THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

ELECTRICITY TARIFF USAGE PREDICTION VIA DATA MINING

Master's Thesis

ONUR KARLIDAĞ

ISTANBUL, 2016

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
INFORMATION TECHNOLOGIES

ELECTRICITY TARIFF USAGE PREDICTION VIA DATA MINING

Master's Thesis

ONUR KARLIDAĞ

Supervisor: PROF. DR. ADEM KARAHOCA

İSTANBUL, 2016

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
INFORMATION TECHNOLOGIES

Name of the thesis: Electricity Tariff Usage Prediction via Data Mining
Name/Last Name of the Student: Onur Karlıdağ
Date of the Defense of Thesis:

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assoc. Prof. Dr. Nafiz ARICA
Graduate School Director
Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Sciences.

Prof. Dr. Adem KARAHOCA
Program Coordinator
Signature

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Sciences.

Examining Comittee Members	Signature
Thesis Supervisor Prof Dr. Adem KARAHOCA	----------
Member Prof.Dr. İbrahim Pınar	----------
Member Asst.Prof.Dr. Dilek Karahoca	----------

ACKNOWLEDGMENTS

I would like to express my appreciation and thanks to my thesis supervisor, Prof. Dr. Adem Karahoca, for his support in preparation of this study.

İstanbul,2016 Onur KARLIDAĞ

ABSTRACT

ELECTRICITY TARIFF USAGE PREDICTION VIA DATA MINING

Onur Karlıdağ

Information Technologies

Thesis Supervisor: Prof. Dr. Adem Karahoca

April 2016, 45 Pages

Continuity of electrical energy production is the most important principle for electricity companies. Produced electrical energy is transmitted to residence, business and industry areas. Production and consumption of electrical energy should be balanced. Prediction of daily consumption is very important for both meeting the demand and preventing waste of energy resources.

Energy tariff is an method for balancing demand and supply. Power outages which cause of immediate energy demand size that system can not supply, prevented with energy tariff, because this tariff encourages customers to consume certain time interval.

In this thesis energy-related dataset with tariff information was used. Transitions between tariffs have direct effect on electricty companies. This thesis provides prediction of tariff transition possibility from flat to multiple.

In this study, consumption of Istanbul European side data which contains first three months of 2015 was used. This dataset includes tariff information, invoice amount and consumption basis kWh for each three-time periods which are day, peak and off peak. It has attributes: flat-time bill amount (thkodnolanasis), flat-time bill amount group (thkodnolanasisgroup), consumption (kWh) group of day (t1group), consumption (kWh) group of peak (t2group), consumption (kWh) group of off-peak (t3group), tariff information (tarife), multiple-time bill amount (thkodnolanesnek) which are input columns and multiple-time bill amount group (thkodnolanesnekgroup) is determined output column.

Prediction of the tariff transition possibility calculated with following classification algorithms Logistic Regression, RBF Network, SMO, Naive Bayes, Naive Bayes Net,

Naive Bayes Updatable, J48, NBTree using WEKA. J48, NBTree and Bayes Net classification algorithms have highest accuracy rates which are 100 percent, 100 percent and 99.98 percent.

This thesis, unlike the literature, provides prediction of tariff transition possibility from flat which means one-time to multiple which means three-time.

Keywords: Electricity Energy, Electricity Energy Tariff Data, Prediction Algorithms, Weka, Logistic Regression.

ÖZET

VERİ MADENCİLİĞİ İLE ELEKTRİK TARİFESI KULLANIMI TAHMİNLEMESİ

Onur Karlıdağ

Bilgi Teknolojileri

Tez Danışmanı: Prof. Dr. Adem Karahoca

Nisan 2016, 41 Sayfa

Elektrik enerjisi günümüzde en çok kullanılan enerji türüdür. Elektrik enerjisi üretiminde süreklilik esastır. Üretilen elektrik enerjisi meskenlere, ticarethanelere, sanayilere aktarılır. Elektrik enerjisi üretiminin, günlük gerçekleşen tüketime yakın tahmin edilmesi, tüketim talebi karşılayabilmek ve talep edilenden fazla enerji üretiminin neden olacağı, gereksiz kaynak kullanımının önüne geçilmesi açısından çok önemlidir.

Elektrik enerjisi arz ve talebi dengede tutabilmek için kullanılan yöntemlerden biri enerji tarifeleridir. Bu tarifeler sayesinde, sistemden sunamayacağı büyüklükte ani enerji talebinin neden olacağı elektrik kesintilerinin önüne geçilmiştir çünkü bu tarifler tüketimleri belirli zaman aralıklarında kullanımına teşvik eder.

Bu tez, elektrik enerjisi sektörü için çok önemli bir konu olan tek zamanlı ve üç zamanlı tarifeleri merkez alır. Müşteriler tarafından kullanılan bu tarifeler arasındaki geçişler ya da değişimler elektrik şirketlerini doğrudan etkilemektedir. Bu tez, tek zamanlı tarife kullanan müşterilerin, üç zamanlı tarifeye geçme ihtimallerinin ne olduğunu tahmin eden çalışma sunar.

Bu çalışmada, İstanbul Avrupa yakası 2015 yılı ocak, şubat ve mart ayı verileri kullanılmıştır. Bu veriler, tarife, fatura tutarı ve kWh bazında tüketim bilgilerini içerir. Öznitelikler arasından, tek zamanlı tarife fatura tutarı (thkodnolanasis), tek zamanlı tarife fatura tutarı grubu (thkodnolanasisgroup), gündüz tüketimi (t1group), puant tüketimi (t2group), gece tüketimi (t3group), tarife bilgisi(tarife), üç zamanlı tarife fatura tutarı (thkodnolanesnek) alanları giriş verileri olarak kullanılmıştır. Çıkış verisi olarak ise üç zamanlı tarife fatura tutarı grubu (thkodnolanesnekgroup) alanı kullanılmıştır.

Müşterilerin Üç zamanlı tarifeden tek zamanlı tarifeye geçme ihtimalini tahmin edebilmek için WEKA aracı ile Logistic Regression, RBF Network, SMO, Naive Bayes, Naive Bayes Net, Naive Bayes Updatable, J48, NBTree algoritmaları kullanılmıştır. J48, NBTree ve Bayes Net sınıflandırma algoritmaları en yüksek doğruluk oranlarına sahiptir, bunlar sırası ile yüzde 100, yüzde 100 ve yüzde 99.98.

Bu tezin, literatürdeki çalışmalardan farkı, müşterilerin, enerji tarifeleri arasındaki geçiş ihtimalini tahmin etmesidir.

Anahtar Kelimeler: Elektrik Enerjisi, Elektrik Enerjisi Tarife Verisi, Tahmin Algoritmaları, Weka, Logistic Regression.

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

Bayes Net : Bayesian Network

FP Rate : False Positive Rate

MAE : Mean Absolute Error

NBTree : Function that combines Naive Bayes with a Decision Tree classifier.

