Machine Learning and Statistical Approaches to Classification – A Case Study

Imo Eyoh, Robert John

ASAP Research Group
School of Computer Science
University of Nottingham, UK
Email: (ije, rij)@cs.nott.ac.uk

***Abstract*—The advent of information technology has led to the proliferation of data in disparate databases. Organisations have become data rich but knowledge poor. Users need efficient analysis tools to help them understand their data, predict future trends and relationships and generalise to new situations in order to make proactive knowledge-driven decisions in a competitive business world. Thus, there is an urgent need for techniques and tools that intelligently and automatically transform these data into useful information and knowledge for effective decision making. Data mining is considered to be the most appropriate technology for addressing this need. Datamining is the process of extracting or "mining" knowledge from large amounts of data. Regression analysis and classification are two datamining tasks used to predict future trends. In this study, we investigate the behaviour of a statistical model and three machine learning models (artificial neural network, decision tree and support vector machine) on a large electricity dataset. We evaluate their predictive abilities based on this dataset. Results show that machine learning models, for this real world dataset, outperform statistical regression while artificial neural network outperforms support vector machine and decision tree in the classification task. In terms of comprehensibility, decision tree is the best choice. Although not definitive this research indicates that certainly these machine learning methods are an alternative to regression with certain datasets.**

I. Introduction

The task of predictive modelling is to build classifiers to predict future occurrence of an events, also referred to as supervised learning. These classifiers are datamining algorithms that group data into related classes. They first fit a model as a function of the other attributes of the dataset and later use the fitted model to classify previously unseen instances [7]. Among the datamining classifiers are linear regression, decision tree, support vector machine, artificial neural network, K-nearest neighbour (K-NN), naive bayes and many others [13],[4]. Data mining tasks such as regression and classification have been the mainstay of predictive modelling. While regression tries to model the relationships between data, classification approaches attempt to group data into predefined classes (categorical labels). In these experiments, we evaluate the performances of regression and classification tasks on four classifiers namely: multiple linear regression, decision tree, support vector machine and artificial neural networks. The criteria for evaluation are mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE) and predictive accuracy. The dependent and independent variables for the regression analysis are continuous while the dependent variable for the classification task is categorical.

The rest of the paper is organised as follows: In section II some related literatures are reviewed. Section III highlights the aim and objectives of the study. In section IV, we provide a description of the dataset used in the experiments. Section V gives a brief discussion on multiple linear regression, decision tree, support vector machine, artificial neural network and performance metrics used for comparison. In section VI, we compare the result of these models and summarise the findings. Section VII concludes the study and describes future work.

II. Related Literature

Much research related to the performance of classification methods has been reported. Several authors have compared various classifiers with each other based on various performance metrics and evaluation results have been reported.

Byvatov et al [2] compared support vector machine with artificial neural networks for the classification of drug/nondrug chemical compounds. Evaluation results indicate a slightly higher prediction accuracy of SVM (82%) over ANN (80%).

Timor et al [10] also compared the performance of artificial neural network with support vector machines to model stock price selection problem in Istanbul stock exchange. Different models of support vector machines and artificial neural networks were employed for the analysis. Artificial neural networks exhibit overall higher performance of 81.34% against SVM (75.56%).

Chandra et al [3] present the combination and comparison of artificial neural network with decision tree for classification of wine data. The authors first trained the neural network and later used decision tree to extract IF$\cdots$ THEN rules from the network. In the second stage they used both classifiers for classification and compared their performances. ANN demonstrates high generalisation performance in wine classification with accuracy of 98.7% against decision tree with accuracy of 96.8%.

Kumar et al [6] carry out a performance evaluation of decision tree and artificial neural network based classifiers in diversity of datasets. Three decision tree models - CHAID,

QUEST and C5.0 are compared against artificial neural network with predictive accuracy, training time and comprehensibility as their performance metrics. The classifiers were evaluated on mushroom, vote, nursery and credit datasets. The authors report that artificial neural network and C5.0 have higher predictive accuracies, also decision tree based classifiers displays high comprehensibility and low training time. Artificial neural network on the other hand shows zero comprehensibility and significant training time especially with large datasets.

