining more precise predictors that significantly define and affect the output. In another words, we intended to find the effect of the most correlated variables, as predictors, to the dependent variable on the prediction accuracy rates for the aforementioned prediction techniques. The overall prediction percentage agreement will be the main performance metric which will be measured under different dependent variable values distributions (distribution of "0"s and "1"s) in the dataset. This is to identify the effect of the dependent variable values distribution on the overall prediction accuracy for the three prediction techniques. The experiment will exploits a historical dataset about the Palestinian Labor Force. The dependent attribute will be the individual's "Labor Force Status", that have values: 0 as Employed and 1 as Unemployed. The source of such data is the Palestinian Central Bureau of Statistics (PCBS).

This paper is organized as follows: The literature and related work will be discussed in Section 2. The research methodology to perform the experiment will be presented in Section 3. Experimental results are presented and discussed in Section 4. Finally, Conclusion is given in the last Section 5.

2. Related Work

Many studies have been done across countries on data mining. Applications of data mining were used in a large number of fields, especially for business and medical purposes.

As data mining is a new technology field, it is important and very helpful in predicting and detecting underlying patterns from large volumes of data, many researches were published, comparing results of data mining algorithms in several areas. A research by Rochana Lahiri (2006) performed a performance comparison of several data mining and statistical techniques for classification model. She used a database from Louisiana Motor Vehicle Traffic Crash. The performance was measured in terms of the classification agreement %. The effect of Decision Tree, Neural Network, and Logistic Regression models for different sample sizes and sampling methods on three sets of data had been investigated. The study concluded that a very large training dataset is not required to train a Decision Tree or a Neural Network model or even for Logistic Regression models to obtain high classification accuracy and the overall performance reached a steady value at the sample size of 1000, irrespective of the total population size. The information content of a training dataset is an important factor influencing classification accuracy and is not governed by the size of the dataset. Another important result was that the sampling method has not affected the classification accuracy of the models. She also stated that the overall classification accuracy of the all three methods were very much comparable and no one method over performed any other. She tried to find the effect of the "0"s and "1"s distribution of dependent variable values in the dataset but because the data was very skewed, she failed to do this. As a future work, the study recommends to apply the same study on a dataset were the relationships between the dependent variable and the independent variables are more rigid. i.e.: to select predictors that strongly describe the dependent attribute, and to study the effect of "0"s and "1"s dependent variable values distribution. [1]

The data mining methods comparison were targeted as a second objective in some studies that mainly aimed to develop a prediction model in a critical fields, like medicine, by investigating several data mining methods, intending to get the model that have the highest prediction accuracy. This type of studies has been addressed by Delen D. et al. (2005) in the context of predicting breast cancer survivability. Multiple prediction models, using Artificial Neural Networks, Decision Trees, and Logistic Regression, for breast cancer survivability using a large dataset had been developed. The comparison among the three models had been conducted depending on measuring three prediction performance metrics: classification accuracy, sensitivity and specificity. The k-Fold cross-validation test was used to minimize the bias associated with the random sampling of the training and missing data. The results of the study showed that the Decision Tree (C5) preformed the best of the three models evaluated. Sensitivity analysis, which provides information about the relative importance of the input variables in predicting the output field, was applied on Neural Network models and provided them with the prioritized importance of the prediction factors used in the study. [2]

Another related study in medicine by Bellaachia A. et al. (2006) also in the context of predicting breast cancer survivability. The researchers took the study of Delen D. et al. [2] as the starting point with the same dataset source but with a newer version and different set of data mining techniques. For modeling and comparison, three data mining techniques had been investigated: the Naïve Bayes, the back-propagated Neural Network, and the C4.5 Decision Tree algorithms. The main goal was to have a prediction model with high prediction accuracy, besides high precision and recall metrics for patients' data retrieval. They used other performance metrics: specificity and sensitivity to compare the prediction models. The results presented that C4.5 algorithm has a much better performance than the other two techniques. The obtained results differed from the study of Delen D. et al. [2] due to the facts that they used a newer database (2000 vs. 2002), a different pre-classification (109,659 and 93,273 vs. 35,148 and 116,738) and different toolkits (industrial grade tools vs. Weka). [3]