RBF Network : Radial Basis Function Network

RMSE : Root Mean Absolute Error

ROC : Receiver Operating Characteristic

SMO : Sequential Minimal Optimization

SVM : Support Vector Machines

TP Rate : True Positive Rate

1. INTRODUCTION

Electricity market legislation is determined by EMRA which stands for Energy Market Regulatory Authority in Turkey. Electricity Market Law ("Electricity Market Law", 2001) established by EMRA to ensure stable, financially strong and transparent electricity market.

Electricity Company which is subject of the thesis consists of distribution and retail sale companies. Distribution companies carry electricity from electric energy generation plants to consumption points, retail sale companies sell electricity to consumers. Distribution companies are responsible for quality of service and retail sale companies are responsible for commercial quality ("Regulation on Service Quality in Electricity Distribution and Retail Sale", 2012).

Electricity Company serves in 3.573 km² area and has 4.3 million customers. The company makes 25 billion kWh electricity distributions. These features make the company the biggest electricity distribution company in Turkey. The company is adopting qualified and continuous serve understanding. With 13 percent market share, the company is flagship of the electricity sector ("The Energy Sector: A Quick Tour for the Investor", 2013).

The company has two tariffs for their customers, in the one-time tariff consumption of customers billed with constant price, in the three-time tariff, consumption billed by three periods which are day, peak and off peak. Customers can determine their tariff by considering their consumption habits.

In this thesis energy-related dataset with tariff information was used. Transitions, between tariffs directly effects on electricity companies. This thesis provides prediction of tariff transition possibility from flat to multiple.

In this thesis, consumption of Istanbul European side data that first three months of 2015 were used. Dataset contains tariff information, invoice amount and consumption

basis kWh for each three-time periods which are day, peak and off peak. The data set contains following attributes flat-time bill amount (thkodnolanasis), flat-time bill amount group (thkodnolanasisgroup), consumption (kWh) of day (t1group), consumption (kWh) of peak (t2group), consumption (kWh) of off-peak (t3group), tariff information (tarife), multiple-time bill amount (thkodnolanesnek) which are input columns and multiple-time bill amount group (thkodnolanesnekgroup) is determined output column.

Prediction of the tariff transition possibility calculated with following classification algorithms Logistic Regression, RBF Network, SMO, Naive Bayes, Naive Bayes Net, Naive Bayes Updatable, J48, NBTree using with WEKA (Witten and Frank, 2005).

In Section 2, concise literature researches which is related to prediction algorithms in electricity field provided. In Section 3, firstly features of the electric energy data set explained secondly discretization of the attributes shown. Thirdly, prediction algorithms which are used for prediction of tariff information defined. In Section 4, outcome of the algorithms which are used for prediction of tariff information evaluated. In Section 5, outcome of the algorithms compared according to accuracy rate and listed from highest to lowest accuracy rate.

2. LITERATURE REVIEW

In this section, brief literature review provided.

2.1 ELECTRICITY TARIFF USAGE PREDICTION MODELS

Energy-related data is examined under various topics in literature. Topics can be mainly grouped as follows: relationship between customers' expectation and their preferences, consumption behavior and tariff design, consumption behavior based on customer segmentation, effect of the price on consumption. By these categories, some of the related studies can be given.

To determine energy efficiency solutions, firstly customers' consumption behaviors should be examined. To find out correlation between customers' characteristics and consumption, customer information segmented according to: social class, contracted power, number of rooms, family size and type of tariff. There is strong correlation between tariff and consumption (Pombeiro, Pina and Silva, 2012).

Electricity suppliers need to design proper tariffs which respond to consumers' energy usage characteristics to gain new market share and to create more competition within their markets (Stephenson, Lungu, Paun, Silvas and Tupu, 2001).

Proper tariff design requires knowledge of consumers' energy usage characteristics. Load diagram is becoming a key for proper tariff design (Chicco, Napoli, Piglione, Postolache, Scutariu and Toader, 2002).

Electricity tariff prices are important factor that drive the consumers' consumption behaviors. Consumers avoid consuming electricity during period which has peak price (Kirschen,Strbac,Cumperayot and Mendes,2000).

Electricity customers are aware of tariffs and price of them and this knowledge affects their tariff preferences (Slavickas,Alden and El-Kady,1999).

3. DATA AND METHODS

In this section, firstly, features of the data set detailed. Secondly, discretization approach explained. Finally, classification algorithms which are used for prediction described.

3.1 DATA SET

The energy-related data set contains tariff information, invoice amount and consumption basis kWh for each three-time periods which are day, peak and off peak. Energy-related data set consist of 8 attributes and 100000 instances.

The attributes which are equivalent to columns as follows;

i. Thkodnolanasis: flat-time bill amount
ii. Thkodnolanasisgroup: flat-time bill amount group
iii. t1group: consumption (kWh) group of day
iv. t2group: consumption (kWh) group of peak
v. t3group: consumption (kWh) group of off-peak
vi. tarife: tariff information
vii. thkodnolanesnek: multiple-time bill amount
viii. thkodnolanesnekgroup: multiple-time bill amount group

i. **Thkodnolanasis**: This attribute contains flat-time bill amount which is calculated based on tariff and unit price, data type of it is numeric there are 1173 distinct thkodnolanasis value. Min value of this attributes is 0 and maximum value of it is 119110.
ii. **Thkodnolanasisgroup:** This attribute contains flat-time bill amount groups there are 5 distinct thkodnolanasisgroup values which means 5 different groups. The group of number begins with 1 and end with 5. Data type of this attribute is numeric.
iii. **T1group:** This attribute contains group of consumption within day period of three-time tariff, there are 5 distinct t1group values which mean 5 different

groups. t1group number begins with 1 and end with 5. Data type of this attribute is numeric

iv. **T2group:** This attribute contains group of consumption within peak period of three-time tariff, there are 5 distinct t2group values which mean 5 different groups. t2group number begins with 1 and end with 5. Data type of this attribute is numeric.

v. **T3group:** This attribute contains group of consumption within off peak period of three-time tariff, there are 5 distinct t3group values which mean 5 different groups. t3group number begins with 1 and end with 5. Data type of this attribute is numeric

vi. **Tarife:** This attribute contains the tariff information of customers. There are 2 distinct values which are one –time and three-time tariff. Data type of this attribute is nominal.

vii. **Thkodnolanesnek**: This attribute contains multiple-time bill amount which is calculated based on three-time tariff and unit price, data type of this numeric, there are 1140 distinct thkodnolanesnek value. Min value of this attributes is 0 and maximum value of it is 119110.

viii. **Thkodnolanesnekgroup:** This attribute contains multiple-time bill amount groups, there are 5 distinct thkodnolanasisgroup values which means 5 different groups consist of a, b, c, d, f. Data type of this attribute is nominal.

3.1.1 Discretization

The dataset composed of many different values, in order to find out correlation between the attributes, values of them must be clustered. The following attributes were clustered flat-time bill amount group (Thkodnolanasisgroup), consumption (kWh) group of day (T1group), consumption (kWh) group of peak (T2group), consumption (kWh) group of off-peak (t3group), flat-time bill amount group (thkodnolanesnekgroup).

Number of cluster was determined as five for these attributes, in other words, these attributes was divided into five clusters by value range. Member count of the clusters for an attribute should be close to each other for efficient clustering, while dividing attributes into clusters this principle applied.