Tso and Yau [11] compared classification accuracy of regression, decision and artificial neural network models in predicting electricity energy consumption in Hong Kong during winter and summer. Their performance metric is based on square root of average squared error (RASE). The authors claim that the decision tree model and neural network model perform slightly better than the regression models in the summer and winter phases, respectively. They pointed out that the differences in RASE between the three types of models are quite small.

Maliki et al [8] present a comparison of regression and artificial neural network models using electrical power generation in Nigeria. Their performance metrics are based on mean squared error (MSE), mean absolute error (MAE) and root mean squared error (RMSE). Artificial neural network exhibits least error compared to the regression model.

Razi and Athappilly [9] perform comparative analysis of artificial neural networks, non-linear regression and classification and regression tree (CART) using dataset on smoking habits of people. The analysis is evaluated based on the mean absolute percentage error (MAPE), mean squared error (MSE), large prediction error (LPE) and mean absolute error MAE). Results show that neural network and CART exhibit better predictive accuracies than the non-linear regression model.

III. Aim and objectives of Experiments

The aim of these experiments is to model a publicly available electricity dataset using statistical regression and machine learning techniques. The objectives are:

1) To compare the performance of statistical technique with the machine learning techniques.
2) To compare the performance of each machine learning technique in both regression and classification problems.

IV. Dataset Description

The electricity dataset [1] is utilised for these experiments. It consists of seven independent variables and one dependent variable.

The dataset is based on observations obtained from Australian New South Wales (NSW) electricity market. In this market, the electricity prices vary and depend on demand and supply of the market. The dataset consists of 45312 observations (17762 instances with missing values) with 8 attributes namely: (day of the week, time of day, NSW price (NSP), NSW demand (NSD), Victoria region price (VP), Victoria region demand (VD), scheduled electricity transfer between states (ET). Changes in the price based on a moving average of the last 24 hours is used to evaluate the output and the assigned class (1 or -1) for the output is a reflection of the deviations of the price on a one day average. The attributes for the prices and demand are numeric and are scaled using Z-score normalisation scheme with mean 0 and standard deviation of 1. Normalising numeric values helps to speed up the learning process especially when neural network is used for classification. The time of day and day of the week are not applied in this analysis as there are factor variables. For building predictive models, the normalised data set is randomly split into training, testing and validation datasets in the ratio 2/3, 1/6 and 1/6 respectively to ensure that each dataset is a representative of the universal dataset. The validation dataset also known as design dataset is used to evaluate the model performance. The testing dataset is used to evaluate the generalization performance of the models on instances or examples not seen during training. Using the testing dataset to evaluate model performance gives an unbiased estimate of the model error.

V. Predictive Models

In this section we provide a brief description of the models utilised in the analysis.

A. Multiple Linear Regression

Regression analysis is a modelling strategy that predicts continuous variable based on the contribution of other variables in the dataset. The electricity dataset has more than one independent variable and therefore is a multiple linear regression problem with a response variable labelled "ET" and four (4) predictor variables. The model helps to explore the individual effect each predictor variable has on the response variable.

For the regression task, electricity transfer (ET) is adopted as the response variable while NSW price (NSP), NSW demand (NSD), Victoria region price (VP), Victoria region demand (VD) are the independent variables. Both dependent and independent variables are continuous. The class attribute is a factor variable and not contributory to regression analysis. In regression analysis, the ideal is for all independent variable to be correlated with the dependent variables. The independent variables must not be linearly correlated with each other (multicollinearity risk). From Table I, the variables NSD and VD are highly correlated. We may not be sure which one of the variable is explaining the variation in the dependent variable. This can also be seen in the scatter plot in Figure 1. Using both NSD and VD variables, may not provide a distinctive contribution of the variables to the variation of the dependent variable, however, for this experiment, we retain both. From the information in Figure 1, regression analysis can also act as a feature analysis tool to select relevant variables for any modelling tasks.

A multiple linear regression model is developed to explore the relationship between the dataset attributes and the response variable.

