International Journal of Future Generation Communication and Networking

Vol. 13, No.2s, (2020), pp. 1609–1621

Prediction Of The Future Electricity Consumption And Production Using The Most Efficient Machine Learning Algorithm

Sohum R. Dhavale[1], Sakshi D. Mane Deshmukh[1], Khyati P. Shah[1], Pradyot J. Itkurkar[1]
[1]BE Student, Dept. Of Computer Engineering, NBN Sinhgad School of Engineering, Ambegaon (Bk), Pune-411041, Maharashtra,India

Abstract

The prediction of electrical energy demand is a matter of concern for many countries as the forecast of consumption of electricity is crucial for policy making. This paper presents an efficient way to predict the future load on the system by using various machine learning approaches. The data-set, consisting of numerous patterns of electricity consumed from various commercial and domestic areas, are the readings collected by the Axonet Smart Energy meter. To accomplish this task, this survey introduces the most efficient algorithm among these machine learning algorithms viz. Naïve Bayes Classifier Algorithm, Decision Tree Algorithm and Random Forest Algorithm to predict the electricity consumed using the historical data.

Keywords*: Electricity Consumption Prediction, Smart Energy Meter, Forecasting, Naïve Bayes Classifier, Decision Tree, Random Forest, Machine Learning, Data Analysis.*

I. INTRODUCTION

The demand for electricity is increasing drastically day by day in the household as well as in the industries. These needs being naturally fluctuating in nature, a lot of electricity produced is wasted. Thus, need to predetermine the electricity needs has become a matter of utmost importance nowadays. The proposed system introduces a electricity prediction model that uses the historical data of electricity consumed by the industrial as well as household users. A superior demand prediction is essential for building effective energy forecasting system. Therefore, accuracy in energy consumption forecast is very crucial [1]. In this new era, energy being a vital pillar for economic growth, precise forecast of electricity needs have become extremely crucial owing to the continuously changing rate of electricity consumption [2].

II. LITERATURE SURVEY

Paper [14] compares the results of data mining techniques like regression model, SVM model and Neural Nets for electricity consumption forecast and thus concludes the most efficient algorithm amongst them. The attributes of the dataset comprising electricity consumption data in the Islamic Republic of Iran-Mazandaran province considered are moisture, population, temperature and electricity units' price. After analysis, based on the relative error and correlation rates, the regression model was found to be more accurate and had a lesser error rate compared to the other two. The prediction results on an average showed increase in electricity consumption by 3.2% per annum, population being a major factor in the increment. Conducting analysis on monthly or daily data can give more accurate results.

According to paper [4] the energy to be generated by the photo-voltaic cells that are installed in distributed regions is predicted by using the Naïve Bayes classifier. The dataset which has the readings collected for the interval of one day is continuous-valued data, which is then converted into categorical -valued data with class labels viz. like 'very low', 'low', 'medium', 'high' and 'very high'. The attributes considered are daily global solar radiation, daily sunshine duration, daily average temperature and daily photo-voltaic energy generation, where 'daily photo-voltaic energy generation' being the input attribute and ' daily total global solar radiations', 'daily total sunshine duration' and 'daily average temperature' be the predicted output attributes. This also resulted in improvement of

ISSN: 2233-7857 IJFGCN

Sensitivity and Accuracy. Enhancing Sensitivity and Accuracy of the Naïve Bayes algorithm using different input parameters can be considered as future scope.

Research paper [9] proposes to predict the direct desired level in absence of regression using pattern recognition. It uses simple classifier models. The biggest challenge in forecasting the energy load is that even the consumer is unaware of the amount of consumption of the energy in the next hour. The client has no assurance weather the load will be low or high. The method mentioned below is the solution to the same problem. The data points of the dataset are equally partitioned into ordinal bins which represents three energy levels that are preprocessed numerical values carried out by Machine learning Random forest classifier. The above mentioned method of using the Random Forest Classifier serves as a better alternative to the Regression model. It gives more accurate results but at the same time is also time consuming.

