

Electric Power Load Forecast Using Decision Tree Algorithms

*Hambali, M., Akinyemi, A., Oladunjoye, J. & Yusuf, N

Department of Computer Science

Federal University Wukari

Wukari, Taraba State, Nigeria

*Corresponding Author: hamberlite@gmail.com +234-7065868393

ABSTRACT

In the past, the large population of people are rural dwellers, who may not need electricity for their day-to-day activities. However, currently there is an increase in urban drift population to providing quality and quantity of power for day-to-day activities because of the emerging technologies which operation depends solely on electricity. Hence, this study of forecasting electric power load required by the people. One way to achieve quality power generation, transmission, distribution and its marketing is by being able to forecast accurately the energy need of the population; so as to reduce operating cost and maximize usage of the electric power generated at all times. Therefore, unlike many reviewed papers, this paper presents an up-to-date by experimenting using Decision Tree Algorithms – (Classification and Regression Tree) CART, (Reduced Error Pruning Tree) REPTree and Decision Stump for electric power load forecast. The work revealed that, REPTree Decision Tree Technique is suitable to forecast electric load and outperformed other decision tree algorithms. This work will be of considerable usefulness to Yola/Jimeta Power Transmission Company and others involved in Power transmission, Generation, Distribution and Marketing industry to enable them forecast electric power load and provide appropriate advising and decisions in timely manner.

Keywords- Data Mining; REPTree; Load Forecast; Artificial Neural Networks; Decision Tree Algorithms.

CISDI Journal Reference Format

*Hambali, M., Akinyemi, A., Oladunjoye, J. & Yusuf, N. (2016); Electric Power Load Forecast Using Decision Tree Algorithms. Computing, Information Systems, Development Informatics & Allied Research Journal. Vol 7 No 4. Pp 29-42
Available online at www.cisdijournal.net

1. INTRODUCTION

The fundamental function of an electric power company is to provide customers with high quality electric energy in secured and economical manner. In order to do so, an electric power company faces challenges in economic and technical problems in planning, control and operation of electric power system. For the purpose of optimal planning and operation of electric power system, there is need for proper evaluation of present day and future electric power load (Samsher and Unde, 2012; Mohammed and Sanusi, 2012; Olagoke, Ayeni and Hambali, 2015). Load forecasts are very important for energy suppliers and other participants in electric energy generation, transmission, distribution and markets. Precise and accurate models for electric power load forecasting are crucial to the operation and planning of a utility company. Load forecasting assists an electric transmission company to make vital decisions including decisions on acquiring and generating electric power, load switching, and infrastructure development. The issue on load forecasting has been in existence for few decades to forecast the future demand of electricity. This includes the accurate prediction of both the magnitudes and geographical locations of electric load over the periods of time. Electricity load forecasting is considered as one of the critical factors for economic operation of power systems, and accurate load forecasting holds a great saving potential for electric transmission corporations. The maximum benefits can be attained when load forecasting is employed to control operations and decisions such as economic dispatch, unit commitment and maintenance, and fuel allocation. (Arunesh, Ibraheem and Muazzam, 2013).

Electric load forecasts can be grouped into three types based on the planning perspective of the duration: Short Term Load Forecasts (STLF): This is usually for one hour up to one week. Medium Term Load Forecasts (MTLF): This is usually for one week up to few months. Long Term Load Forecasts (LTLF): This is longer than a year and more (Olagoke, Ayeni and Hambali, 2015). The forecasts for different time horizons are essential for different operations in electric power company. The nature of these forecasts are different as well. For instance, for a particular region, it is possible to predict the next day load with an accuracy of approximately 1-3%. But, it is impossible to predict the next year peak load with similar accuracy, since accurate long-term weather forecasts are not available (Eugene and Dora, 2006). STLF can help electric load planners to estimate load flows and make decisions that prevent overloading. Timely implementations of such decision lead to the improvement of network reliability and reduce occurrences of equipment failures and blackouts. MTLF is useful in unit maintenance and determining the quantity of fuel to purchase in power plants. LTLF used to supply electric power company management with prediction of future needs for expansion, equipment purchases and inter-tie tariff setting (Eugene and Dora, 2006).

However, load forecast is a complex task because the consumption is influenced by many factors, such as day type, anomalous days, weather conditions, vacations, economy factors, status and idiosyncratic of individual customers' habit (Nahi *et al.*, 2006). Weather conditions influence the load forecasting. In fact, forecasted weather parameters are the most important factors in load forecast. Various weather variables could be considered for load forecasting. Temperature and humidity are the most commonly used load predictors (Olagoke, Ayeni and Hambali, 2015). In the recent past years Artificial Intelligent (AI) and numerous searching algorithms are employed in this domain to improve the accuracy of load forecast. According to Eugene and Dora (2006), erroneous load forecasting can lead to rise in operating costs. For example, over forecast may result in a redundant reserve of electric power therefore, increase the operating cost. On the contrary, under forecast causes failure in providing sufficient electric power. A poor load forecast deludes planners and often results in erroneous and expensive expansion plans (Islam, et al., 2014).

Data mining techniques, including different methods of artificial intelligence (AI), machine learning (ML) and statistics, are computational processes of discovering the valuable information from large data sets. Decision tree (DT) algorithm has gained prominent interests because it not only provides the intuition information of data sets with minimal computational burden, but also divulges the principles learnt by DTs for further interpretation (Chengxi, et al., 2013). The focus of this work is to identify the optimal decision tree algorithms for electric load forecasting. Three decision tree algorithms (Classification and Regression Tree (CART), Decision Stump and Reduced Error Pruning Tree (REPTree)) were applied on the data set and their results were compared using 10-fold cross validation in terms of classification accuracy, errors and execution time. The optimal algorithm will be implemented STLF for Yola Power Transmission Company, Jimeta/Yola, Adamawa State Nigeria.

The rest of this paper is organized as follows. Section 2 presents related work to this study. In section 3, we described the methodology used. Section 4 contains discussion of the results. Finally, conclusions and future work are presented in section 5.

2. RELATED WORKS

For decades now, the researchers had been striving hard on the ways of improving the accuracy of load forecasting. Research in this area in the last few years has resulted in the development of numerous forecasting methods (Gwo-Chung and Ta- Peng, 2006). These methods are mainly classified into two categories: Classical approaches and artificial intelligence (AI) based techniques. Classical approaches including various statistical modeling methods such as time- series, regression, exponential smoothing, Box-Jenkins model and Kalman filters. One of these classical methods is a weather-insensitive approach which used historical load data to forecast the future electric load. It is well-known as Box-Jenkins' ARIMA (Box and Jenkins, 1970; Vemuri, Hill and Balasubramanian, 1973; Chen, Wang and Huang, 1995; Wang and Schulz, 2006), which based on theoretical univariate time sequences. However, these classical methods cannot properly represent the complex non-linear relationships that exist between the load and series of other factors that influence it (Samsher and Unde, 2012).