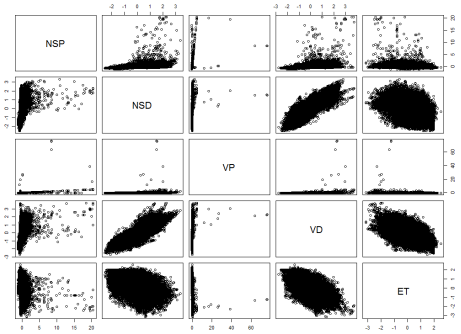

Fig. 1. Electricity scatter Plot

Definition 1. *Let x_1, x_2,...,x_n be a set of attributes in an n-dimensional vector describing each example in the dataset. Then the relationship between the target variable, Y and independents variables x_1, x_2,...,x_n can be represented as a multiple regression model as:*

$$Y = \alpha_0 + \alpha_1 x_1 + ... + \alpha_i x_i + ... + \alpha_n x_n + \varepsilon \quad (1)$$

where i=1,2,...,n are the number of attributes in the dataset α_i are the model coefficients and ε is the random error. The model correspond to the sum of the model parameters:

$$Y = \alpha_0 + \sum_{i=1}^{n} \alpha_i x_i + \varepsilon \quad (2)$$

The correlation among each individual variable is shown in Figure 1 and Table I.

TABLE I
ATTRIBUTE CORRELATION RESULTS

	NSP	NSD	VP	VD	ET	Class
NSP	1.000	0.341	0.321	0.345	-0.215	0.326
NSD	0.341	1.000	0.108	0.839	-0.461	0.355
VP	0.321	0.108	1.000	0.127	-0.094	0.082
VD	0.345	0.839	0.127	1.000	-0.624	0.298
ET	-0.215	-0.461	-0.094	-0.623	1.000	-0.162
Class	0.326	0.355	0.082	0.298	-0.162	1.000

From Table 1, NSD and VD exhibit very high correlation. This shows that these two variables are bounded together and their relationship is almost tending to a perfect line. There is also a strong negative relationship between the two predictor variables with the response variable.

Using electricity dataset, we define a multiple relationship between the response variable labelled "ET" and the attributes of the electricity dataset - NSP, NSD, VP and VD. The multi-variable linear regression is modelled as shown below:

$$ET = \alpha_0 + \alpha_1 NSP + \alpha_2 NSD + \alpha_3 VP + \alpha_4 VD + \varepsilon \quad (3)$$

First the model is fitted to learn the joint relationships among the four predictor variables. To explain the fitted model, the summary of the model is generated as shown in Table II.

TABLE II
REGRESSION MODEL SUMMARY

	Estimate	Std.Error	t value	Pr($>\|t\|$)
Intercept	0.003023	0.005689	0.531	0.5952
NSP	-0.009760	0.006281	-1.554	0.1202
NSD	0.214770	0.010486	20.483	<2e-16
VP	-0.016055	0.006436	-2.494	0.0126
VD	-0.796792	0.010506	-75.842	<2e-16

Residual standard error is 0.7709 on 18355 degrees of freedom. Multiple R-squared is 0.4037, which explains the proportion of the variability in electricity transfer that is explained by the model. Adjusted R-squared is 0.4036. F-statistics is 3107 on 4 and 18355 DF, P−value is <2.2e-16. The model has intercept at 0.003, with coefficients of the four independent variables NSP, NSD, VP and VD at -0.01, 0.215, -0.02 and -0.8 respectively. The estimated Y (output) fitted value becomes:

$$\hat{ET} = 0.003 - 0.01NSP + 0.215NSD - 0.02VP - 0.8VD \quad (4)$$

The negative slope of NSP implies that one unit increase in the New South Wales price leads to a decrease of 0.01 unit in the estimated electricity transfer when all other predictors are held constant. The p-values for each independent variable test the null hypothesis that the coefficients of the variable is equal to zero. A low p-value ($p < 0.05$ level of significance) implies rejection of the null hypothesis and addition of the variable to the model as changes in the value of the variable are related to changes in the response variable. In the regression model summary in Table 2, the independent variables - NSD,VP and VD are significant with p-values < 0.05. The null hypothesis is rejected. However, the p-value for variable NSP is 0.1202 which is greater than 0.05, meaning that we can get rid of NSP as addition of NSP to the model will not be useful. The intercept of the estimated model is 0.003.

B. Decision Tree

Decision tree models are the most widely used traditional analysis tools in predictive modelling that can handle both classification and regression tasks and also facilitate decision making [12]. Their popularity lies in the fact that they are transparent models, easy to understand and interpret as in Figure 2 using the electricity dataset.