Figure 3.1: Discretization result for t1group

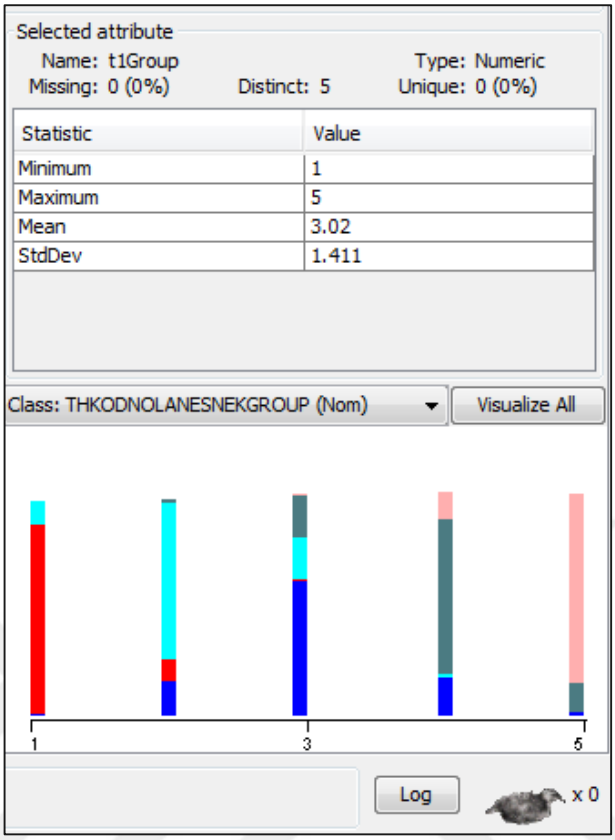

Figure 3.1 illustrates distribution of t1group to 5 groups which have very close count of member each other. Count of members for each group listed in Table 3.1 which also illustrated with column in this figure from left to right.

Table 3.1: Count of Instance Distribution for t1group Discretization

Discretization of t1group	1	2	3	4	5
Counts	19518	19787	20071	20424	20200

Table 3.2: Range of Discretization for t1group

t1group		
Group Name	Beginning Value	End Value
1	0	45
2	46	71
3	72	96
4	97	137
5	138	168577

Table 3.2 shows groups' range for t1group attribute.

Figure 3.2: Discretization result for t2group

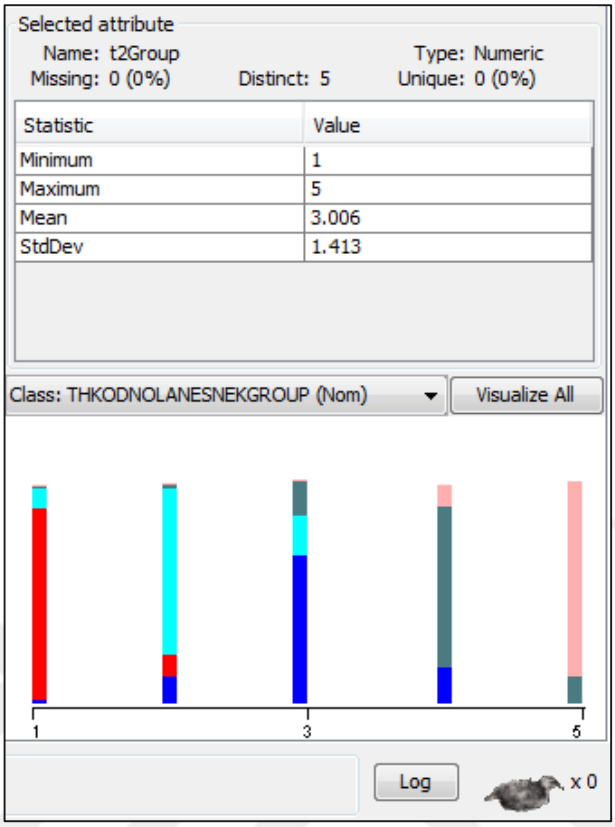

Figure 3.2 illustrates distribution of t2group to 5 groups which have very close count of member each other. Count of members for each group listed in Table 3.3 which also illustrated with column in this figure from left to right.

Table 3.3: Count of Instance Distribution for t2group Discretization

Discretization of t2group	1	2	3	4	5
Counts	19867	19864	20259	19861	20149

Table 3.4: Discretization result for t2group

t2group		
Group Name	Beginning Value	End Value
1	0	27
2	28	44
3	45	60
4	61	85
5	86	90033

Table 3.4 shows groups' range for t2group attribute.

Figure 3.3: Discretization result for t3group

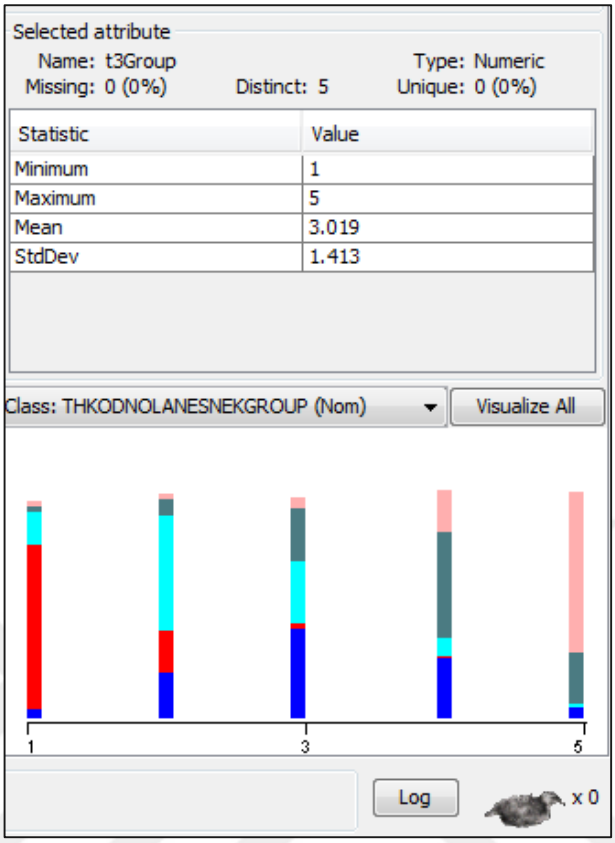

Figure 3.3 illustrates distribution of t3group to 5 groups which have very close count of member each other. Count of members for each group listed in Table 3.5 which also illustrated with column in this figure from left to right.

Table 3.5: Count of Instance Distribution for t3group Discretization

Discretization of t3group	1	2	3	4	5
Counts	19462	20188	19662	20331	20357

Table 3.6: Discretization result for t3group

t3 group		
Group Name	Beginning Value	End Value
1	0	28
2	29	47
3	48	64
4	65	92
5	93	84245

Table 3.6 shows groups' range for t3group attribute.