As from 1990s, researchers began to employ different approaches for load forecasting other than classical approach. The emphasis shifted to the implementation of various AI techniques for load forecasting (LF). AI techniques such as neural network, fuzzy logic and expert systems have been applied to deal with the non-linearity, large data sets requirement in implementing the LF modeling and other difficulties in modeling of classical methods used for the application of LF (Samsher and Unde, 2012; Olagoke, Ayeni and Hambali, 2015). Christianse (1971) and Park et al. (1991) employed exponential smoothing models by Fourier series transformation to forecast electric load. Douglas et al. (1998) verifying the impacts of forecasting model in terms of temperature. They combined Bayesian estimation with dynamic linear model into load forecasting. The outcome of experiment show that the presented model is suitable for predicting load with imperfect weather information. The drawback of these methods is time consuming, especially in a situation where the number of variables is increased. Recently, to prevent a lot of variables selection problem, Azadeh et al. (2008) apply fuzzy system to provide an ideal rule base in order to select the type of ARMA models to use, and the results also show that the integrated approach outperform those novel intelligent computing models. Wang et al. (2008) proposed hybrid ARMAX (auto-regressive and moving average with exogenous variables) model with particle swarm optimization to efficiently proffer solution to the problem of trapping into local minimum which is caused by exogenous variable (e.g., weather condition). Their results indicate that the proposed approach has superior forecasting accuracy.

To improve the accuracy of load forecasting, state space and Kalman filtering technologies were developed to reduce the difference between actual loads and prediction loads (random error) for load forecasting model. This approach introduces the periodic component of load as a random process. It needs historical data between 3–10 years to build the periodic load variation and to evaluate the dependent variables (load or temperature) of power system (Brown, 1983; Gelb, 1974; Trudnowski, McReynolds and Johnson, 2001). Moghram and Rahman (1989) proposed a model based on this technique and verified that the proposed model outperforms four other forecasting methods (multiple linear regression, time series, exponential smoothing, and knowledge-based approach). The drawback of these methods is difficult to evade the noise observation in the forecasting process specially when dealing with multivariable. Recently, Al-Hamadi and Soliman (2006) use fuzzy rule-based logic, by utilizing current weather data as well as the recently past history of load and weather data, to recursively estimate the optimal fuzzy parameters for each hour load of the day.

Amjady (2007) employs hybrid model of forecast-aided state estimator (FASE) and multi-layer perceptron (MLP) neural network to forecast short-term bus load of power systems. The proposed hybrid model was used on a real power system, and the results indicate that the hybrid method has better prediction accuracy compare to other models, such as MLP, FASE, and the periodic auto-regression (PAR) model. Another concept employed is regression models which build causal-effect relationships between electric load and independent variables. The most famous models are linear regression, proposed by Asbury (1975) that include weather variable into forecasting model. Papalexopoulos and Hesterberg (1990) added holiday and temperature as another factors in their proposed model. This proposed model used weight least square method to achieve robust parameter estimation encountering with the heteroskedasticity. Soliman et al. (1997) designed a multivariate linear regression model for load forecasting, they used temperature, wind cooling/humidity factors on their model.

The experimental results show that the proposed model outperforms the harmonic model as well as the hybrid model. Similarly, Mirasgedis et al. (2006) also integrate weather meteorological variables, like relative humidity, heating, and cooling degree-days to forecast electricity demand in Greece. Mohamed and Bodger (2005) incorporate economic and geographic variables (such as GDP, electricity price, and population) to forecast electricity consumption in New Zealand. Their model is based on linear assumption, though, these independent variables are unjustified to be considered because their model is known to be nonlinear. Recently, Tsekouras et al. (2007) present a nonlinear multivariable regression approach to forecast annual load, they used correlation analysis with weighting factors to choose appropriate input variables.

In the recent decade, lots of researches had tried to apply the AI techniques to improve the accuracy of the load forecasting issue. Knowledge-based expert system (KBES) and artificial neural networks (ANNs) are the prevalent methods used (Rahman and Bhatnagar, 1998; Chio, Kao and Cook, 1997; Rahman and Hazim, 1993). Recently, applications of fuzzy inference system and fuzzy theory in load forecasting are also receiving attentions, Ying and Pan (2008) present adaptive network fuzzy inference system (ANFIS), by mapping relation between the input and output data to determine the optimal distribution of membership functions, to forecast regional load. Pai (2006) and Pandian et al. (2006) also employ fuzzy approaches to obtain superior performance in terms of load forecasting.

In Figueiredo *et al.* (2005) works on the load forecast issue of EDP (Portuguese Distribution Company). Firstly, they used unsupervised learning (clustering) to obtain partitions of historical data into a set of consumer classes, then supervised learning (DT) is implemented to define each class by rule-based classifications and create a DT model to assign consumers to the existing classes. The objective of their research is to find the relevant knowledge about how and when consumers use electricity. In Chien and Yuafi (1994), DT is applied to estimate the line flows and bus voltage following an outage event in an efficient manner. The approach has been successfully tested by Taiwan system in China. In Ugedo, et al. (2005) designed an approach for southern Spanish generation company to determine the daily load patterns and their associated probability of non-connected unit. DT is used to identify the load pattern so as to approximately predict when its generating units are connected to improve the network constraints.

Guo and Niu (2008) propose a new model which first identifies the different patterns of daily load using data mining technology of classification and regression tree (CART), it considered features such as weather and data type by means of recognition. It then sets up pattern bases which are composed of daily load data sequence and employ artificial neural network to forecast model based on a pattern base which matches the forecasting day. This simplified model is said to reflect the daily load accurately and improve forecasting precision. Prakash Ranganathan and Nygard (2011) suggested an approach that uses M5 decision tree classifiers to predict the electric load demand. However, researchers did not take into consideration other factors such as weather and nature of the day when predicting.