The decision tree above generates the rules below:
Rule 1: IF NSP <0.29 THEN class is -1
Rule 2: IF NSP >=0.29 THEN class is 1
The decision tree uses NSP as the splitting criterion to split the training dataset into the two classes (-1 and 1). The root node error is 0.42 while the leaf nodes errors are 0.30 (node 2) and 0.17 (node 3) respectively. According to [12], decision trees do not offer the best performance and represent a trade-off between performance and simplicity of explanation. As shown in Figure 2, a decision tree can be viewed as a hierarchical tree

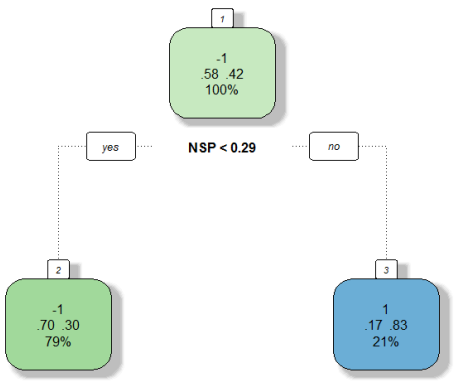

Fig. 2. Decision Tree Classification for electricity dataset

structure with internal nodes representing a test on an attribute, the branch denoting the outcome of the test with leaf nodes (terminal nodes) representing the class labels [4]. The top node represents the parent node from which all descendant nodes are derived. Learning a decision tree involves partitioning the training examples in a recursive (divide-and-conquer) manner to reach a conclusion. During the process of tree construction, some heuristics, called attribute selection measures (splitting rule or criterion) are employed to choose the attribute that best partitions the dataset into distinct classes. This criterion consists of a splitting attribute and either a split point or splitting subset. The most commonly used attribute selection measures are the information gain, gain ratio and gini index. The procedure for learning a decision tree starts as a single node, N, which represents the training examples in the dataset. If the attributes are all of the same class, the learning ends and node N is returned as a leaf node with that class label. If this is not the case, the attribute selection method is called to determine the appropriate splitting criterion at that node. The splitting criterion provides three important information on tree learning:

1) it determines which attribute to test at node say N in order to best partition the dataset into individual classes.
2) it determines which branch to grow from node N based on the outcome of the test.
3) it indicates the splitting attribute and either a split point or splitting subset.

Using the splitting criterion ensures that the partitions obtained are as pure as possible (i.e. all instances in that partition belong to the same class) at each branch. The decision tree implemented is the CART (classification and regression tree). An important feature of CART is its ability to generate both classification and regression trees. In case of regression, CART looks for splits that minimize the prediction squared error (least squared deviation). The prediction in each leaf is based on the weighted mean for node. The attribute selection measure utilise by CART is gini index or information gain. For this experiment, we utilise the information gain.

Definition 2. *Given a p-dimensional training set, N, with n distinct classes, C_i (i= 1,2, ... n). Let $C_{i,N}$ be instances in class C_i, $|N|$ and $|C_{i,N}|$ be the number of instances in N and $C_{i,N}$ respectively. The information required to classify an example in N is computed as follows:*

$$Info(N) = -\sum_{i=1}^{n} p_i log_2(p_i) \tag{5}$$

where $p_i = |C_{i,N}|/|N|$ is the probability that a training instance in the training set belongs to class C_i. This gives the information on the proportion of instances in each class (0 or 1). For each attribute, X with distinct values $(x_1, x_2, x_3...x_j)$, the training set is partitioned into $N_1, N_2...N_j$ subset on X such that N_m represents examples in N with outcome x_m of X, $m = 1, 2...j$. $N_1, N_2...N_j$ are the different branches grown from the parent node (any node that holds the data partition). The information required in order to obtain pure partitions to arrive at exact classification of an instance in N based on the partitioning by X is computed as below:

$$Info_X(N) = \sum_{m=1}^{j} |N_m|/|N| * Info(N_m) \tag{6}$$

where $|N_m|/|N|$ represents the weight of the jth partition. The information gain is then calculated as the difference between original information requirement and the new requirement as:

$$Gain(X) = Info(N) - Info_X(N) \tag{7}$$

The attribute with the highest information gain is adopted as the splitting attribute at the decision node.