Paper [8] proposes a less time consuming version of the Distributed Decision tree algorithm without even affecting the accuracy of the Decision tree algorithm. The time consuming tasks are made to perform in parallel to improve performance and also make scalable design. This Horizontal data parallelism is implemented in Spark environments' shared-nothing architecture. The datasets like click on ad prediction, wine quality, credit card fraud detection and skin segmentation were used. Spark's features like map-reduce, Resilient Distributed Dataset and its in-memory computation time and speed allows any scalable algorithm to perform easily. Thus, if the speedup factor is neglected, the performance of Spark is better than Planet which is a Hadoop implementation.

Paper [3] proposes a k-means algorithm approach to predict the thermal comfort states of the residents, which is divided into two main tasks viz. Modeling and Optimization. At first, the modelings of the total energy consumption, air and skin temperature and the skin temperature gradient are verified after their implementation. Then for optimizing, all the undesired frequencies are filtered out and optimal frequencies are located using the Augmented Firefly Algorithm i.e. AFA. As a result, the energy efficiency improved considerably, while maintaining the indoor temperature. While this approach fulfills the objectives of a smart building, the differing thermal comfort conditions of every single room may still not be satisfied

III. EXISTING SYSTEM

The traditional electric meters which are used today have rotors which helps in calculating the electricity used. They are inefficient, less accurate and lack in communication with the environment i.e. the user or the energy provider. They also have less privacy and minimal security compared to the smart energy meters[6]. As it is difficult for a consumer to predict its future electricity requirements, the producer thus can never get at least an approximate idea of the amount of electricity to be produced. This leads to either scarcity of electricity or wastage of the produced electricity and the resources involved in producing this electricity.

Also currently, personnel has to physically go to the sites wherever the meters are installed and collect the meter readings manually. This system is very exhausting in terms of money, time and manpower. So with the increasing population, the need to predetermine the electricity needs is drastically increasing for efficient utilization and wastage avoidance of the electrical energy

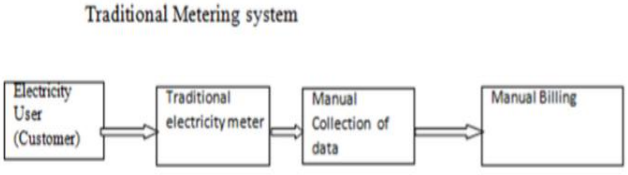

Fig -1: Traditional Electricity Metering System [6]

ISSN: 2233-7857 IJFGCN

IV. PPROPOSED SYSTEM

The smart energy meter has an embedded chipset which helps in calculating the current flow which then sends the collected data to the database. This database can be accessed by both - the electricity consumer as well as the energy providers, thus discarding the need of manual collection of meter readings. The network which allows the meters to communicate with the environment i.e users and energy providers, are of WAN (wide area network) and HAN (home area network). The proposed system will thus help the consumer predict its future electricity consumption and the producer can also foresee the future electricity requirements.This will also give insight to the producer what will be the future requirement of energy

IV. Working

[12] The data generated is collected by the Axonet Smart meters i.e. the smart energy meters, which are installed on-site and are generally used in domestic sector and industrial sectors. These meters are connected with a central system despite of being geologically distributed. The raw data from the location is sent to the Datonis API. It is the platform through which the data is channelized and managed using the SiteWare Instance. This JSON parsed data is stored in the Amazon Web Server thereafter.

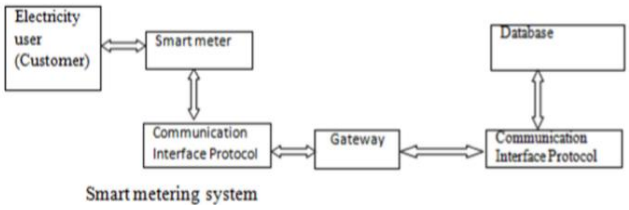

Fig -2: Smart Electricity Metering System [6]

The dataset from the AWS has different readings which is collected by different smart energy meters. The readings specify the amount of electric energy consumed by the user on a particular day and for a particular amount of time on which pre-processing techniques such as removing inconsistent data, data- cleaning, transformation and categorization are applied and the dataset for efficiently processed.

This dataset, which is initially partitioned into 2 parts - training dataset and testing dataset, generates a prediction model or classifier based on the training set. This model generated is now capable of predicting the future electricity requirements of a particular user based on its past consumption data as measured by the smart energy meters. This prediction model is thus is very beneficial electricity producers as well as its consumers.