3. METHODOLOGY

The proposed methodology used in this work for load forecast pattern for Yola power Transmission Company using decision tree algorithms belongs to the process of Knowledge Discovery and Data Mining. The stages in the process include the following:

Data gathering: The historical data collected for three month at the Yola power transmission company office along Numan road Jimeta/Yola, Adamawa State, Nigeria. These data were for the month of September, October and November 2015. The parameters found in this dataset are as follows: Date, Time (hourly record), Temperature for 24 hours daily, Input voltage and Output voltage. The input voltage is in coming voltage or load into the Transmission Company which is high voltage of 132kv and before it step-down into the lower voltage levels of 33kv/11kv/0.415kv. The first two months (8weeks) was used for training and the remaining one month is used for both validation and testing of the algorithms. Table 1 shows a sample of a day dataset.

Note:

Oil Temperature is the temperature of the oil used by the transformer and it is recorded on an hourly bases which varies with time and weather. Wind Temperature is the temperature of the windings of the coil inside the transformer it is also recorded on an hourly bases.

Table 1: Electric Company Dataset

DATE	HOURS	TEMP	OIL TEMP	WIND TEMP	INPUT VOLTAGE	OUTPUT VOLTAGE
30/9/2015	9	25	40	42	128	32
30/9/2015	10	25	40	42	132	33
30/9/2015	11	25	40	42	132	33
30/9/2015	12	25	40	42	128	32
30/9/2015	13	25	42	44	128	32
30/9/2015	14	25	44	46	132	33
30/9/2015	15	25	44	46	132	33
30/9/2015	16	33	44	46	132	33
30/9/2015	17	33	44	46	122	30.5
30/9/2015	18	33	44	46	120	30
30/9/2015	19	33	44	46	128	32
30/9/2015	20	33	44	46	116	29
30/9/2015	21	33	44	46	120	30
30/9/2015	22	23	44	46	128	32
30/9/2015	23	23	44	46	128	32

Fig. 1 shows spreadsheet interface of the data set used for training the decision tree algorithms.

DATE	HOURS	WEATHER TEMP	OIL TEMP	WIND TEMP	INPUT VOLTAGE	OUTPUT VOLTAGE
01/09/2015	1	22	38	40	124	31
01/09/2015	2	22	38	40	124	31
01/09/2015	3	22	38	40	124	31
01/09/2015	4	22	38	40	116	29
01/09/2015	5	22	38	40	112	28
01/09/2015	6	24	38	40	128	39
01/09/2015	7	24	38	40	124	31
01/09/2015	8	24	38	40	120	30
01/09/2015	9	24	38	40	124	31
01/09/2015	10	31	42	44	128	32
01/09/2015	11	31	42	44	126	31.5
01/09/2015	12	31	42	44	128	32
01/09/2015	13	31	42	44	126	30
01/09/2015	14	31	42	44	128	31
01/09/2015	15	31	42	44	135	33.8
01/09/2015	16	29	42	44	134	33.5
01/09/2015	17	29	42	44	136	33.5
01/09/2015	18	29	42	44	124	34
01/09/2015	19	29	42	44	136	31
01/09/2015	20	29	42	44	136	34
01/09/2015	21	29	42	44	128	34

Fig. 1: Spreadsheet Interface of the Training Data Set

- **Pre-processing:** This stage involve dataset preparation before applying data mining techniques. At this stage, traditional pre-processing methods such as data cleaning, transformation of variables and data partitioning were applied. Also, other techniques such as attributes selection and re-balancing of data were employed in order to solve the problems of high dimensionality and imbalanced data that may be present in the dataset.
- **Data Mining:** In this stage, data mining algorithms are applied in order to forecast electric load. In doing this, decision tree (DT) algorithms such as CART, REPTree and Decision Stump are employed and compared.
- **Interpretation:** At this stage, the obtained models are analyzed to determine the pattern in load forecasting model.

This section is done by observing the factors that appeared (in the rules and decision trees) and how they are related for consideration and interpretation.

3.1 Experimental Tool Used

The experimental tool used was WEKA. WEKA (Waikato Environment for Knowledge Analysis) is used for forecasting electric load data in this work. Weka is one of the popular suites of machine learning software developed at the University of Waikato. It is open source software available under the GNU General Public License. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality.

3.2 Decision Tree Algorithms

A decision tree is a flow-chart tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals (Abdulsalam et al., 2015). Each of the internal nodes have two or more children node and the internal nodes contain splits, which test the value of an expression of the attributes. Arc from an internal node to its children are labelled with district outcomes of the test. Each leaf node has a class label associated with it.

The decision tree classifier has two phases:

- Growth phase or Build phase.
- Pruning phase.

The tree is constructed in the first stage by recursively splitting the training set based on local optimal criteria until all or most of the records belonging to each of the time is partitioned and assigned the same class label. The tree may over-fit the data (Han and Kamber, 2006). The pruning phase resolves the problem of over-fitting the data in the decision tree. The prune phase generalizes the tree by eliminating the noise and outliers. The accuracy of the classification improving in the pruning phase. Pruning phase accesses only the fully grown tree. The growth phase requires multiple passes over the training data. The time required for pruning the decision tree is very less compared to build the decision tree.

The decision tree algorithms used in this work are briefly described in the following section.

3.3 Classification and Regression Tree (CART)

CART was introduced by Breiman et al. (1984). It is a non-parametric decision tree learning technique that yields either classification or regression trees, depending on the nature of the variable (categorical or numeric). Decision trees are generated by a group of rules based on variables in the modeling data set (Steven, 2014). Rules based on variables' values are chose to develop the best split in order to differentiate observations based on the dependent variable. Once a rule is selected and splits a node into two, each "child" node also undergo the same process (i.e. it is a recursive procedure of splitting a node). Splitting halts when CART detects no further gain can be made, or some pre-set stopping rules are met. Alternatively, the data are split as much as possible and then the tree is later pruned. Each branch of the tree ends in a terminal node. Each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

CART is characterized by the fact that it produces binary trees, each internal node has exactly two outgoing edges, while both ID3, C4.5 algorithms yield the decision trees with variable branches per node. CART is a special Hunt's based algorithm due to the fact that, it is used for regression analysis with the help of regression trees. The regression analysis feature is employed in forecasting a dependent variable (result) given a set of predictor variables over a given period of time. The CART decision tree is a binary recursive partitioning procedure proficient in processing continuous and nominal attributes both as targets and predictors. In CART, trees are grown using Gini Index attribute selection for building and splitting procedure, to a maximum size without the use of topping rule and then pruned back (essentially split by split) to the root through cost-complexity pruning. The cost complexity pruning used is to eliminate the unreliable branches from the decision tree to improve the accuracy. The CART mechanism is intended to produce not one, but a sequence of nested pruned trees, all of which are candidate optimal trees.

$$\text{Gini Index: } 1 - \sum_j p_j^2 \quad (1)$$

Gini index of a pure table which consist of single class is zero because the probability is 1 and $1-1=0$. Similar to Entropy, Gini index also reaches maximum value when all classes in the table have equal probability (Abdulsalam, et al, 2015).