In network security, data mining techniques used heavily in predicting network intrusion detection systems to protect computing resources against unauthorized access. Several studies were performed in this area and some of them addressed the prediction performance comparison of different data mining techniques like the study by Panda M. et al. (2008). A dataset of 10% KDDCup'99 intrusion detection has been generated and used in the experiment. Three popular data mining algorithms had been used in the experiment: Decision Trees ID3, J48 and Naïve Bayes. The prediction performance metrics used in the study were the time taken to build the model and the prediction error rate. For the evaluation of prediction error rate, the 10-fold cross validation test was used. As a result of the experiment, the Decision Trees had proven their efficiency in both generalization and detection of new attacks more than the Naïve Bayes. But this maybe dependence on the contents and characteristics of the data which allows single algorithm to outperform others. [4]

Amooee G. et al. (2011) used data mining techniques to identify defective parts manufactured in an industrial factory and to maintain high quality products. A data of 1000 records was collected from the factory and 10% (100 records) of the data was about a defective parts. Prediction accuracy and processing time of the prediction techniques were the comparison performance metrics. The results showed that SVM and Logistic regression prediction algorithms has the best processing time with high overall prediction accuracy. The decision tree with its tree different branching algorithms (CRT, CHAID, and QUEST) achieved the highest prediction accuracy rates but needed more time. Neural network achieved the least prediction accuracy rate with medium processing time. [5]

Data mining concept was the most appropriate to the study of student retention from sophomore to junior year than the classical statistical methods. This was one main objective of the study addressed by Ho Yu C. et al. (2010) in addition to another objective that identifying the most affecting predictors in a dataset. The statistical and data mining methods used were classification tree, multivariate adaptive regression splines (MARS), and neural network. The results showed that transferred hours, residency, and ethnicity are crucial factors to retention, which differs from previous studies that found high school GPA to be the most crucial contributor to retention. In Ho Yu C. et al. research, the neural network outperformed the other two techniques. [6]

The prediction techniques RIPPER, decision tree, neural networks and support vector machine were used to predict cardiovascular disease patients. The performance comparison metrics were the Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate. Kumari M. et al. study showed that support vector machine model outperforms the other models for predicting cardiovascular disease. [7]

The neural network was found to achieve better performance compared to the performance rates of Naive Bayes, K-NN, and decision tree prediction techniques in a study performed by Shailesh K R et. al. (2011) to predict the inpatient hospital length of stay in a super specialty hospital. [8]

The same result was seen that the neural network outperformed both the decision tree and linear regression models when the performance for the students' academic performance in the undergraduate degree program was measured by predicting the final cumulative grade point average (CGPA) of the students upon graduation. The correlation coefficient analysis was used to identify the relationship of the independent variables with the predictors. Ibrahim Z. et al. (2007) [9]

Social network data, using data mining techniques and the prediction error rates were the comparison metric, was studied by Nancy P. et al. (2011). The tree based algorithms such as RndTree, ID3, C-RT, CS-CRT, C4.5, CS-MC4 and the k-nearest neighbor (k-NN) algorithms were used in the study. The RndTree algorithm achieved least error rate and outperformes the other algorithms. [10]

C. Deepa et al. (2011) compared the prediction accuracy and error rates for the compressive strength of high performance concrete using MLP neural network, Rnd tree models and CRT regression. The results showed that neural network and Rnd tree achieved the higher prediction accuracy rates and Rnd tree outperforms neural network regarding prediction error rates. [11]

The Rand tree algorithm also outperforms the other algorithms, C4.5, C-RT, CS-MC4, decision list, ID3 and naïve bayes, in a study of vehicle collision patterns in road accidents by S.Shanthi et al. (2011). Selection algorithms were used including CFS, FCBF, Feature Ranking, MIFS and MODTree, to improve the prediction accuracy. Feature Ranking algorithm was found the best in improving the prediction accuracy for all algorithms. [12]

In this study, we have used a Labor Force Data (LFS 2009) in the Palestinian's territories as a training dataset with 38,037 records. For testing, we used the same dataset of Labor Force Data (LFS 2009) and Labor Force Data (LFS 2010). The two datasets was processed and cleaned against missing values and inconsistency.