ISSN: 2233-7857 IJFGCN

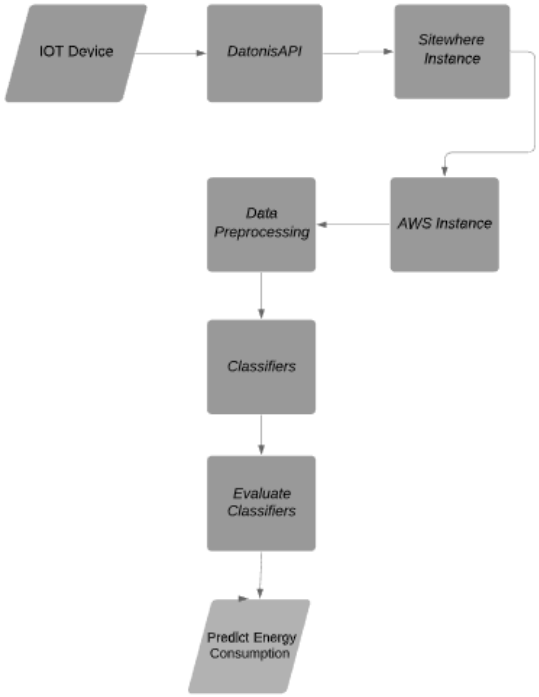

Fig -3: System Flow Diagram

To analyze the performance of the model to estimate how accurate the model will work in real-time the cross validation technique is used. The testing data is used to compute the confusion matrix. It shows the accuracy of the model in terms of Positive (P) and Negative (N) observation such that:

True-Positive (TP): Here the observation as well as prediction is positive.

False-Positive (FP): Prediction is positive but the observation is negative.

True-Negative (TN): Prediction as well as observation is negative.

False-Negative (FN): Prediction is negative but here the observation is positive.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Fig -4: Classification Rate/Accuracy

V. Methodologies

This section provides a quick summary of different Machine Learning Approaches.

Naïve Bayes classifier is a simple Bayesian classifier which is commonly used in the area of data mining. It is a generative model which returns probability values as output. Naive Bayes algorithm works contrary to the classification strategy of a one rule classifier. The Naïve Bayes classifier uses the assumption of conditional independence. That means, every attribute is conditionally not dependent on other attribute[7]. The Bayes' Theorem is represented as follow,

ISSN: 2233-7857 IJFGCN

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad P(C_i|\mathrm{X}) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Fig 5-: Naïve Bayes classifier

where, P(A/B) is the probability of event A in case B occurs, P(B/A) is the probability of event B in case A occurs, P(A) is the probability of A and P(B) is the probability of B [4]. The naive Bayes makes a prediction using Bayes' Theorem, which derives the probability of a prediction from the underlying evidence, as observed in the labeled data.

The Naive Bayes algorithm works as follows:

1. Calculate the prior probability
2. Find Likelihood probability with each attribute for each class.
3. Calculate posterior probability using Bayes formula.

Decision tree, a type of supervised learning algorithm, is a flowchart-like structure in which every internal node represents a test on any attribute, the branches denote the outcome of the tests, and the leaf nodes or terminal nodes holds the decision taken after computing all attributes i.e. the class label. The process of prediction of the class label for a record is started from the root of the tree by comparing values of the root attribute with that of the records' attribute. Based on that comparison, a branch is selected, thus reaching the corresponding internal node. This process is recursively carried out until the leaf node i.e. the target class is obtained.

The decision tree algorithm flow is as follows :

1. Pick the best attribute or feature which will split the dataset and place it at the root of the tree.
2. Perform a test on that feature and split the training dataset based on the test results.
3. Repeat steps 1 and 2 until leaf node is reached in all the branches.