Algorithm 1 Pseudocode for CART tree construction by exhaustive search

1. Start at the root node.
2. For each X , find the set S that minimizes the sum of the node impurities in the two child nodes and choose the split $\{X^* \in S^*\}$ that gives the minimum overall X and S .
3. If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn.

3.3 Reduced Error Pruning Tree (REPTree)

REPTree algorithm is founded on the principle of calculating the information gain with entropy and reducing the error arising from variance (with back-fitting) (Witten and Frank, 2005). The benefit of this method is that, complexity of decision tree model is decreased by reduced error pruning method and the error arising from variance is reduced (Bouckaert et al, 2008). It is a fast decision tree learner. The algorithm only sorts values for numeric attributes once. The missing values are tackled by splitting the corresponding instances into pieces.

REPTree uses the regression tree logic and builds multiple trees in different iterations. Subsequently it selects the best one from all generated trees and consider it as the representative of the generated trees. In pruning the tree, the mean square error measure is used on the predictions made by the tree.

3.4 Decision Stump

Decision stumps are special decision trees with a single layer. As contrast to a tree which has multiple layers, a stump mostly stops after the first split. Decision stumps are typically used in population segmentation for huge data. Seldom, they are also used to help build simple yes/no decision model for smaller data with little data (Murphy, 2010). Decision stumps (DS) are generally simple to build as compared to decision tree. This is due to the fact that the DS is just one single run of the tree algorithm and thus does not require to prepare data for the successive splits which make renaming of output simpler to manage.

That is, DS are one level DTs that classify instances by sorting them based on feature values (Brighton and Mellish, 2002). Individual node in a DS denotes a feature in an instance to be classified, and each branch signifies a value that the node can take. Instances are classified starting at the root node and sorting them based on their feature values. At worst scenario, a DS will replicate the most common sense baseline; and it might do better if the selected feature is informative. DSs are usually used as components called "weak learners" or "base learners" in machine learning ensemble techniques such as bagging and boosting (Reyzin and Schapire, 2006).

Subject to the nature of the input feature, several variations are possible. For example, in nominal features, one may build a stump that represents each possible feature value with a leaf (Loper, et al, 2009) or a stump with the two leaves, one for some chosen category, and the other leaf to represent all the other categories. For binary features these two schemes are identical and a missing value may be treated as another category. For continuous features, usually, a threshold value is specified, and the stump contains two leaves for values below and above the threshold. Though, seldom, multiple thresholds may be specified and the stump therefore contains three or more leaves.

A DS makes a prediction based on the value of a particular input feature. It sometimes called 1-rules. It is a tree with only one split, therefore it is a stump. DS algorithm searches all possible value for each attribute. It chooses best attribute based on minimum entropy. Entropy is measure of uncertainty. We measure entropy of dataset (S) with respect to each attribute. For each attribute A , one level computes a score measuring how well attribute A separates the classes (Wyne and Pat, 1992).

$$\text{Score}(A) = \frac{\max\{H(A|C), H(A|\bar{C})\}}{H} \quad (2)$$

4. RESULT AND DISCUSSION

In this study, three decision tree classification models were proposed for the purpose of forecasting electric power load for Yola/Jimeta Power Transmission station and performance evaluation of three models were made using both 10-fold cross validation method based classification accuracy, error reports and execution time. Table 2 shows the results of the experiment and Figures 2, 3 and 4 depict graphical representation of the results.

4.1 Evaluation Metrics

In selecting the appropriate algorithms and parameters that best model the Electric power load forecasting variable were briefly discuss; the following performance metrics were used:

- **Time:** This is referred to as the time required to complete training or modelling of a dataset. It is represented in seconds
- **Kappa Statistic:** A measure of the degree of non-random agreement between observers or measurements of the same categorical variable.
- **Mean Absolute Error (MAE):** Mean absolute error is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error.
- **Root Mean Squared Error (RMSE):** Mean-squared error is one of the famous used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value. The mean-squared error is simply the square root of the mean squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.
- **Root Relative Squared Error (RRSE):** Relative squared error is the total squared error made relative to what the error would have been if the prediction had been the average of the absolute value. As with the root mean-squared error, the square root of the relative squared error is taken to give it the same dimensions as the predicted value.
- **Relative Absolute Error (RAE):** Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values.

Table 2: Results of the Experiments

Performance Metrics	DECISION TREE ALGORITHMS		
	CART	REPTree	Decision Stump
Correctly Classification (%)	87.5625	87.9914	48.1158
Incorrectly Classification (%)	12.4375	12.0086	51.8942
Kappa Statistics	0.8459	0.8511	0.3155
MAE	0.0219	0.0213	0.0615
RMSE	0.1064	0.105	0.1754
RAE (%)	27.063	26.3239	75.9651
RRSE (%)	52.9656	52.265	87.298
Time (second)	17.17	0.43	0.04