Figure 3.4: Discretization result for thkodnolanasisgroup

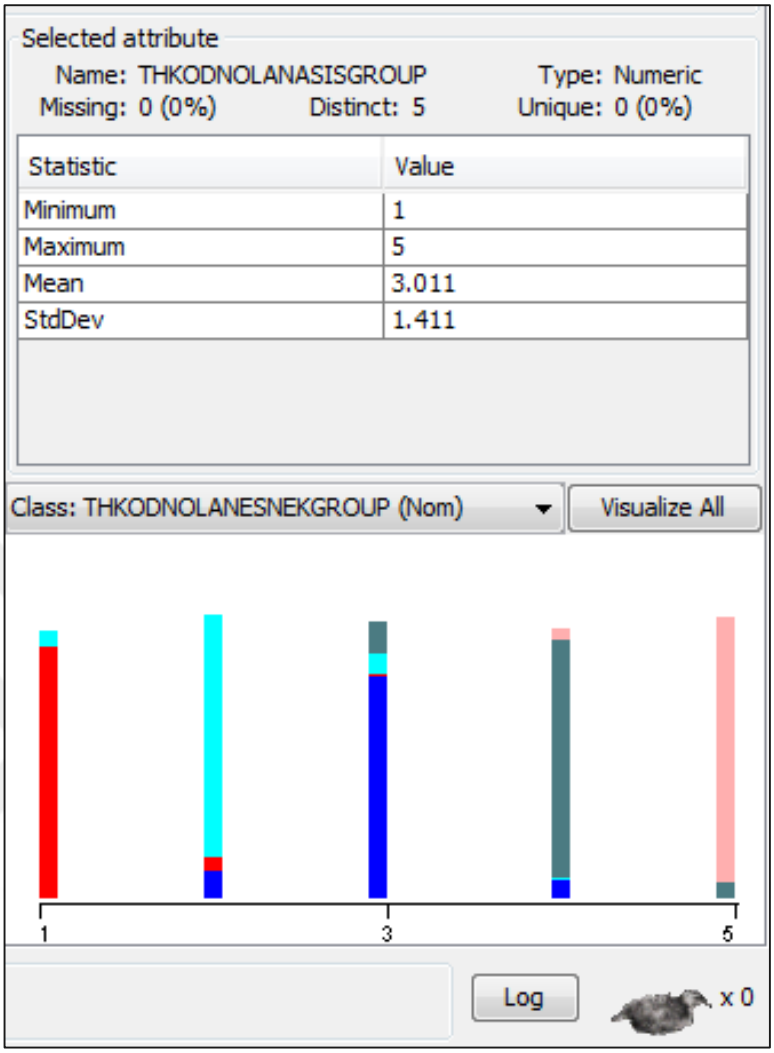

Figure 3.4 illustrates distribution of thkodnolanasisgroup to 5 groups which have very close count of member each other. Count of members for each group listed in Table 3.7 which also illustrated with column in this figure from left to right.

Table 3.7: Count of Instance Distribution for thkodnolanasisgroup Discretization

Discretization of thkodnolanasisgroup	1	2	3	4	5
Counts	19387	20540	20056	19657	20360

Table 3.8: Discretization result for thkodnolanasisgroup

thkodnolanasisgroup		
Group Name	Beginning Value	End Value
1	0	39
2	40	60
3	61	79
4	80	109
5	110	119110

Table 3.8 shows groups' range for thkodnolanasisgroup attribute.

Figure 3.5: Discretization result for thkodnolanesnekgroup

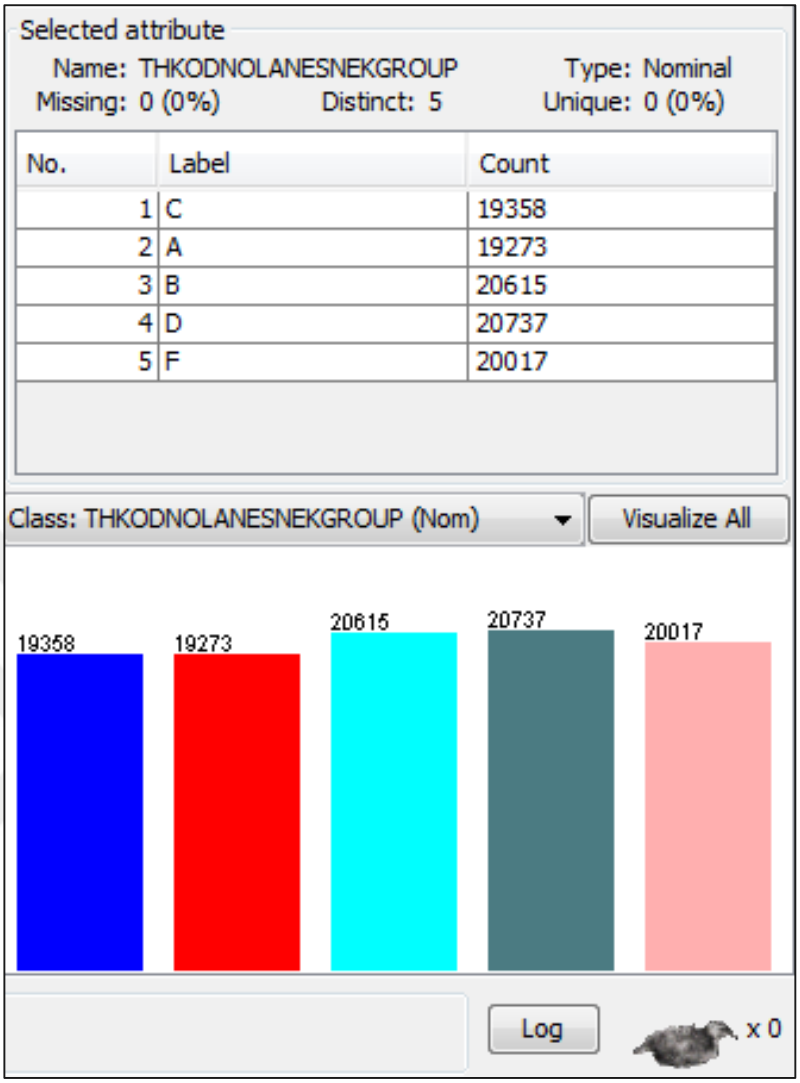

Figure 3.5 illustrates distribution of thkodnolanesnekgroup to 5 groups which have very close count of member each other. Count of members for each group listed in Table 3.9 which also illustrated with column in this figure from left to right.

Table 3.9: Count of Instance Distribution for thkodnolanesnekgroup Discretization

Discretization of thkodnolanasisgroup	c	a	b	d	f
Counts	19358	19273	20615	20737	20017

Table 3.10: Discretization result for thkodnolanesnekgroup

thkodnolanesnekgroup		
Group Name	Beginning Value	End Value
A	1	37
B	38	58
C	59	76
D	77	107
F	108	119110

Table 3.10 shows groups' range for thkodnolanesnekgroup attribute.

3.2 METHODS

Machine learning is used to reach information among huge volume data. Prediction of future is provided with machine learning system. Machine learning converts data into information using with algorithms and make prediction with them. Machine learning focused on data analysis mostly. WEKA is one of the used machine learning packet (Witten and Frank, 2005). In this subsection, brief information is provided for the following algorithms Logistic Regression, RBF Network, SMO, Naive Bayes, Naive Bayes Net, Naive Bayes Updateable, J48, ID3, NBTree.