Decision Tree Classifier class from Scikit learn can train the binary decision tree with Gini and cross-entropy impurity measure. Here let's consider the case with three features and three classes:

```
from sklearn.datasets import make_classification
>>> nb_samples = 500
>>> X, Y = make_classification(n_samples=nb_samples, n_features=3,
n_informative=3, n_redundant=0, n_classes=3, n_clusters_per_class=1)
```

Let's first consider a classification with default Gini impurity:

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import cross_val_score
>>> dt = DecisionTreeClassifier()
>>> print(cross_val_score(dt, X, Y, scoring='accuracy', cv=10).mean())
0.970
```

Using built-in-function export_graphviz():

```
from sklearn.tree import export_graphviz
>>> dt.fit(X, Y)
>>> with open('dt.dot', 'w') as df:
 df = export_graphviz(dt, out_file=df,
 feature_names=['A','B','C'],
 class_names=['C1', 'C2', 'C3'])
```

Following are the graph for the features of decision tree:

ISSN: 2233-7857 IJFGCN

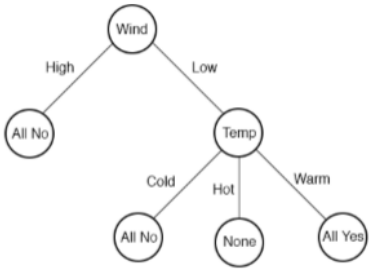

Fig -6: Sample Decision Tree

Random forest, a supervised learning algorithm which is mainly used for classification problems, is a model which consists of many individual decision trees that operate as an ensemble. The prediction results of each individual decision tree is considered and the final prediction is made by means of average voting.

The Random Forest algorithm works as follows :

1. Select test features from the given dataset and construct a decision tree for every sample.
2. Store the prediction results of each individual decision tree and perform voting for every predicted outcome.
3. Select the highest voted predicted outcome as the final prediction.

VI. DATASET

The case study of this research is done by the input provided by the smart energy meter. Therefore, the research data was collected from this region. The dataset employed in this research was collected annually. Moreover, the electricity consumption prediction results are based on the long-term study. The dataset variables are shown in table 1. These variables are classified into two groups.

Independent Variables	Temperature, Humidity, Pressure
Dependent Variable	Electricity consumption Rate (Use)

Table 1. The dataset variables

The group one of the dataset includes the independent variables that are effective on the energy consumption prediction rate. These variables include temperature, humidity and pressure. The other group has only a single dependent variable also known as the prediction, the electricity use (total consumption). The above statistics were collected from the smart energy meters installed by the Axonet Emsys. Before performing data analysis, the data is to be preprocessed. Data preprocessing is nothing but preparing the data for the main process, namely knowledge discovery, to start. In this stage, the data will be analyzed in terms of having errors such as the lost variables, noisy data, repetitive data, and false data. In case any of such errors exist, it should be resolved . Data integration, data selection and data conversion are other tasks that are to be performed in the preprocessing stage. In figures 7 to 10, the dataset is depicted as a graph.