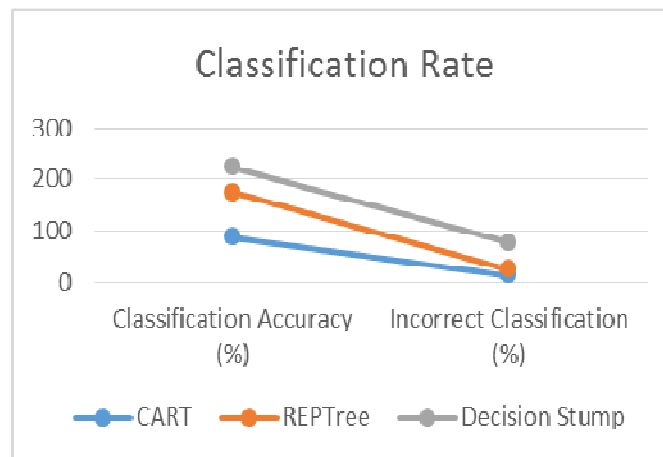


Fig. 2: Correct and Incorrect classification Result

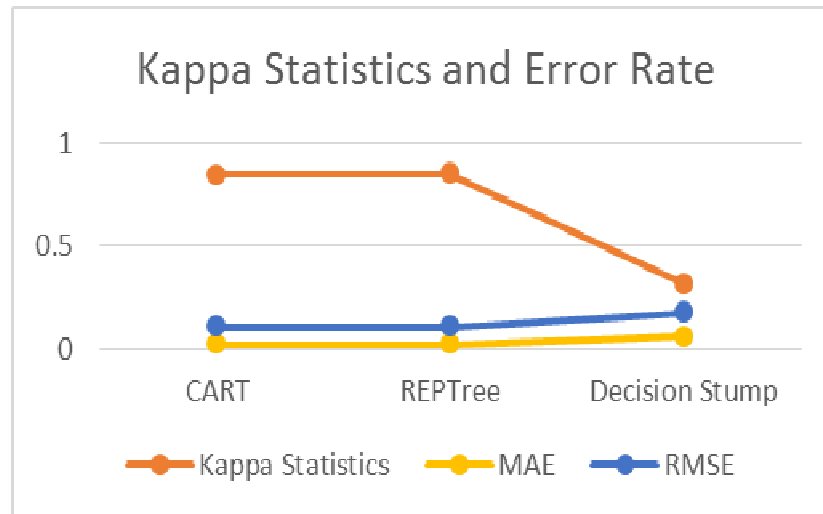


Fig. 3: Kappa statistics and Error Rate

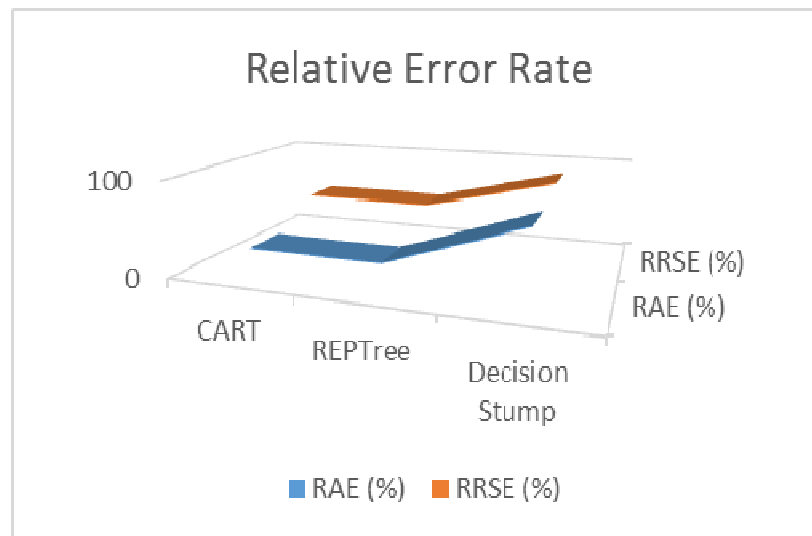


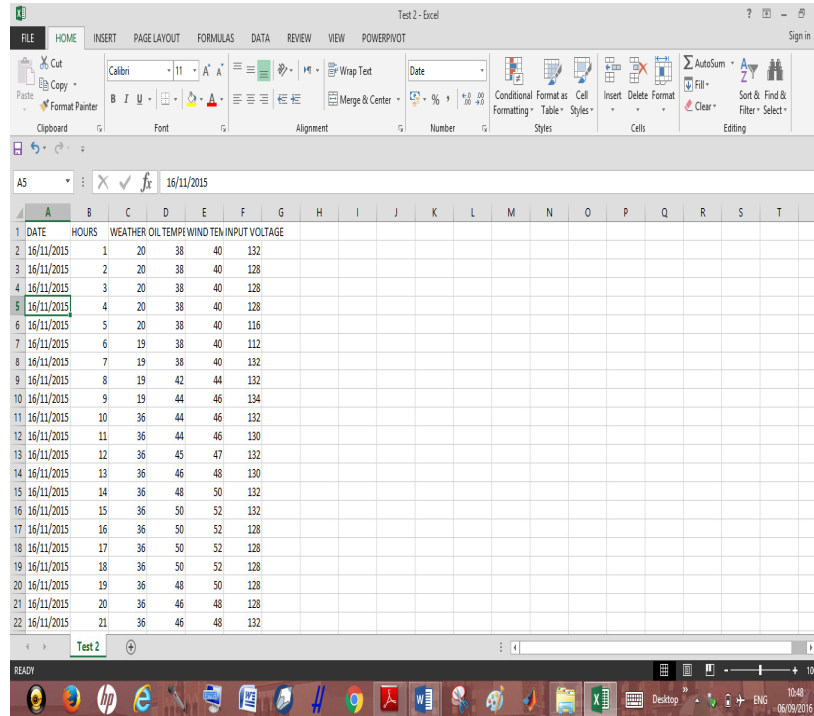
Fig. 4: Relative Error Rate

From the table 2, three Algorithms were used for Decision Tree Model (CART, REPTree and Decision Stump). For the Electric Power load forecasting, CART algorithm used 17.17 secs to model, with correctly classification of 87.56%, kappa statistic of 0.8459, mean absolute error of 0.0219 and root mean square error of 0.1064 while REPTree algorithm was modeled within 0.43 secs, with correctly classification of 87.99%, kappa statistic of 0.8511, mean absolute error of 0.213 and root mean square error of 0.105. Also, Decision Stump algorithm used 0.04 sec to modeled, with kappa statistic of 0.3155, mean absolute error of 0.0615 and root mean square error of 0.1754.

Finally, from the result analysis by comparing the techniques, REPTree performs better than other DT algorithms used based on the error report, number of correctly classified instances and accuracy rate generated. Though there is very close tight between REPTree and CART except in the time require to build the model that REPTree used 0.43 secs and CART used as much as 17.17 secs to build the model.

4.2 Implementation of REPTree to Forecast Electric Power Load

At this stage, two week dataset (Fig 5) were used to test and validate the results. The Dataset is pre-processing using NumericToNormal method in order to reduce the problems of high dimensionality and imbalanced data that may be present in the dataset.