C. Support Vector Machines

A support vector machine (SVM) is a classifier that support classification for both linear and non linear data [4]. It takes a set of labeled examples as inputs and produces predicted labels as outputs. Support vector machines operate in two phases [2]:

1) First, the original data vectors are mapped to a very high dimensional space. Given an appropriate non linear mapping with a high dimension, the SVM can separate data from two classes by a hyperplane.
2) Second, the algorithm finds a hyperplane in the new high dimensional feature space with the largest margin separating the classes of data. The hyperplanes are found using support vectors (training examples) and margins which are defined by the support vectors.

Thus, the so-called Maximal Margin Classifiers are the simplest form of the SVMs and handles data that are linearly separable. The separating hyperplane is defined as [4]:

$$f(x) = (w.x) + b_0 \tag{8}$$

where w (the weight vector) and b_0 (the bias) are the parameters of the hyperplane to be estimated by the SVM, x is the input vector that is mapped to a high dimensional space. Given

a training tuple, D in two dimension, D= (x_1, x_2),and bias as additional weight, w_0, the hyperplane can be formulated as:

$$w_0 + w_1 * x_1 + w_2 * x_2 = 0 \tag{9}$$

such that all points above the hyperplane satisfy:

$$w_0 + w_1 * x_1 + w_2 * x_2 > 0 \tag{10}$$

and points lying below the dividing hyperplane satisfy:

$$w_0 + w_1 * x_1 + w_2 * x_2 < 0 \tag{11}$$

The hyperplanes defining the sides of the margin can be re-written as:

$$P_1 : w_0 + w_1 x_1 + w_2 x_2 \geq 1, y_i = +1 \tag{12}$$

where P_1 is the decision boundary for class 1 and any instance falling on or above this hyperplane belongs to class +1

$$P_2 : w_0 + w_1 x_1 + w_2 x_2 \leq 1, y_i = -1 \tag{13}$$

P_2 is the hyperplane for class 2 and any instance falling on or above this hyperplane belongs to class -1. All instances falling on the two hyperplanes are called support vectors and these provide the most information for the classification.

D. Artificial Neural networks(ANNs)

Artificial neural networks are interconnected set of neurons that exhibit some of the behaviours of biological neural networks. Artificial neural networks are used extensively in classification and regression problems because they can model both simple and complex relationships between many input and output variables [8]. Studies show that ANNs can often outperform many statistical and other ML approaches in terms of classification accuracy, generalization, robustness to noise and its ability to model non-linear relationships efficiently where most traditional models struggle. The basic building block of artificial neural networks is the neuron which performs mathematical operations on input data in the form of learning to obtain the corresponding outputs. The structure of ANN consists of the input layer, the hidden layer and the output layer. Before applying ANN, the model is first trained with sample data set which is often split into three sets namely training, testing and validation datasets. The aim of training artificial neural network is basically to obtain optimal weights and biases that minimise some cost function such as mean square error. The trained network is evaluated on the validation sets to ascertain model accuracy while the test set verifies the generalization ability of the model. The actual training of a neural network involves presenting a set of input vectors to the network input layer units. The activation of the inputs is fed forward through the weighted connections to the hidden layer units and finally to the output layer unit(s).

Definition 3. *Given X input vectors in n-dimensional input space, and a unit j in hidden or output layer, the net input, I_j, to unit j is defined as [4]:*

$$I_j = b_j + \sum_{i=1}^{n} w_{ij} O_i \tag{14}$$

where w_{ij} is the connection weight from unit i in the previous layer to unit j in the next layer, O_i is the output of unit i from the previous layer and b_j is the bias of the unit.