3.2.1 Logistic Regression

Logistic regression was developed by David Cox in 1958 (Cox, 1958). The binary logistic model is used to estimate the probability of a binary response based on one or more predictor variables. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function.

3.2.2 RBF Network

Radial basis function network is an artificial neural network that uses radial basis functions (Broomhead and Lowe, 1988). Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control. Radial basis function (RBF) networks typically have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer.

3.2.3 SMO

Sequential minimal optimization was developed by John Platt in 1998 (Platt, 1998). SMO is widely used for training support vector machines. SMO is an iterative algorithm for solving optimization problem. SMO breaks problem into a series of smallest possible sub-problems, which are then solved analytically.

3.2.4 Naive Bayes

In machine learning, Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem (Russell and Norvig, 2003). Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.

3.2.5 Naive Bayes Net

Bayesian network is member of probabilistic graphical models (Bouckaert,1995). Graphical model structures are used to represent knowledge about an uncertain domain.

3.2.6 Naive Bayes Updateable

This is the updateable version of Naive Bayes (Russell and Norvig, 2003).

3.2.7 J48

Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances (Quinlan, 2014). J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc.

3.2.8 ID3

ID3 (Iterative Dichotomiser 3) decision tree algorithm is developed by Quinlan in 1986. In the decision tree method, information gain approach is generally used to determine suitable property for each node of a generated decision tree.

3.2.9 NBTree

The NBTree algorithm is a hybrid between decision-tree classifiers and Naive Bayes classifiers (Kohavi, 1996). It represents the learned knowledge in the form of a tree which is constructed recursively.

4. FINDINGS

In this section, results of algorithms that are used for prediction of tariff transition possibility from flat which means one-time to multiple which means three-time are provided. Success of algorithms which were applied to energy-related data set that contains 3 months data evaluated over outcomes of the algorithms.

The energy-related data set contains tariff information, invoice amount and consumption basis kWh for each three-time periods which are day, peak and off peak. Energy-related data set consist of 8 attributes and 100000 instances.

Prediction algorithms applied to the data set in WEKA platform. WEKA is most popular machine learning software which is free written in Java, developed at Waikato University, New Zealand (Witten and Frank, 2005). WEKA involves algorithms for data classification. Results of the algorithms are given as summary of outcome, detailed accuracy by class and confusion matrix.

In this study, two different data sets were prepared. First data set called training and it has 100000 instances and it contains 50000 rows that belong to flat tariff consumptions and the rest of 50000 rows that belong to multiple tariff consumptions. Second data set called test and it has 50000 rows this data set only contains flat time tariff data which was used for prediction. WEKA builds a model over training data set, then this model was saved to use later. The saved model is loaded to WEKA then test data loaded, finally the model is re-evaluated on current test set. After this process WEKA returns outcomes that analyzed in this section.

4.1 BAYES NET

Bayes Net classification algorithm applied to the data set using with WEKA platform and summary of outcome tabulated in Table 4.1.

Table 4.1: Summary of Bayes Net Outcome

Correctly Classified Instances	99979	99.98%
Incorrectly Classified Instances	21	0.02%
Kappa statistic	0.9997	
Mean absolute error	0.0003	
Root mean squared error	0.0091	
Relative absolute error	0.10%	
Root relative squared error	2.27%	
Total Number of Instances	100000	

99979 instances among 100000 instances are correctly classified that equivalent to 99.98 percent of the total instances.

MAE value is 0.0003.

RMSE value is 0.0091.

The value for Kappa is 0.9997 indicating almost complete level of agreement.

Kappa value is calculation which based on agreement of predicted class with actual class. Value of kappa statistic varies from 0 to 1. If the value is 0 it means there is no relation between class label and attributes, relation increase while the value approaches to 1.

MAE stands for Mean Absolute Error that means average size of errors as the name suggests. Value of the absolute error is difference between predicted value and actual value.

RMSE stands for Root Mean Squared Error. This calculation based on mean error like MAE. RMSE calculation is average of the square of error.

Value of MAE and RMSE varies from 0 to ∞. Difference between MAE and RMSE indicates variation in the errors.

Table 4.2: Bayes Net Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	C (59-76)
	1	0	1	1	1	1	A (1-37)
	1	0	1	1	1	1	B (38-58)
	1	0	1	1	1	1	D (77-107)
	1	0	0.999	1	1	1	F (108-119110)
Weighted Avg.	1	0	1	1	1	1	

In Table 4.2

TP Rate is 1 for all classes.

FP Rate is 0 for all classes.

Precision value decrease while bill amount increase for this classifier method.

ROC Area is 1 for all classes.

True Positive Rate is the rate of correctly classified instances in a class.

False Positive Rate is the rate of incorrectly classified instances in a class.

Precision value is proportion; calculation is correctly classified instances for that class divided by total number of instances classified as that class.

Formulation is: True Positive / (True Positive + False Positive)

Recall value is proportion: calculation is correctly classified instances for that class divided by total number of instances in that class.

Formulation is: True Positive / (True Positive + False Negative)

F-Measure is harmonic mean of recall and precision. Formulation is: 2 * Precision * Recall / (Precision + Recall).

ROC stands for Receiver Operating Characteristic and used for performance of the classifiers.

Success of classifier method increases while the ROC area approaches to 1 and decrease while the ROC area approaches 0.5.

Table 4.3: Bayes Net Confusion Matrix

A	b	c	d	e	classified as
19354	0	0	0	4	a = C
0	19273	0	0	0	b = A
0	0	20607	5	3	c = B
0	0	2	20731	4	d = D
0	0	0	3	20014	e = F

Confusion matrix shows counts of correctly and incorrectly classified instance for each class.

4.2 Naive Bayes

Naive Bayes classification algorithm applied to the data set using with WEKA platform and summary of outcome tabulated in Table 4.4.

Table 4.4: Summary of Naive Bayes Outcome

Correctly Classified Instances	81287	81.29%
Incorrectly Classified Instances	18713	18.71%
Kappa statistic	0.7659	
Mean absolute error	0.0799	
Root mean squared error	0.244	
Relative absolute error	24.96%	
Root relative squared error	61.01%	

81287 instances among 100000 instances are correctly classified that equivalent 81.29 percent of the total instances.

MAE value is 0.0799.

RMSE value is 0.244.

The value for Kappa is 0.7659 indicating good level of agreement.

Table 4.5: Naive Bayes Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.914	0.058	0.791	0.914	0.848	0.98	C(59-76)
	0.942	0.012	0.949	0.942	0.945	0.996	A(1-37)
	0.839	0.023	0.906	0.839	0.871	0.985	B(38-58)
	0.886	0.142	0.621	0.886	0.73	0.946	D(77-107)
	0.488	0	0.997	0.488	0.656	0.995	F(108-119110)
Weighted Avg.	0.813	0.048	0.851	0.813	0.809	0.98	

In Table 4.5

TP rate reach maximum value which is 0.92 on class A. TP rate is increase while bill amount decrease for this classifier method.