DATE	HOURS	WEATHER	OIL TEMPE	WIND TEN	INPUT VOLTAGE
16/11/2015	1	20	38	40	132
16/11/2015	2	20	38	40	128
16/11/2015	3	20	38	40	128
16/11/2015	4	20	38	40	128
16/11/2015	5	20	38	40	116
16/11/2015	6	19	38	40	112
16/11/2015	7	19	38	40	132
16/11/2015	8	19	42	44	132
16/11/2015	9	19	44	46	134
16/11/2015	10	36	44	46	132
16/11/2015	11	36	44	46	130
16/11/2015	12	36	45	47	132
16/11/2015	13	36	46	48	130
16/11/2015	14	36	48	50	132
16/11/2015	15	36	50	52	132
16/11/2015	16	36	50	52	128
16/11/2015	17	36	50	52	128
16/11/2015	18	36	50	52	128
16/11/2015	19	36	48	50	128
16/11/2015	20	36	46	48	128
16/11/2015	21	36	46	48	128
16/11/2015	22	36	46	48	132

Fig 5: Testing Dataset

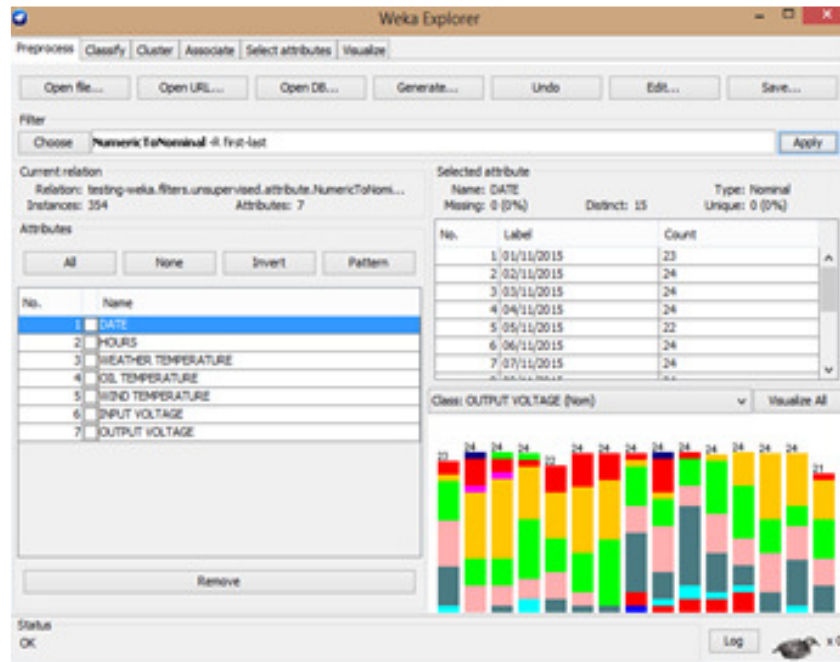


Fig 6: Pre-Processed Dataset

After that, the pre-processed dataset is supply into the WEKA using 10-fold cross validation, the following result were obtained (Table 3):

Table 3: Result Summary of Test Experiment

Size of the tree : 15													
Time taken to build model: 0.02 seconds													
=== Predictions on test data ===													
inst#, actual, predicted, error, probability distribution													
1	6:32	6:32	0	0	0	0	0.013	*0.933	0.013	0	0.04	0	0
2	6:32	6:32	0	0	0	0	0.013	*0.933	0.013	0	0.04	0	0
3	6:32	6:32	0	0	0	0	0.013	*0.933	0.013	0	0.04	0	0
4	6:32	6:32	0	0	0	0	0.013	*0.933	0.013	0	0.04	0	0
5	6:32	6:32	0	0	0	0	0.013	*0.933	0.013	0	0.04	0	0
6	6:32	6:32	0	0	0	0	0.013	*0.933	0.013	0	0.04	0	0
7	6:32	6:32	0	0	0	0	0.013	*0.933	0.013	0	0.04	0	0
8	6:32	6:32	0	0	0	0	0.013	*0.933	0.013	0	0.04	0	0
9	6:32	6:32	0	0	0	0	0.013	*0.933	0.013	0	0.04	0	0
10	4:30	4:30	0	0	0	0	*0.906	0.019	0.038	0	0	0	0.038
....													
...													
=== Stratified cross-validation ===													
=== Summary ===													
Correctly Classified Instances	316	89.2655 %											
Incorrectly Classified Instances	38	10.7345 %											
Kappa statistic	0.8661												
K&B Relative Info Score	27677.2555 %												
K&B Information Score	740.8283 bits	2.0927	bits/instance										
Class complexity order 0	927.1212 bits	2.619	bits/instance										
Class complexity scheme	13069.3542 bits	36.9191	bits/instance										
Complexity improvement (Sf)	-12142.2329 bits	-34.3001	bits/instance										
Mean absolute error	0.033												
Root mean squared error	0.1322												
Relative absolute error	22.3708 %												
Root relative squared error	48.7697 %												
Total Number of Instances	354												

From Table 3, correctly classification is improve from 87.99% to 89.27%, as well as all other error reports. This can be explained as effect NumericToNormal pre-processing method and small size of dataset used in testing the model.

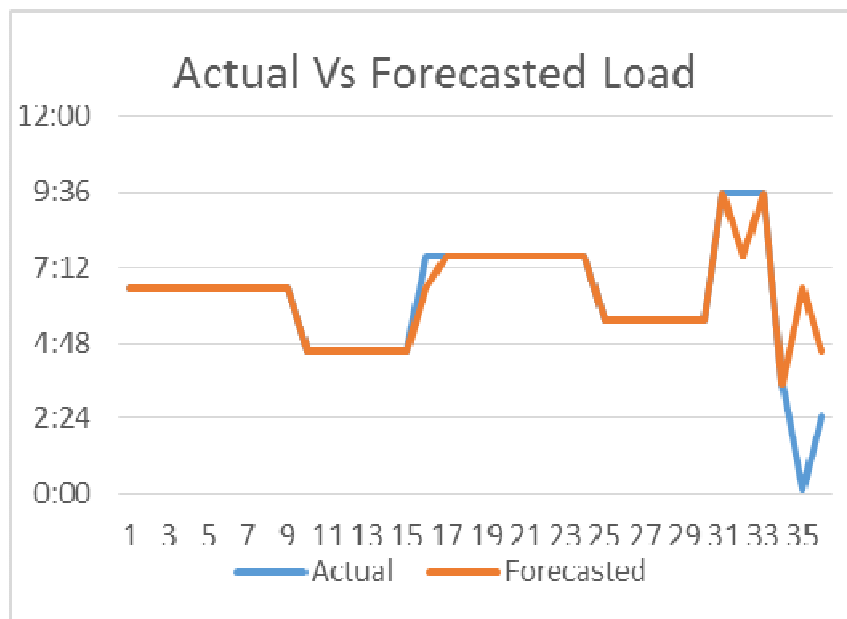


Fig 6: Comparison of Forecasted Load and Actual Load

Figure 6 shows the comparison between actual load data obtained from the power transmission company which is very close to the results obtained from the trained REPTree model output data.