ISSN: 2233-7857 IJFGCN

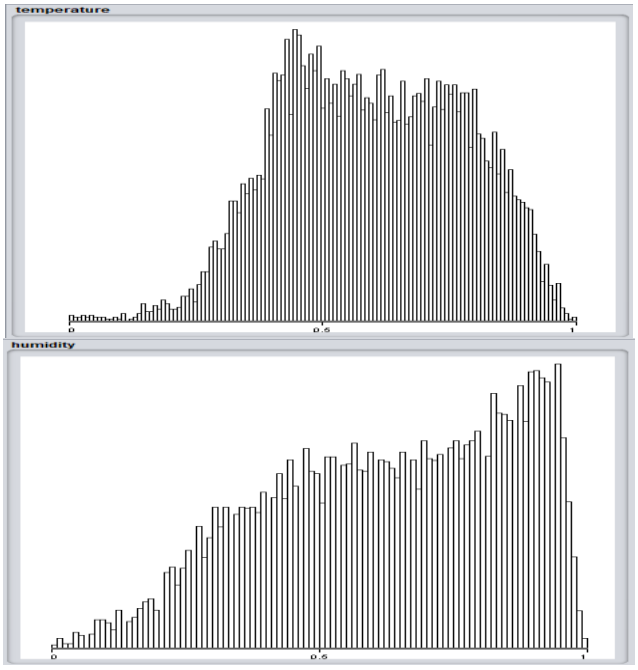

Fig -7. The normalized temperature
Fig -8. The normalized humidity

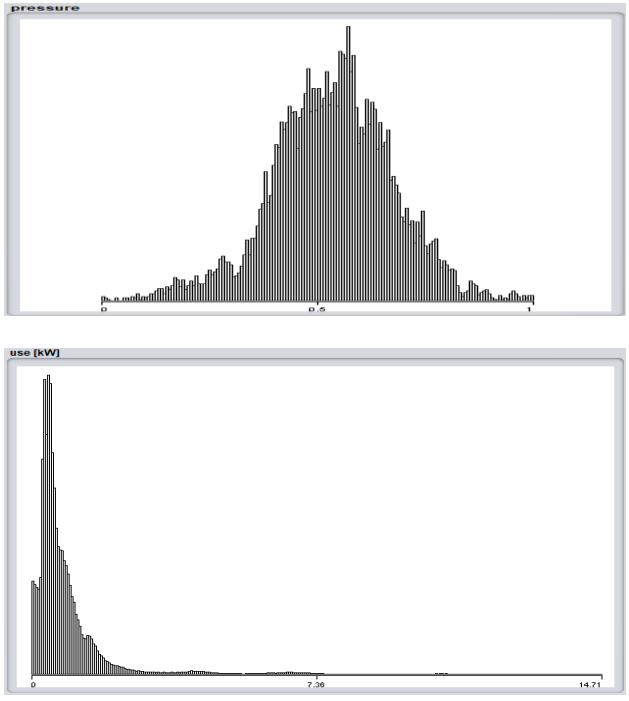

Fig -9. The normalized pressure
Fig -10. The average usage

Fig 11 shows the overall distribution of the energy consumption in watts on the y-axis to the linear time on x-axis. The line graph clearly shows some pattern in the usage of the energy with respect to the linear time.

ISSN: 2233-7857 IJFGCN

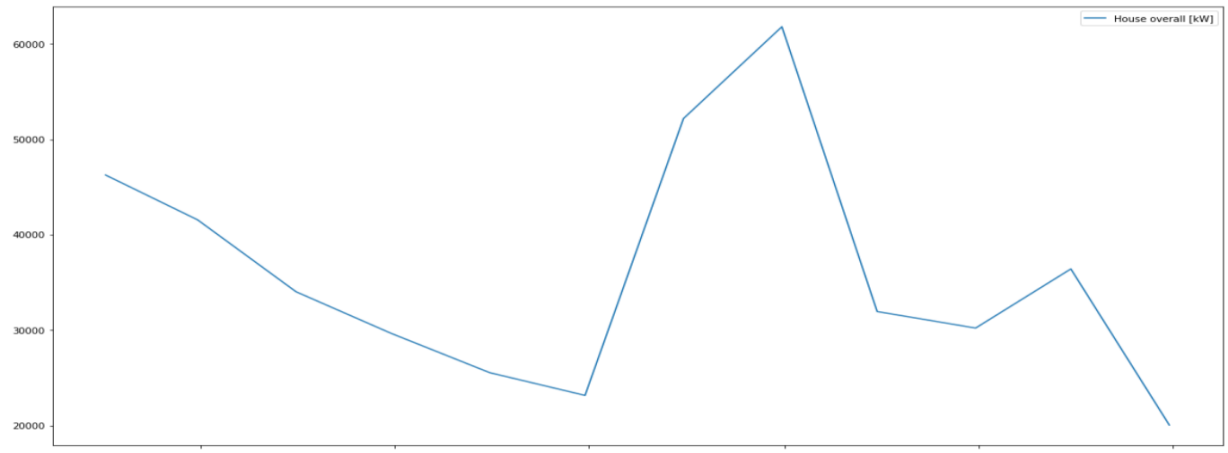

Fig -11. Overall usage of energy

VII. LIBRARIES

1. Numpy

This library contains mathematical tools for model building. Basically this library we need to include any kind of mathematics in our code. Numpy library adds support for large multidimensional arrays and matrices.

We can import this library using :

import numpy as np

2. Matplotlib

Matplotlib is a visualization library in python for 2D plots.This library allows us visual access to large amounts of data. This library consists of several plots such as line, bar, scatter and histogram.

We can import this library using :

import matplotlib.pyplot as plt

3. Pandas

Pandas library helps us for managing and importing the datasets. This is an open source library provides easy to use data structures and data analysis tools. It provides structures and operations for manipulating numerical tables and time series.

We can import this library using :

import pandas as pd.

VIII. DATA PREPROCESSING

Dealing with Missing Values

When we consider a real world dataset, it may consist of missing data i.e. incorrectly encoded data or such type of data that is inappropriate for modelling, so there are a few options that can be taken into account

1) Case wise deletion of missing data. In this, the cases or rows which consist of missing values are deleted permanently from the dataset. This approach suits for large dataset which have very few missing features.

2) Creating a sub model to predict those features. This approach is difficult because it is required to determine a supervised strategy to train that model with each feature and then to predict their value.

3) Replace the missing values with the mean or median value of the feature in which they occur. This is used in case of numerical feature values. Scikit learn offers the class *Imputer* which strategically fills the missing values based on mean, median or frequency. Here mean is the default choice.

ISSN: 2233-7857 IJFGCN

The following code shows an example using these approaches.