FP rate reach maximum value which is 0.14 on class D.

Precision reach maximum value which is 0.94 on class A. Precision value increase while bill amount decrease for this classifier method.

ROC Area reach maximum value which is 0.996 on class A. ROC Area rate is increase while bill amount decrease, Average value of ROC Area is 0.98 so this classifier method work successfully for the data set.

Table 4.6: Naive Bayes Confusion Matrix

a	b	c	d	e	classified as
17701	0	658	993	6	a = C
2	18149	1122	0	0	b = A
2333	975	17294	10	3	c = B
2338	0	11	18365	23	d = D
16	0	0	10223	9778	e = F

4.3 Naive Bayes Updateable

Naive Bayes Updateable classification algorithm applied to the data set using with WEKA platform and summary of outcome tabulated in Table 4.7.

Table 4.7: Summary of Naive Bayes Updateable Outcome

Correctly Classified Instances	81287	81.29%
Incorrectly Classified Instances	18713	18.71%
Kappa statistic	0.7659	
Mean absolute error	0.0799	
Root mean squared error	0.244	
Relative absolute error	24.96%	
Root relative squared error	61.01%	

81287 instances among 100000 instances are correctly classified that equivalent 81.29 percent of the total instances.

MAE value is 0.0799.

RMSE value is 0.244.

The value for Kappa is 0.7659 indicating good level of agreement.

Table 4.8: Naive Bayes Updateable Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.914	0.058	0.791	0.914	0.848	0.98	C(59-76)
	0.942	0.012	0.949	0.942	0.945	0.996	A(1-37)
	0.839	0.023	0.906	0.839	0.871	0.985	B(38-58)
	0.886	0.142	0.621	0.886	0.73	0.946	D(77-107)
	0.488	0	0.997	0.488	0.656	0.995	F(108-119110)
Weighted Avg.	0.813	0.048	0.851	0.813	0.809	0.98	

In Table 4.8

TP rate reach maximum value which is 0.94 on class A. TP rate is increase while bill amount decrease for this classifier method.

FP rate reaches maximum value which is 0.14 on class D.

Precision reach maximum value which is 0.94 on class A. Precision value increase while bill amount decrease for this classifier method.

ROC Area reach maximum value which is 0.996 on class A. ROC Area rate is increase while bill amount decrease, Average value of ROC Area is 0.98 so this classifier method work successfully for the data set.

Table 4.9: Naive Bayes Updateable Confusion Matrix

a	b	c	d	e	classified as
17701	0	658	993	6	a = C
2	18149	1122	0	0	b = A
2333	975	17294	10	3	c = B
2338	0	11	18365	23	d = D
16	0	0	10223	9778	e = F

4.4 Logistic

Logistic classification algorithm applied to the data set using with WEKA platform and summary of outcome tabulated in Table 4.10.

Table 4.10: Summary of Logistic Outcome

Correctly Classified Instances	91544	91.54%
Incorrectly Classified Instances	8456	8.46%
Kappa statistic	0.8943	
Mean absolute error	0.0547	
Root mean squared error	0.1612	
Relative absolute error	17.11%	
Root relative squared error	40.30%	

91544 instances among 100000 instances are correctly classified that equivalent 91.54 percent of the total instances.

MAE value is 0.0547.

RMSE value is 0.1612.

The value for Kappa is 0.8943 indicating almost complete level of agreement.

Table 4.11: Logistic Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.867	0.045	0.824	0.867	0.845	0.976	C(59-76)
	0.956	0.008	0.966	0.956	0.961	0.999	A(1-37)
	0.91	0.031	0.885	0.91	0.897	0.99	B(38-58)
	0.875	0.02	0.921	0.875	0.898	0.986	D(77-107)
	0.971	0.003	0.989	0.971	0.98	0.999	F(108-119110)
Weighted Avg.	0.915	0.021	0.917	0.915	0.916	0.99	

In Table 4.11

TP rate reach maximum value which is 0.97 on class F. TP rate is increase while bill amount increase for this classifier method.

FP rate reach maximum value which is 0.14 on class C.

Precision reach maximum value which is 0.98 on class A. Precision value increase while bill amount increase for this classifier method.

ROC Area reach maximum value which is 0.999 on both class A and F. Average value of ROC Area is 0.99 so this classifier method work successfully for the data set.

Table 4.12: Logistic Confusion Matrix

a	b	c	d	e	classified as
16786	2	1563	1007	0	a = C
1	18418	854	0	0	b = A
1206	649	18752	8	0	c = B
2361	0	9	18149	218	d = D
28	5	13	532	19439	e = F

4.5 RBF Network

RBF Network classification algorithm applied to the data set using with WEKA platform and summary of outcome tabulated in Table 4.13.

Table 4.13: Summary of RBF Network Outcome

Correctly Classified Instances	89989	89.99%
Incorrectly Classified Instances	10011	10.01%
Kappa statistic	0.8749	
Mean absolute error	0.0595	
Root mean squared error	0.1734	
Relative absolute error	18.61%	
Root relative squared error	43.35%	

89989 instances among 100000 instances are correctly classified that equivalent 89.9 percent of the total instances.

MAE value is 0.0595.

RMSE value is 0.1734.

The value for Kappa is 0.8749 indicating almost complete level of agreement.

Table 4.14: RBF Network Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.875	0.047	0.816	0.875	0.844	0.982	C(59-76)
	0.954	0.013	0.945	0.954	0.949	0.997	A(1-37)
	0.888	0.03	0.886	0.888	0.887	0.988	B(38-58)
	0.849	0.026	0.894	0.849	0.871	0.987	D(77-107)
	0.937	0.008	0.966	0.937	0.951	0.996	F(108-119110)
Weighted Avg.	0.9	0.025	0.901	0.9	0.9	0.99	

In Table 4.14

TP rate reach maximum value which is 0.95 on class A. TP rate is increase while bill amount decrease for this classifier method.

FP rate reach maximum value which is 0.04 on class C.

Precision reach maximum value which is 0.96 on class F. Precision value increase while bill amount increase for this classifier method.

ROC Area reach maximum value which is 0.997 on both class A and F. Average value of ROC Area is 0.99 so this classifier method work successfully for the data set.

Table 4.15: RBF Network Confusion Matrix

a	b	c	d	e	classified as
16938	0	1475	942	3	a = C
16	18385	872	0	0	b = A
1224	1080	18308	2	1	c = B
2469	0	1	17609	658	d = D
117	0	0	1151	18749	e = F

4.6 SMO

SMO classification algorithm applied to the data set using with WEKA platform and summary of outcome tabulated in Table 4.16.

Table 4.16: Summary of SMO Outcome

Correctly Classified Instances	91046	91.05%
Incorrectly Classified Instances	8954	8.95%
Kappa statistic	0.8881	
Mean absolute error	0.2437	
Root mean squared error	0.322	
Relative absolute error	76.16%	
Root relative squared error	80.50%	

91046 instances among 100000 instances are correctly classified that equivalent 91.05 percent of the total instances.

MAE value is 0.2437.

RMSE value is 0.322.

The value for Kappa is 0.8881 indicating almost complete level of agreement.