5. CONCLUSION

Forecasting the electric power load is a difficult task. Different approaches to forecast electric power load take into accounts properties of different weather variables; such as temperature, humidity and geographical locations, which are the commonly used load predictors. Hence, a new method for electric load forecasting was developed using Decision Tree pre-processed to reduce effect of missing data and data imbalance.

The work revealed that, REPTree Decision Tree Technique is suitable to forecast electric load and outperformed other decision tree algorithms used with lower error metrics and higher correctly classification instances. This method saves much laboratory needed effort, time and operating costs. Tools used have a more intuitive and easy to use interface, with parameter free data mining algorithms to simplify the configuration and executing and with good visualization facilities to make the results meaningful to take vital decision in equipment acquisition and electric power generation, load switching and infrastructural development.

Future work can be focused on how these processes can be carried out in a beneficial mode with less amount of time and resources. Standardization of data, and the pre-processing, discovering and post-processing tasks needed can be improved upon.

REFERENCE

1. Abdulsalam S. O., Babatunde A. N., Hambali M. A. and Babatunde R. S. (2015). Comparative Analysis of Decision Tree Algorithms for Predicting Undergraduate Students' Performance in Computer Programming. *Journal of Advances in Scientific Research & Its Application (JASRA)*, 2, Pg. 79 – 92.
2. Alpaydin E. (2004) *Introduction to Machine Learning*. The MIT Press, Printed and Bound in the United States of America. ISBN: 0-262-01211-1
3. Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA.
4. Bouckaert R. R., Frank E., Hall M., Kirkby R., Reutemann P., Seewald A. and Seuse D. (2008). WEKA Manual for 3.6.0, [http://prdownloads.sourceforge.net/weka/WekaManual 3.6.0.pdf?download](http://prdownloads.sourceforge.net/weka/WekaManual%203.6.0.pdf?download). [Access: 24 June 2015].
5. Brighton, H. and Mellish, C. (2002). *Advances in Instance Selection for Instance Based Learning Algorithms*, Data Mining and Knowledge Discovery, Pg. 153–172.
6. Loper, E. L., Bird, S. and Klein, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly. ISBN 0-596-51649-5.
7. Murphy C. (2010). *Building Decision Trees from Decision Stumps*. University College Dublin Peter Flom, Consulting.
8. Reyzin, L. and Schapire, R. E (2006). How Boosting the Margin Can Also Boost Classifier Complexity, in *ICML'06: Proceedings of the 23rd International Conference on Machine Learning*, Pg. 753-760.
9. Witten I. H., Frank E. (2005). *Data Mining Practical Machine Learning Tools and Techniques –2nd ed.* the United States of America, Morgan Kaufmann Series in Data Management Systems.
10. Ying-Chun G. and Dong-Xiao N. (2008). Intelligent Short-Term Load Forecasting Based on Pattern-Base. In *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics*, Kunming.
11. Zhang J. G., Deng H. W. (2007). Gene Selection for Classification of Microarray Data Based on the Bayes Error. *BMC Bioinformatics*, 8, Pg. 370
12. Samsher K. S. and Unde M. G. (2012). Short Term Load Forecasting Using ANN Technique. *International Journal of Engineering Sciences and Emerging Technologies*, 1 (2), Pg. 99-107.
13. Mohammed B. and Sanusi S. A. (2012). Short- Term Load Forecasting Using ANN. *Proceedings of the International Multi-conference of Engineers and Computer Scientist IMECS 2012, Hong Kong Vol. 1* Pg. 978-988.
14. Mahrufah D. Olagoke, A. A. Ayeni and Moshood A. Hambali (2016). Short Term Load Forecasting Using Neural Network and Genetic Algorithm. *International Journal of Applied Information Systems*, 10(4): Pg. 22 – 28, January 2016. Published by Foundation of Computer Science (FCS), NY, USA. DOI: 10.5120/ijais2016451490.
15. Arunesh K. Singh, Ibraheem, S. Khatoon, Md. Muazzam (2013). An Overview of Electricity Demand Forecasting Techniques. *Network and Complex Systems, National Conference on Emerging Trends in Electrical, Instrumentation & Communication Engineering*, 3 (3). www.iiste.org. ISSN 2224-610X (Paper) ISSN 2225-0603 (Online)
16. Eugene A. Feinberg and Dora Genethliou (2006). *Load Forecasting. Applied Mathematics for Power System Weather*, Issue: August, Publisher: Springer, Chapter 12, Pg. 269 -285
17. Islam B., Baharudin Z., Raza Q. and Nallagownden P. (2014). A Hybrid Neural Network and Genetic Algorithm Based Model for Short Term Load Forecast. *Research Journal of Applied Sciences, Engineering and Technology* 7(13), Pg. 2667-2673. ISSN: 2040-7459;e-ISSN: 2040-7467.
18. Nahi, K., W. René, S. Maarouf and G. Semaan, (2006). An Efficient Approach for Short Term Load Forecasting Using Artificial Neural Networks. *Int. J. Elect. Power Energ. Syst.*, 28(8), Pg. 525-530.
19. Chengxi Liua, Zakir Hussain Ratherab, Zhe Chena, and Claus Leth Baka (2013). An Overview of Decision Tree Applied To Power Systems. *International Journal of Smart Grid and Clean Energy*, 2 (3), Pg. 413 -418.
20. Gwo-Chung Liao and Ta- Peng Tsao (2006). Application of a Fuzzy Neural Network Combined With Chaos Genetic Algorithm and Simulated Annealing to Short Term Load Forecasting. *Evolutionary Computation IEEE Transaction Vol. 10(3)*, Pg. 330 -340.
21. Park, D. C., El- Sharkawi, M. A., Marks, R. A. II, Atlas, L. E. and Danborg, M. J. (1991). Electric Load Forecasting Using an Artificial Neural Network. *IEEE Transactions of Power Engineering*, Vol. 6, Pg. 442-449.