Definition 4. *Given the net input I_j to unit j, using the sigmoid function, the output, O_j is expressed as:*

$$O_j = \frac{1}{1 + e^{-I_j}} \tag{15}$$

The parameters of the neural network model are estimated by minimising the cost function (error). The commonly used cost function is the mean square error which is computed as:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (A_i - O_i)^2 \tag{16}$$

where A_i is the actual output, O_i is the network predicted output for each value of the ith input and N is the number of observations. The error obtained is back-propagated from the output unit to the hidden units to obtain the error of the network's prediction. Error on each j output unit is computed as below:

$$Error_j = O_j(1 - O_j)(A_j - O_j) \tag{17}$$

where O_j is the output of unit j, and A_j is the known actual value of the input. The error on each hidden layer is computed with respect to the higher layer as follows:

$$Error_j = O_j(1 - O_j) \sum_{p} Error_p w_{jp} \tag{18}$$

where w_{jp} is the connection weight from unit j to a unit p in the next higher layer and $Error_p$ is the error of unit p.
Weights are updated as follows:

$$\delta w_{ij} = (l) Error_j O_i \tag{19}$$

$$w_{ij} = w_{ij} + \delta w_{ij} \tag{20}$$

Biases are updated as shown:

$$\delta bias_j = (l) Error_j \tag{21}$$

$$bias_j = bias_j + \delta bias_j \tag{22}$$

A 5-4-1 multilayer artificial neural network architecture trained with backpropagation algorithm is adopted for the analysis. The number of hidden units is calculated using Equation 23 [5]

$$h = 2rand\sqrt{v * c} \tag{23}$$

where v is the number of input and c is the number of classes. Rand represent a random number in the interval [0,1]

VI. Performance Comparison of Models

The performance of statistical and machine learning approaches on regression analysis is based on their Mean Square Error (MSE - a measure of the average of the squares of errors), Root Mean Square Error (RMSE - a measure of the spread of the actual observed input values about the predicted values) and Mean Absolute Error (MAE - a measure of the average magnitude of the errors in a set of prediction) as used in [8] and define in [4] as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{24}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{25}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{26}$$

where x_i is the actual observed value for the ith observation, y_i is the predicted value and N is the number of observations in the testing dataset.

Predictive Accuracy is the percentage of the test set correctly classified [4] given as:

$$PredAcc = \frac{True\ Predictions\ from\ all\ classes}{Total\ number\ of\ observations} * 100 \tag{27}$$

Machine learning classifiers: decision tree, artificial neural network and support vector machine are used for the classification task which is a two class problem with the output classified as either 1 or -1 representing increase or decrease in electricity price between states. The purpose of the classification model is to learn a mapping: $y = f(x)$, that separates data into different classes, with x as the input and y as the output (class label) [4] . Ten runs of the experiments were conducted for decision tree, support vector machine and artificial neural network respectively and their averages computed. A confusion matrix (error matrix) is used to calculate the performance of the machine learning techniques on classification task. The models are evaluated using the testing dataset in all cases in order to obtain an unbiased estimate of the results.

A. Regression Analysis Models Comparison

TABLE III
Regression Analysis Results

	MSE	MAE	RMSE
Multiple Regression	0.607	0.78	0.64
Decision Tree	0.549	0.741	0.595
Support Vector	0.390	0.624	0.493
Artificial Neural Network	0.280	0.529	0.421

From the regression results in Table III, machine learning techniques outperform the statistical regression model while ANN outperforms both the support vector and decision tree models. The reason that one classifier outperforms another maybe dependent on the data and parameter settings of the model.

B. Classification Models Comparison

The predictive accuracy is formulated using the confusion matrix (CM). A sample confusion matrix for decision tree, support vector machine and artificial neural network for one run of the experiments are shown in Tables (V, VI, and VII). Each table depicts the actual count of the observations with their corresponding percentages in bracket. Table IV shows

TABLE IV
Misclassified observations from 10 samples

ANN	Decision Tree	SVM
1134	1306	1231
1116	1231	1217
1136	1303	1187
1114	1244	1222
1116	1240	1235
1136	1299	1229
1108	1302	1226
1160	1307	1216
1150	1231	1192
1107	1242	1235

the total misclassified observations taken from each run of the experiments using testing dataset. Artificial neural network displays lowest number of total misclassified instances followed by support vector machine and decision tree. The graph of misclassification rate for each sample is shown in Figure 3

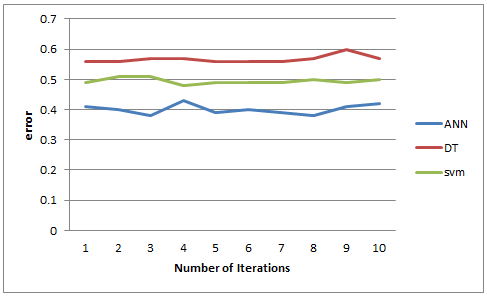

Fig. 3. Misclassification rate from each sample run

TABLE V
Decision Tree Confusion Matrix

	Predicted (cv)	
Actual	-1	1
-1	2534(55)	137(3)
1	1169(25)	752(16)

Table VIII summarises the overall classification accuracies of the three machine learning classifiers on ten (10) different samples of the dataset. Artificial neural network outperforms both the support vector machine and decision tree algorithms.