```
from sklearn.preprocessing import Imputer

>>> data = np.array([[1, np.nan, 2], [2, 3, np.nan], [-1, 4, 2]])

>>> imp = Imputer(strategy='mean')
>>> imp.fit_transform(data)
array([[ 1. ,  3.5,  2. ],
       [ 2. ,  3. ,  2. ],
       [-1. ,  4. ,  2. ]])

>>> imp.fit_transform(data)
array([[ 1. ,  3.5,  2. ],
       [ 2. ,  3. ,  2. ],
       [-1. ,  4. ,  2. ]])

>>> imp = Imputer(strategy='most_frequent')
>>> imp.fit_transform(data)
array([[ 1.,  3.,  2.],
       [ 2.,  3.,  2.],
       [-1.,  4.,  2.]])
```

Fig -12. Preprocessing Imputer

Here the default value for missing feature entry is np.nan, however we can use a different placeholder for missing_value feature.

Splitting the dataset into training and testing set

from sklearn.model_selection import train_test_split
>>> X_train, X_test, Y_train, Y_test = train_test_split(X,
test_size=0.25, random_state=1000)

In the case above, the ratio for training dataset is 75 percent and the same for testing dataset is 25 percent. To accept the NumPy RandomState generator or an integer seed, random_state is one of the important parameters.

from sklearn.utils import check_random_state
from sklearn.utils import check_random_state
>>> rs = check_random_state(1000)
<mtrand.RandomState at 0x12214708>
>>> X_train, X_test, Y_train, Y_test = train_test_split(X, Y,
test_size=0.25, random_state=rs)

Feature Scaling

A numerical dataset made up of different values which are from different distributions, different scales. A machine learning algorithm is unable to differentiate among these various situations, and hence, it is always preferred to standardize datasets before processing them.

ISSN: 2233-7857 IJFGCN

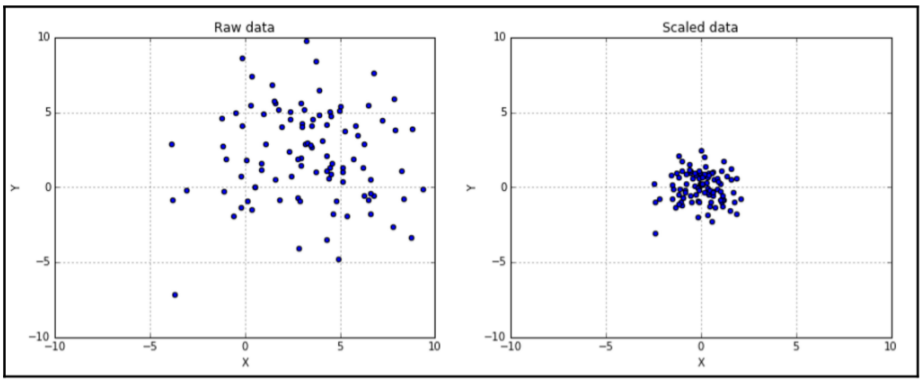

Fig -13. Comparison between a raw dataset and the same dataset scaled

The result shown in above figure can be achieved by using *StandardScalar* class.