3. Methodology

To achieve the objectives of this research, we have started data preparation (LFS 2009 and LFS 2010) in a suitable form for the experiment. The data contains the following variables: person's labor status (dependent variable), Sex, Age at last Birthday, Does he currently attending school?, Years of schooling, Educational Attainment (higher qualification), Refugee Status, District, Locality Type, Region, and Marital Status. All these variables are numeric data type and nominal measure, see Table 1. We should mention that for the training of the prediction algorithms we used LFS 2009 dataset and for testing we used both LFS 2009 and LFS 2010.

TABLE 1: SPECIFICATION OF THE LABOR FORCE DATASET VARIABLES

Variable	Values	Measure
Sex	1 or 2	Nominal
Age at last Birthday	0-100	Nominal
Does he currently attending school	1-4	Nominal
Years of schooling	0-40	Nominal
Educational Attainment (higher qualification)	1-10	Nominal
Refugee Status	1-3	Nominal
District	16 Category	Nominal
Locality Type	1-3	Nominal
Labor Force Status	0: Employed 1: Unemployed	Nominal
Region	1 or 2	Nominal
Marital Status	1-3	Nominal

We selected the "*person's labor status"* as the dependent variable that has only two expected values, 1 as Unemployed and 0 as Employed. Because we need to find the effect of dependent variable values distributions (distribution of "0"s and "1"s) on the overall prediction percentage agreement, we prepared different copies of the dataset and changed the dependent variable values distributions for three values ratios, see Table 2.

TABLE 2: DEPENDENT VARIABLE VALUES DISTRIBUTION

Ratio	0: Employed	1: Unemployed	Total
1:1	9,292	9,292	**1,8584**
2:1	20,122	9,292	**29,414**
3:1	28,745	9,292	**38,037**

The original dataset is the set with ratio 3:1, and from this dataset we extract the other two datasets by reducing the number of records of the "Employed" persons in a way to have the required ratio. This was done by deriving a stratified sample for each ratio type using "District" as stratifying variable.

Selecting the independent variables (inputs) that have tight relationships to the dependent variable (output) to get more accurate results when applying the aforementioned prediction techniques, is one of the main objectives of this study. To verify this, we applied *"Spearman's Bi-variate correlation"* analysis to identify the exact coefficient percentage and significance of this effect between the variables of the dataset on each other. The correlation analysis was performed on the three ratio datasets and only the correlation results of the dependent variable *"Labor Force Status"* with other variables was selected, see Table 3. As shown in Table 3, we selected the variables that have correlation coefficient of 10% and more, and significance value of 0.05 and less. This results the variables, marked in bold, that are the most related to the dependent variable, they are (ordered by importance): Region, Age, District, Marital Status, and Refugee Status.

After this step, we run the prediction techniques: Binary Logistic Regression, Multilayer Perceptron Neural Network (MLP), and CRT Decision Tree on the three ratio datasets at two stages. In the first stage we assigned the all variables as independent variables, apply the prediction algorithm and recording the results. The same procedure was followed in stage two but we assigned the independent variables depending on the correlation analysis results as shown in Table 2.

TABLE 3: LABOR FORCE STATUS CORRELATIONS

N		1,8584	29,414	38,037
Unemployment : Employment Ratio		**1:1**	**2:1**	**3:1**
Sex	Coefficient	.001	.004	.001
	Significance	.866	.457	.790
Age at last Birthday	**Coefficient**	**-.242**	**-.219**	**-.203**
	Significance	**.000**	**.000**	**.000**
Does he currently attending school	Coefficient	.017	.014	.011
	Significance	.023	.015	.026
Years of schooling	Coefficient	.017	.019	.016
	Significance	.023	.001	.002
Educational Attainment (higher qualification)	Coefficient	-.015	-.019	-.018
	Significance	.035	.001	.001
Refugee Status	**Coefficient**	**-.117**	**-.107**	**-.099**
	Significance	**.000**	**.000**	**.000**
District	**Coefficient**	**.238**	**.221**	**.205**
	Significance	**.000**	**.000**	**.000**
Locality Type	Coefficient	-.008	-.003	-.003
	Significance	.267	.622	.539

TABLE 3 CONT.