22. Figueiredo V., Rodrigues F., Vale Z., and Gouveia J. B. (2005). An Electric Energy Consumer Characterization Framework Based On Data Mining Techniques. *IEEE Trans. Power Syst.*, 20 (2), Pg. 596-602.
23. Chien Y. C. and Yuafi H. Y. (1994). Estimation of Line Flows and Bus Voltages Using Decision Trees. *IEEE Trans. Power Syst.*, 9 (3):1569-1574.
24. Ugedo A, Lobato E, Peco J, Rouco L. (2005). Decision Trees Applied to The Management of Voltage Constraints. In The Spanish Market. *IEEE Trans. Power Syst.*, 2005, 20(2):963-972.
25. Box, G.E.P. and Jenkins G.M. (1970). Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco.
26. Chen J.F., Wang W.M. and Huang C.M. (1995). Analysis of an Adaptive Time-Series Autoregressive Moving-Average (ARMA) Model For Short-Term Load Forecasting, *Electric Power Syst. Res.* 34 (3), pg. 187-196.
27. Vemuri S., Hill D. and Balasubramanian R. (1973). Load Forecasting Using Stochastic Models. In Proceeding of the 8th Power Industrial Computing Application Conference, Pg. 31-37.
28. Wang H. and Schulz N.N. (2006). Using AMR Data for Load Estimation for Distribution System Analysis. *Electric Power Syst. Res.* 76 (5), Pg. 336-342.
29. Christianse, W.R. (1971). Short Term Load Forecasting Using General Exponential Smoothing, *IEEE Trans. Power Apparatus Syst.* PAS-90 (1971) 900-911.
30. Park J. H., Park Y. M. and Lee, K.Y. (1991). Composite Modeling for Adaptive Short-Term Load Forecasting, *IEEE Trans. Power Syst.* 6 (1), Pg. 450-457.
31. Douglas, A.P., Breipohl, A.M., Lee, F.N. and Adapa, R. (1998). The Impact of Temperature Forecast Uncertainty on Bayesian Load Forecasting. *IEEE Trans. Power Syst.* 13 (4), Pg. 1507-1513.
32. Azadeh, A., Saberi, M., Ghaderi, S.F., Gitiforouz, A. and Ebrahimipour, V. (2008). Improved Estimation of Electricity Demand Function by Integration of Fuzzy System and Data Mining Approach. *Energy Convers. Manage.* 49 (8), Pg. 2165-2177.
33. Wang, B., Tai, N.-L., Zhai, H.-Q., Ye, J., Zhu, J.-D. and Qi, L.-B. (2008). A new ARMAX Model Based on Evolutionary Algorithm and Particle Swarm Optimization For Short Term Load Forecasting. *Electric Power Syst. Res.* 78 (10), Pg. 1679-1685.
34. Brown, R. G. (1983). Introduction to Random Signal Analysis and Kalman Filtering, John Wiley & Sons, Inc., New York.
35. Gelb, A. (1974). Applied Optimal Estimation. The MIT Press, MA.
36. Trudnowski, D.J., McReynolds, W.L. and Johnson, J.M. (2001). Real-Time Very Short-Term Load Prediction for Power-System Automatic Generation Control, *IEEE Trans. Control Syst. Technol.* 9 (2), Pg. 254-260.
37. Moghram, I. and Rahman, S. (1989). Analysis and Evaluation of Five Short-Term Load Forecasting Techniques. *IEEE Trans. Power Syst.* 4 (4), Pg. 1484-1491.
38. Al-Hamadi, H. M. and Soliman, S. A. (2006). Fuzzy Short-Term Electric Load Forecasting Using Kalman Filter. *IEEE Proc. Conf.* 153 (2), Pg. 217-227.
39. Amjady, N. (2007). Short-term Bus Load Forecasting of Power Systems by a New Hybrid Method. *IEEE Trans. Power Syst.* 22 (1), Pg. 333-341.
40. Asbury, C. (1975). Weather Load Model for Electric Demand Energy Forecasting. *IEEE Trans. Power Apparatus Syst.* PAS-94, Pg. 1111-1116.
41. Papalexopoulos, A.D. and Hesterberg, T.C. (1990). A Regression-Based Approach to Short-Term System Load Forecasting, *IEEE Trans. Power Syst.* 5 (4), Pg. 1535- 1547.
42. Soliman, S.A., Persaud, S., El-Nagar, K. and El-Hawary, M.E. (1997). Application of Least Absolute Value Parameter Estimation Based On Linear Programming to Short Term Load Forecasting. *Int. J. Electrical Power Energy Syst.* 19 (3), Pg. 209-216.
43. Mirasgedis, S., Safaridis, Y., Georgopoulou, E., Lalas, D.P., Moschovits, M., Karagiannis, F. and Papakonstantinou, D. (2006). Models for Mid-Term Electricity Demand Forecasting Incorporating Weather Influences, *Energy* 31 (2-3), Pg. 208-227.
44. Mohamed, Z. and Bodger, P. (2005). Forecasting Electricity Consumption In New Zealand Using Economic And Demographic Variables, *Energy* 30 (10), Pg. 1833- 1843.
45. Tsekouras, G.J., Dialynas, E.N., Hatziaargyriou, N.D. and Kavatzas, S. (2007). A Non-Linear Multivariable Regression Model For Midterm Energy Forecasting of Power Systems, *Electric Power Syst. Res.* 77 (12), Pg. 1560-1568.

46. Rahman, S. and Bhatnagar, R. (1998). An Expert System Based Algorithm for Short-Term Load Forecasting. IEEE Trans. Power Syst. 3 (2), Pg. 392–399.
47. Chiu, C.C., Kao, L.J. and Cook, D.F. (1997). Combining a Neural Network with a Rule-Based Expert System Approach for Short-Term Power Load Forecasting in Taiwan. Expert Syst. Appl. 13 (4), Pg. 299–305.
48. Rahman, S. and Hazim, O. (1993). A Generalized Knowledge-Based Short-Term Load-Forecasting Technique. IEEE Trans. Power Syst. 8 (2) Pg. 508–514.
49. Ying, L.-C. and Pan, M.-C. (2008). Using Adaptive Network Based Fuzzy Inference System to Forecast Regional Electricity Loads. Energy Convers. Manage. 49 (2) Pg. 205–211.
50. Pai, P.-F. (2006). Hybrid Ellipsoidal Fuzzy Systems in Forecasting Regional Electricity Loads. Energy Convers. Manage. 47 (15–16), Pg. 2283–2289.
51. Pandian, S.C., Duraiswamy, K., Rajan, C.C.A. and Kanagaraj, N. (2006). Fuzzy Approach for Short Term Load Forecasting. Electric Power Syst. Res. 76 (6–7), Pg. 541–548.
52. Wyne Iba and Pat Langley (1992). Induction of One –Level Decision Trees. Machine Learning.
53. Sushilkumar Kalmegh (2015). Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News. IJISSET - International Journal of Innovative Science, Engineering & Technology, 2 (2).
54. Jayanth, S K and Sasikala, S. (2013). Reptree Classifier for Identifying Link Spam in Web Search Engines. ICTACT Journal on Soft Computing, 3 (2), Pg. 498- 505.