TABLE VI
SVM CONFUSION MATRIX

	Predicted (cv)	
Actual	-1	1
-1	2400(52)	260(6)
1	975(21)	957(21)

TABLE VII
ANN CONFUSION MATRIX

	Predicted (cv)	
Actual	-1	1
-1	2294(50)	376(8)
1	732(16)	1190(26)

TABLE VIII
SUMMARY OF CLASSIFICATION RESULTS

	Accuracy	Overall error
Decision Tree	72.2%	0.57
Support Vector	73.5%	0.49
Artificial Neural Network	75.5%	0.40

VII. CONCLUSION

From the regression results, machine learning techniques perform better than the statistical multiple regression model on this problem while ANN performs better than the support vector and decision tree regression models. Artificial neural network gives the lowest RMSE which is an indication of a better fit of the dataset.

For the classification task, artificial neural networks outperform the support vector machine and decision tree. During the experiments, we observed that decision tree returned higher false positive values than support vector machine and artificial neural network. This represents a strength of a decision tree as it captures more of the real cases or instances of the dataset.

Generally, although artificial neural networks and support vector machines are non-comprehensible models, they offer better classification and regression accuracies than the transparent decision tree model in this particular instance.

In future, we intend to enhance the capability of artificial neural networks by extracting rules from the networks in order for them to gain wider acceptance in the data mining and machine learning communities.

REFERENCES

[1] http://moa.cms.waikato.ac.nz/datasets/. Accessed: 2015-01-30.

[2] Evgeny Byvatov, Uli Fechner, Jens Sadowski, and Gisbert Schneider. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of Chemical Information and Computer Sciences*, 43(6):1882–1889, 2003.

[3] Rohitash Chandra, Kaylash Chaudhary, and Akshay Kumar. The combination and comparison of neural networks with decision trees for wine classification. *School of sciences and technology, University of Fiji, in*, 2007.

[4] Han Jiawei and Micheline Kamber. Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann*, 2nd Edition, 2006.

[5] Ulf Johansson, Tuve Löfström, Rikard König, and Lars Niklasson. Why not use an oracle when you got one. *Neural Information Processing-Letters and Reviews*, 10(8-9):227–236, 2006.

[6] Pardeep Kumar, V Sehgal, Durg Singh Chauhan, et al. Performance evaluation of decision tree versus artificial neural network based classifiers in diversity of datasets. In *Information and Communication Technologies (WICT), 2011 World Congress on*, pages 798–803. IEEE, 2011.

[7] Daniel T Larose and Chantal D Larose. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

[8] Olaniyi S Maliki, Anthony O Agbo, Adeola O Maliki, Lawrence M Ibeh, and Chukwuemeka O Agwu. Comparison of regression model and artificial neural network model for the prediction of electrical power generated in nigeria. *Advances in Applied Science Research*, 2(5), 2011.

[9] Muhammad A Razi and Kuriakose Athappilly. A comparative predictive analysis of neural networks (nns), nonlinear regression and classification and regression tree (cart) models. *Expert Systems with Applications*, 29(1):65–74, 2005.

[10] Dincer Hasan Timor, Mehpare and Senol Emir. Performance comparison of artificial neural network (ann) and support vector machines (svm) models for the stock selection problem: An application on the istanbul stock exchange (ise) - 30 index in turkey. *African Journal of Business Management*, (3):1191–1198, 2012.

[11] Geoffrey KF Tso and Kelvin KW Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9):1761–1768, 2007.

[12] Graham Williams. *Data mining with Rattle and R: the art of excavating data for knowledge discovery*. Springer Science & Business Media, 2011.

[13] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.