Region	**Coefficient**	**.252**	**.243**	**.228**
	Significance	**.000**	**.000**	**.000**
Marital Status	**Coefficient**	**-.197**	**-.185**	**-.173**
	Significance	**.000**	**.000**	**.000**

The primary results obtained of the overall prediction accuracy percentage for the three prediction algorithms showed that the differences between assigning the independent variables by selecting all variables in stage one and selecting only the correlated variables in stage two are almost close. Even it is slightly larger by selecting the all variables than selecting the correlated variables. After the training of the models, we tested them using the data of LFS 2009 and LFS 2010.

The *PASW Statistics (SPSS Release 18.0.0)* from IBM was used in all operations and to calculate all the aforementioned statistical and data mining techniques and methods.

4. Experimental Results

The results of the analyses performed on the three different dataset's ratios using the three different prediction techniques have been trained and tested using the LFS 2009 and LFS 2010 datasets. Our concentration will be on the overall prediction accuracy as a metric of performance of the prediction techniques. As we have said in the methodology, we experimented with the three prediction techniques for each ratio dataset two times. One time using the all variables as independent variables and the other time assigning only the variables that are the most correlated with the dependent variable, *"Labor Force Status"*, as seen in Table 2. The overall prediction accuracy for the two iterations in each ratio dataset are too comparable, even with using the all variables as independent variables resulted a bit larger accuracy than using the correlated variables and for all ratio datasets, see Table 4. This results suggest that the three prediction techniques, in some way or another, rely on the independent variables that are most correlated to the dependent variable and with more higher accuracy if some of the other less correlated independent variables, that also have acceptable significance value, added to the analysis. We can imply from this that we can save time by selecting all candidate variables, by common sense, as independent variables without worry to calculate the correlations, unless we can find another way to identify the most correlated variables. Depending on this, we continued the analysis without the results that were based on the correlated variables.

TABLE 4: OVERALL PREDICTION ACCURACY PERCENTAGE OF THE PREDICTION TECHNIQUES FOR BOTH *ALL* AND *CORRELATED* VARIABLES

Method*	**Type**	**Ratio**	**Training**	**Testing 2009**	**Testing 2010**
DT	Corr.	1:1	65.7	67.1	65.3
DT	All	1:1	68.5	67.2	68.4
DT	Corr.	2:1	72.5	77	78.3
DT	All	2:1	73.4	77.1	76.8
DT	Corr.	3:1	77.4	77.4	79
DT	All	3:1	78.1	78.1	78.8
LR	Corr.	1:1	64.2	62.4	62.7
LR	All	1:1	64.6	64.1	66
LR	Corr.	2:1	72.2	76.9	78.6
LR	All	2:1	72.3	77	78.2
LR	Corr.	3:1	77.5	77.5	79
LR	All	3:1	77.6	77.6	78.7
NN	Corr.	1:1	67.5	67.3	65.9
NN	All	1:1	70.5	70.5	70.4
NN	Corr.	2:1	72.6	77	78.3
NN	All	2:1	74.9	78.3	77.6
NN	Corr.	3:1	77.6	77.6	79.1
NN	All	3:1	78.8	78.8	78.4

*: DT: Decision Tree, LR: Logistic Regression, NN: Neural Network

For investigating the effect of dependent variable values distribution we replicated the dataset into three categories each with different "0" to "1" ratio as in Table 2 in the previous section.