Table 4.17: SMO Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.911	0.046	0.827	0.911	0.867	0.966	C(59-76)
	0.951	0.009	0.964	0.951	0.957	0.993	A(1-37)
	0.897	0.026	0.899	0.897	0.898	0.969	B(38-58)
	0.839	0.017	0.928	0.839	0.881	0.959	D(77-107)
	0.958	0.014	0.945	0.958	0.951	0.989	F(108-119110)
Weighted Avg.	0.91	0.022	0.913	0.91	0.911	0.975	

In Table 4.17

TP rate reach maximum value which is 0.95 on class A. TP rate is increase while bill amount decrease for this classifier method.

FP rate reach maximum value which is 0.04 on class C.

Precision reach maximum value which is 0.96 on class F. Precision value increase while bill amount increase for this classifier method.

ROC Area reach maximum value which is 0.993 on both class A and F. Average value of ROC Area is 0,97 so this classifier method work successfully for the data set.

Table 4.18: SMO Confusion Matrix

a	b	c	d	e	classified as
17642	0	1154	526	36	a = C
6	18334	933	0	0	b = A
1429	691	18495	0	0	c = B
2251	0	1	17408	1077	d = D
17	0	0	833	19167	e = F

4.7 NB Tree

NBTree classification algorithm applied to the data set using with WEKA platform and summary of outcome tabulated in Table 4.19.

Table 4.19: Summary of NBTree Outcome

Correctly Classified Instances	99998	100.00%
Incorrectly Classified Instances	2	0.00%
Kappa statistic	1	
Mean absolute error	0.0015	
Root mean squared error	0.0103	
Relative absolute error	0.46%	
Root relative squared error	2.58%	

99998 instances among 100000 instances are correctly classified that equivalent 100 percent of the total instances.

MAE value is 0.0015.

RMSE value is 0.0103.

The value for Kappa is 1 indicating almost complete level of agreement.

Table 4.20: NBTree Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	C(59-76)
	1	0	1	1	1	1	A(1-37)
	1	0	1	1	1	1	B(38-58)
	1	0	1	1	1	1	D(77-107)
	1	0	1	1	1	1	F(108-119110)
Weighted Avg.	1	0	1	1	1	1	

In Table 4.20

TP rate is 1 for all classes.

FP rate is 0 for all classes.

Precision value is 1 for all classes.

ROC Area is 1 for all classes, so this classifier method work successfully for the data set.

Table 4.21: NBTree Confusion Matrix

a	b	c	d	e	classified as
19357	0	0	1	0	a = C
0	19272	1	0	0	b = A
0	0	20615	0	0	c = B
0	0	0	20737	0	d = D
0	0	0	0	20017	e = F

4.8 J48

J48 classification algorithm applied to the data set using with WEKA platform and summary of outcome tabulated in Table 4.22.

Table 4.22: Summary of J48 Outcome

Correctly Classified Instances	100000	100%
Incorrectly Classified Instances	0	0%
Kappa statistic	1	
Mean absolute error	0	
Root mean squared error	0	
Relative absolute error	0%	
Root relative squared error	0%	
Total Number of Instances	100000	

100000 instances among 100000 instances are correctly classified that equivalent 100 percent of the total instances.

MAE value is 0.

RMSE value is 0.

The value for Kappa is 1 indicating almost complete level of agreement.

Table 4.23: J48 Detailed Accuracy by Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	C(59-76)
	1	0	1	1	1	1	A(1-37)
	1	0	1	1	1	1	B(38-58)
	1	0	1	1	1	1	D(77-107)
	1	0	1	1	1	1	F(108-119110)
Weighted Avg.	1	0	1	1	1	1	

In Table 4.23

TP rate is 1 for all classes.

FP rate is 0 for all classes.

Precision value is 1 for all classes.

ROC Area is 1 for all classes, so this classifier method work successfully for the data set.

Table 4.24: J48 Confusion Matrix

a	b	c	d	e	classified as
19358	0	0	0	0	a = C
0	19273	0	0	0	b = A
0	0	20615	0	0	c = B
0	0	0	20737	0	d = D
0	0	0	0	20017	e = F

5. DISCUSSION

In this section, outcomes of the algorithms that stated in section 4 discussed for proper classification.

Table 5.1: Classifying Outcomes of Algorithms

	Bayes Net	Naive Bayes	Naive Bayes Updatable	Logistic	RBF Network	SMO	NBTree	J48
Correctly Classified Instances	99979	81287	81287	91544	89989	91046	99998	100000
Incorrectly Classified Instances	21	18713	18713	8456	10011	8954	2	0
Kappa statistic	0.9997	0.7659	0.7659	0.8943	0.8749	0.8881	1	1
Mean absolute error	0.0003	0.0799	0.0799	0.0547	0.0595	0.2437	0.0015	0
Root mean squared error	0.0091	0.244	0.244	0.1612	0.1734	0.322	0.0103	0
Relative absolute error	0.10%	24.96%	24.96%	17.11%	18.61%	76.16%	0.46%	0%
Root relative squared error	2.27%	61.01%	61.01%	40.30%	43.35%	80.50%	2.58%	0%
Total Number of Instances	100000	100000	100000	100000	100000	100000	100000	100000

J48 is best algorithm to classify the given instances. Correctly classified instances of J48 are 100 percent. According to these results, the rest of the classification algorithms are considered successful.

Kappa value is calculation which based on agreement of predicted class with actual class. Value of Kappa statistic varies from 0 to 1. If the value is 0 it means there is no relation between class label and attributes, relation increase while the value approaches to 1. NBTree and J48 have the highest value.

Table 5.2: Classifying Percentage Outcomes of Algorithms

Rates	Bayes Net	Naive Bayes	Naive Bayes Updatable	Logistic	RBF Network	SMO	NBTree	J48
Correctly Classified Instances	99.98%	81.29%	81.29%	91.54%	89.99%	91.05%	100.00%	100%
Incorrectly Classified Instances	0.02%	18.71%	18.71%	8.46%	10.01%	8.95%	0.00%	0%

Correctly classified instances are directly related to accuracy of classification algorithms.

Table 5.3: Weighted Averages of Detailed Accuracy Outcomes

Weighted Avg.	Bayes Net	Naive Bayes	Naive Bayes Updatable	Logistic	RBF Network	SMO	NBTree	J48
TP Rate	1	0.813	0.813	0.915	0.9	0.91	1	1
FP Rate	0	0.048	0.048	0.021	0.025	0.022	0	0
Precision	1	0.851	0.851	0.917	0.901	0.913	1	1
Recall	1	0.813	0.813	0.915	0.9	0.91	1	1
F-Measure	1	0.809	0.809	0.916	0.9	0.911	1	1
ROC Area	1	0.98	0.98	0.99	0.99	0.975	1	1

Weighted averages of detailed outcomes are tabulated. NBTree, J48 and BayesNet have the best TP Rate value.

All the ROC values are bigger than 0.975.

Table 5.4: Count of Predicted Instance Distribution by Logistic Algorithm

Groups	Counts
1:C	8918
2:A	13457
3:B	11023
4:D	8020
5:F	8582
Total:	50000

According to the Logistic tariff prediction model; 50000 one-time tariff consumer that are the entire test data can be billed from three-time tariff with the tabulated three-time consumption characteristics.