Fig.1 plots three graphs one for each dataset type as training, testing 2009 and testing 2010 respectively. Each graph plots the overall prediction accuracy percent over the three datasets ratios and using the prediction methods. It is seen that, in the three graphs, the prediction performance for the three prediction techniques were comparable and no one prediction technique outperform the other two, except a very small outperformance for the neural network in some conditions. For the training graph, it is seen that when the dependent variable values distribution ratio was 1:1, the overall prediction accuracy rate of neural network, decision tree and logistic regression were around 70%, 68% and 66% respectively. This means that neural network predicted the person's work status accurately more than the other two with small difference among them. When the dependent variable values distribution ratio was 2:1, the same reading was achieved but with more smaller differences among the three prediction techniques' overall prediction performance rates. The behavior was the same when the dependent variable values distribution ratio became 3:1 but this time the overall prediction accuracy rates for the three methods were almost the same with a non significant difference among their results reached a maximum value of 1.2%. It is also seen that by increasing the dependent variable values ratio, the overall prediction accuracy rate for each one of the three techniques plot an increasing curve, starting from 1:1 ratio that scored the lowest overall prediction accuracy rate value with significant difference between this value and the prediction value when the ratio was 2:1. The same behavior was seen between the ratios 2:1 and 3:1.

For testing data the behavior was the same as for the training data. Each prediction technique plots an increasing curve. As the dependent variable values distribution ratio increased, the overall prediction accuracy rate also increased. The difference between the training data curves and the testing data curves was that in the testing curves a plateau was reached when the dependent variable values distribution ratio was 2:1 and increasing the ratio to 3:1 didn't significantly improve the overall prediction accuracy rate more. While the curves of the training data were almost linear and no plateau was reached. This means that in the testing graphs the curves met when the dependent variable values distribution ratio was 2:1 and continued the same level reaching to ratio 3:1, while in the training curves they met when the dependent variable values distribution was 3:1 and not before.

Another exciting results were the results of prediction accuracy rates for the individual values of the dependent variable. Table 5 shows the prediction accuracy rates of the dependent variable values and the overall prediction accuracy for the three prediction techniques over the three dependent variable values distribution ratios and for the testing datasets 2009 and 2010.

It is seen that for all of the prediction techniques and when the dependent variable values distribution were 2:1 and 3:1, the prediction accuracy rates for the "0" and "1" were highly not comparable. The prediction methods predicted the dependent variable value that have the largest frequency in the dataset with high accuracy rate while they fail to predict the other value at approximately the same, or nearby, accuracy rate. For example the neural network in the year 2009 data and for the ratio 3:1, predicted the value of "0" and "1" for 94.9% and 28.9% respectively.

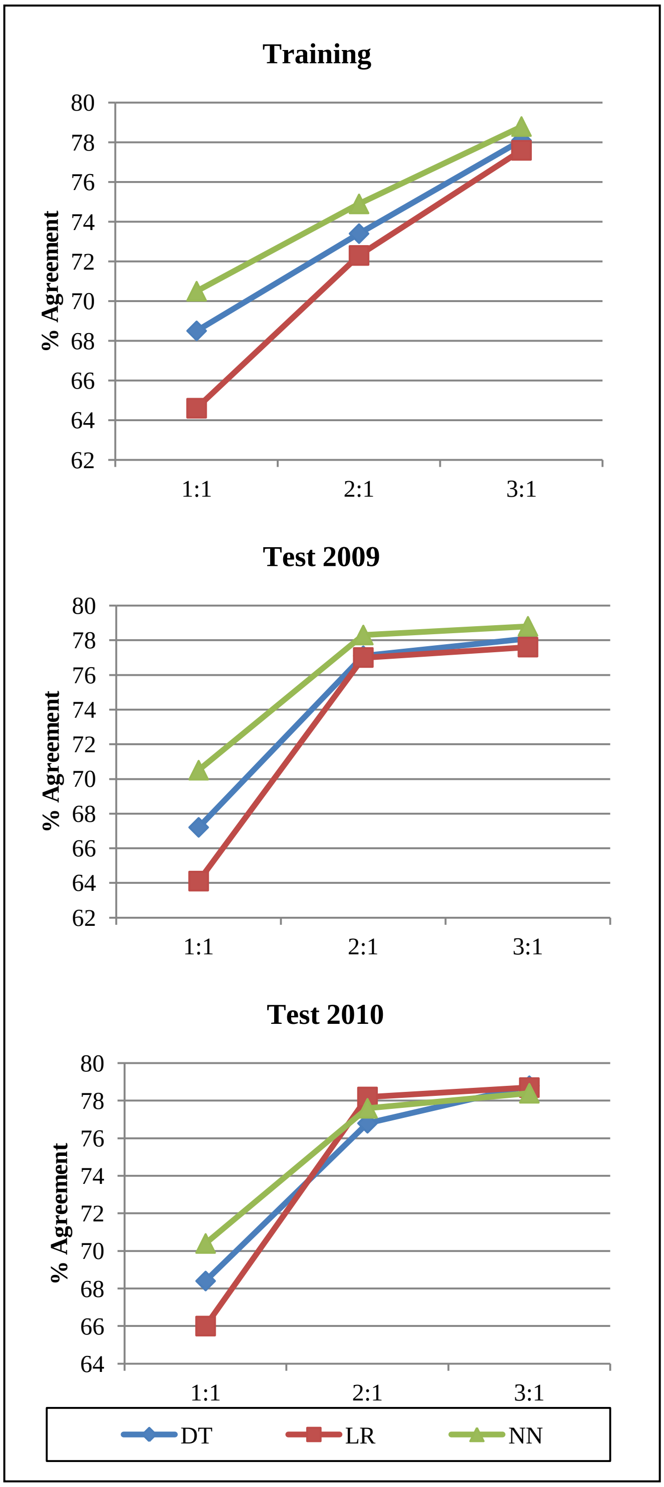

Fig. 1: The overall prediction accuracy rates of the prediction techniques over the three ratio datasets

We can see how much the difference is, and this holds for all of the prediction methods within all of the dependent variable values ratios and in the two years, 2009 and 2010, even for the training data. On the other hand, It is seen that for all of the prediction techniques and when the dependent variable values distribution was 1:1, the prediction accuracy rates for the “0” and “1” were highly comparable. The “0” and “1” values were fairly predicted by the prediction techniques with acceptable accuracy rate and the difference was very small compared with the difference when the ratio was 2:1 or 3:1. If we take the same example, the neural network in the year 2009 data and for the ratio 1:1, predicted the value of “0” and “1” for 70.8% and 69.6% respectively. This holds for all of the prediction methods within all of the dependent variable values ratios and in the two years, 2009 and 2010, even for the training data. The tradeoff in this situation was that the overall prediction accuracy rate was reduced significantly, but still acceptable, when the ratio became 1:1.

TABLE 5: Prediction Accuracy Percentage Of The Prediction Techniques For The Testing Data 2009 And 2010

Method*	Ratio	Test 2009			Test 2010		
		0:Employed	1:Unemployed	Overall	0:Employed	1:Unemployed	Overall
DT	**1:1**	**66.2**	**70.3**	**67.2**	**71.9**	**56.8**	**68.4**
DT	2:1	89.7	38.1	77.1	90.7	31.5	76.8
DT	3:1	95.8	23.4	78.1	96.9	19.7	78.8
LR	**1:1**	**63.7**	**65.3**	**64.1**	**67**	**62.6**	**66**
LR	2:1	93.1	27.2	77	93.8	27.4	78.2
LR	3:1	96.5	19.1	77.6	97.2	18.6	78.7
NN	**1:1**	**70.8**	**69.6**	**70.5**	**71.2**	**69.6**	**70.4**
NN	2:1	90.1	41.7	78.3	90.5	35.5	77.6
NN	3:1	94.9	28.9	78.8	94.4	26.2	78.4

*: DT: Decision Tree, LR: Logistic Regression, NN: Neural Network

5. Conclusion

In this paper we worked on achieving the objectives that can be illustrated by performing the prediction techniques performance comparisons of: Logistic Regression, Neural Network and Decision Tree using dataset of higher level of accuracy regarding the content. We tried to identify more precise predictors that significantly define and affect the output by using the correlation analysis but the results have demonstrated a very small differences, even it was more accurate without the correlation results. We think this is due to the small number of predictors, that limited our chances to get a highly correlated and robust training dataset.