Table 5.5: Count of Predicted Instance Distribution by RBF Network Algorithm

Groups	Counts
1:C	8917
2:A	13467
3:B	10856
4:D	8199
5:F	8561
Total:	50000

According to the RBF Network tariff prediction model; 50000 one-time tariff consumer that are the entire test data can be billed from three-times tariff with the tabulated three-time consumption characteristics.

Table 5.6: Count of Predicted Instance Distribution by SMO Algorithm

Groups	Counts
1:C	9572
2:A	13311
3:B	10480
4:D	7359
5:F	9278
Total:	50000

According to the SMO tariff prediction model; 50000 one-time tariff consumer that are the entire test data can be billed from three-time tariff with the tabulated three-time consumption characteristics.

Table 5.7: Count of Predicted Instance Distribution by Naive Bayes Algorithm

Groups	Counts
1:C	9346
2:A	13454
3:B	10222
4:D	12411
5:F	4567
Total:	50000

According to the Naive Bayes tariff prediction model; 50000 one-time tariff consumer that are the entire test data can be billed from three-time tariff with the tabulated three-time consumption characteristics.

Table 5.8: Count of Predicted Instance Distribution by Bayes Net Algorithm

Groups	Counts
1:C	8470
2:A	13898
3:B	10858
4:D	8242
5:F	8532
Total:	50000

According to the Bayes Net tariff prediction model; 50000 one-time tariff consumer that are the entire test data can be billed from three-time tariff with the tabulated three-time consumption characteristics.

Table 5.9: Count of Predicted Instance Distribution by Naive Bayes Updateable Algorithm

Groups	Counts
1:C	9346
2:A	13454
3:B	10222
4:D	12411
5:F	4567
Total:	50000

According to the Naive Bayes Updateable tariff prediction model; 50000 one-time tariff consumer that are the entire test data can be billed from three-times tariff with the tabulated three-time consumption characteristics.

Table 5.10: Count of Predicted Instance Distribution by NBTree Algorithm

Groups	Counts
1:C	8473
2:A	13897
3:B	10865
4:D	8241
5:F	8524
Total:	50000

According to the NBTree tariff prediction model; 50000 one-time tariff consumer that are the entire test data can be billed from three-time tariff with the tabulated three-time consumption characteristics.

Table 5.11: Count of Predicted Instance Distribution by J48 Algorithm

Groups	Counts
1:C	8474
2:A	13898
3:B	10864
4:D	8240
5:F	8524
Total:	50000

According to the J48 tariff prediction model; 50000 one-time tariff consumer that are the entire test data can be billed from three-time tariff with the tabulated three-time consumption characteristics.

Energy-related data is examined under various topics in literature, topics can be mainly grouped the following relationship between customers' expectation and their preferences, consumption behavior and tariff design, consumption behavior based on customer segmentation, effect of the price on consumption.

This thesis, unlike the literature, provides prediction of tariff transition possibility from flat which means one-time to multiple which means three-time.

Outcomes show that, majority of one-time tariff customers can be billed from three-time tariff with the tabulated three-time consumption characteristics, in other words three-time tariff fit their usage time period and they can change their tariff with it which is cheaper than their current tariff.

6. CONCLUSION

Purpose of this thesis is prediction of tariff transition possibility from flat to multiple.

Energy-related data is examined under various topics in literature, topics can be mainly grouped the following relationship between customers' expectation and their preferences, consumption behavior and tariff design, consumption behavior based on customer segmentation, effect of the price on consumption.

This thesis, unlike the literature, provides prediction of tariff transition possibility from flat which means one-time to multiple which means three-time.

The energy-related data set contains tariff information, invoice amount and consumption basis kWh for each three-time periods which are day, peak and off peak.

The dataset composed of many different values, In order to find out correlation between the attributes, values of them must be divided into manageable number of groups.

Then following classification algorithms apply to data set Logıstıc Regression, RBF Network, SMO, Naive Bayes, Naive Bayes Net, Naive Bayes Updatable, J48, NBTree.

J48, NBTree and Bayes Net classification algorithms have highest accuracy rates which are 100 percent, 100 percent and 99.98 percent. Logistic Regression, SMO and RBF Network have the second best accuracy rates which are 91.54 percent, 91.05 percent and 89.99 percent. The remaining are Naive Bayes and Naive Bayes Updateable, correctly classified instances rates are 81.29 percent for both of them.

Results show that majority of flat tariff customers can change their tariff with multiple which is cheaper than their current tariff.

REFERENCES

Bouckaert, R. R. (2001). Bayesian belief networks: from construction to inference.

Broomhead, D. S., & Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks (No. RSRE-MEMO-4148). ROYAL SIGNALS AND RADAR ESTABLISHMENT MALVERN (UNITED KINGDOM).

Chicco, G., Napoli, R., Piglione, F.,Postolache, P., Scutariu, M. and Toader, C. (2002). A Review of Concepts and Techniques for Emergent Customer Categorisation.

Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological), 215-242.

ELECTRICITY MARKET LAW (2001).
Retrieved from

http://www.emra.org.tr/documents/electricity/legislation/ElectricityMarketLaw.doc

Kirschen, D. S., Strbac, G., Cumperayot, P. and Mendes, D. D. (2000).Factoring the Elasticity of Demand in Electricity Prices, IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 15, NO. 2.

Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: a decision tree hybrid. In Proc. of the 2nd Int. Conf. on Knowledge Discovery and Data Mining, pages 202–207

Platt, J. (1998) . Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.

Pombeiro, H., Pina, A., Silva, C. (2012). Analyzing Residential Electricity Consumption Patterns Based on Consumer's Segmentation. Proceedings of the First International Workshop on Information Technology for Energy Applications, Lisbon, Portugal.

Quinlan, J. R. (1986). Induction of Decision Trees. Mach. Learn. 1 (1), 81-106

Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

REGULATION ON SERVICE QUALITY IN ELECTRICITY DISTRIBUTION AND RETAIL SALE (2012).Retrieved from
http://www.emra.org.tr/documents/electricity/legislation/Elk_Yon_Quality_kH44D1kD6YBV.doc

Russell, S. , Norvig, P. (2003) . Artificial Intelligence: A Modern Approach (2nd ed.). Prentice Hall.

Slavickas, R. A., Alden, R. T. H. and El-Kady, M. A.(1999).Managing customer and distribution utility costs, IEEE Trans. Power Delivery, vol. 14, pp. 205–210.

Stephenson, P., Lungu, I., Paun, M., Silvas,I. and Tupu, G. (2001). Tariff Development for Consumer Groups in Internal European Electricity Markets, Proc. CIRED 2001, Amsterdam, The Netherlands, paper 5.3.

The Energy Sector:A Quick Tour for the Investor (2013). Retrieved from http://www.invest.gov.tr/en-US/infocenter/publications/Documents/ENERGY.INDUSTRY.pdf

Witten, I. H., Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques (Second Edition, 2005). San Francisco: Morgan Kaufmann