As a conclusion for this research, the neural network achieved an overall prediction accuracy rate higher than the decision tree and logistic regression when the dependent variable values distribution ratio was 1:1. The prediction performance of the three prediction methods was almost the same and too close to each other when the dependent variable values distribution ratio were 2:1 and 3:1.

Another interesting conclusion is that a tradeoff should be performed, that whether the needed prediction accuracy is a high overall prediction accuracy rate; an adequate and comparable both “0” and “1” dependent value prediction accuracy rate; or a high single “0” or “1” dependent value prediction accuracy rate. If a high overall prediction accuracy is needed then the dependent variable values distribution in the training data should be skewed and the ratio of 0:1 occurrences (or 1:0) for the dependent variable values should be at least 2:1 or larger. This hold also if the requested high prediction accuracy is one of the two dependent variable values, not both, then it’s distribution in the training data should be at least 2 occurrences or more against to 1 occurrence for the other value. An example of this case is the breast cancer diagnosing in women that a high prediction accuracy is needed to check if the patient is infected like the studies of Delen D. [2] and Bellaachia A. [3].

If both dependent variable values are requested to be predicted in comparable prediction accuracy rate, then the training data should not be skewed and the ratio of dependent variable values occurrences should be equal and no more than 1:1. This holds for all of the three prediction techniques and not affected by the total population size of the data (training or testing), because we tested this on another different datasets with total population sizes ranges between 3000 and 4000 instances. This is true if we consider our dataset, between 18,000 and 38,000 instances, as a large dataset but not necessarily a huge dataset.

This conclusion agrees, in general, with the results and conclusion of Lahiri R. [1], but contradict with the failure of neural network to predict one of the dependent variable values. In this study it is seen that the neural network succeeded and even slightly outperformed the logistic regression and decision tree techniques in predicting the values of the dependent variable values, “0” and “1”.

Finally, as a future work to increase the prediction accuracy, we would like to find other techniques that help in finding optimal choice of predictors and do the same study. Also the total population size is another future area of research to test if the huge dataset, with hundreds of thousands or even millions instances, affected the prediction accuracy of the aforementioned prediction techniques in this study.

References

1. Lahiri R., *Comparison of Data Mining and Statistical Techniques for Classification Model*, A Thesis submitted to the graduate faculty of the Louisiana State University in partial fulfilment of the requirements for the degree of Master of Science in The Department of Information Systems & Decision Sciences. (December 2006).
2. Delen D., Walker G., and Kadam A., *Predicting breast cancer survivability: a comparison of three data mining methods*, Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.
3. Bellaachia A. and Guven E., *Predicting Breast Cancer Survivability Using Data Mining Techniques*, Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM 2006).
4. M. Panda and M. R. Patra, *A comparative study of data mining algorithms for network intrusion detection*, proc. of ICETET, India, 2008.pp.504-507. IEEE Xplore.
5. Amooee G., Minaei-Bidgoli B. and Bagheri-Dehnavi M., *A Comparison Between Data Mining Prediction Algorithms for Fault Detection (Case study: Ahanpishegan co.)*, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.
6. Ho Yu C., DiGangi S., Jannasch-Pennell A., and Kaprolet C., *A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year*, Journal of Data Science 8(2010), 307-325.
7. Kumari M. and Godara S., *Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction*, International Journal of Computer Science and Technology (IJCST) Vol. 2, Issue 2, June 2011.
8. Shailesh K R et. al., *Comparison of Different Data Mining Techniques to Predict Hospital Length of Stay*, JPBMS, 2011, 7 (15).
9. Ibrahim Z. and Rusli D*., Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression*, 21st Annual SAS Malaysia Forum, 5th September 2007.
10. Nancy P. and Geetha Ramani R., *A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data*, International Journal of Computer Applications (0975–8887) Volume 32– No.8, October 2011.
11. C. Deepa, K. Sathiya Kumari, and V. Pream Sudha, *A Tree Based Model for High Performance Concrete Mix Design*, International Journal of Engineering Science and Technology Vol. 2(9), 2010, 4640-4646.
12. S. Shanthi and R. Geetha Ramani, *Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms*, International Journal of Computer Applications (0975–8887) Volume 35–No.12, December 2011.