```
from sklearn.preprocessing import StandardScaler

>>> ss = StandardScaler()
>>> scaled_data = ss.fit_transform(data)
```

Fig -14 .StandardScaler

IX. APPLICATION

This system provides the benefits to the consumer as well as producer, By analysing the historical data of electricity usage by consumer and producer, this system provides benefits to the consumer as well as producers by predicting future consumption of electricity. This will reduce the wastage and shortage of electricity by helping the producers to anticipate the future needs.

With the help of this, consumer will regularly monitor and analyse the consumption pattern which leads to altering the use of consumption of electricity. On the commercial use side of this system, it helps the organization to plan the energy overhead efficiently for particular project or system.

X. RESULTS

We built the desired machine learning models using statistical predictive algorithms and then selected the best model in terms of accuracy, precision and recall. The results obtained are shown in Table 2.

Model	Accuracy (%)
Random Forest	97.00
Decision Tree	96.96
Naive Bayes	96.53

Table 2. Result Obtained

As mentioned, the dataset was split into a training set and a testing set. The training set and testing set was splitted into the ratio of 75 - 25 percent. This train/test split was used for comparing

ISSN: 2233-7857 IJFGCN

the results. Here we have used binning to group the continuous variable i.e. electricity consumption rate (use) into smaller number of bins.

As you can see in table 2, Random Forest is the most suitable model, resulting in greater accuracy.

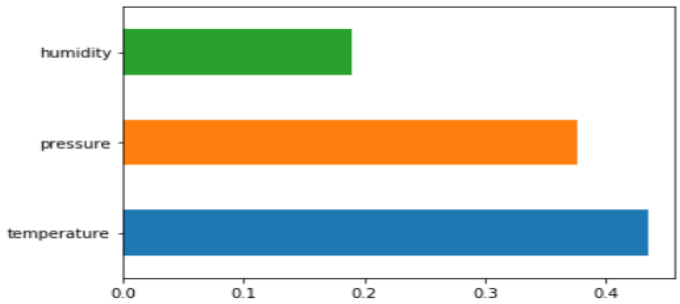

Fig -15 .Importance Factor

Variables	Importance Rate (%)
Temperature	42.26
Pressure	36.72
Humidity	21.00

Table 2. Feature Importance

As we have mentioned earlier, three independent variables were employed for this prediction. These include temperature, pressure and humidity. In table 3, the importance rate of the variable on prediction is shown.The temperature field shows most effect on the electricity consumption rate. As shown in table 2, the model built using the random forest method has greater accuracy rate and hence its performance is better as compared to other methods.

```
Confusion Matrix :
[[120826    535    219     27      0]
 [  1864    749     66     21      0]
 [   664     91    548     52      0]
 [   107     38     67     76      0]
 [     7      3     10      8      0]]
Accuracy Score : 0.9700026988839321
Report :
              precision    recall  f1-score   support

           0       0.98      0.99      0.99    121607
           1       0.53      0.28      0.36      2700
           2       0.60      0.40      0.48      1355
           3       0.41      0.26      0.32       288
           4       0.00      0.00      0.00        28

avg / total       0.96      0.97      0.97    125978
```

Fig -16. Random Forest Output

As we can see in output the accuracy percentage in random forest is 97 percent and average precision rate is 96 percent while the average recall rate is 97 percent. Here we have taken n_estimators = 500 i.e number of decision trees. This output is influenced by votes of 500 decision trees.

ISSN: 2233-7857 IJFGCN

```
Confusion Matrix :
[[120830    524    211     42      0]
 [  1884    729     66     21      0]
 [   675     94    537     49      0]
 [    98     45     84     61      0]
 [     6      0     13      9      0]]
Accuracy Score : 0.9696693073393767
Report :
             precision    recall  f1-score   support

          0       0.98      0.99      0.99    121607
          1       0.52      0.27      0.36      2700
          2       0.59      0.40      0.47      1355
          3       0.34      0.21      0.26       288
          4       0.00      0.00      0.00        28

avg / total       0.96      0.97      0.97    125978
```

Fig -17. Decision tree output

```
Confusion Matrix :
[[121607      0      0      0      0]
 [  2700      0      0      0      0]
 [  1355      0      0      0      0]
 [   288      0      0      0      0]
 [    28      0      0      0      0]]
Accuracy Score : 0.9653034656844846
Report :
             precision    recall  f1-score   support

          0       0.97      1.00      0.98    121607
          1       0.00      0.00      0.00      2700
          2       0.00      0.00      0.00      1355
          3       0.00      0.00      0.00       288
          4       0.00      0.00      0.00        28

avg / total       0.93      0.97      0.95    125978
```

Fig -18. Naive Bayes Output

The above output depicts the result of the decision tree and naive Bayes algorithm, here if we compare the accuracy then we observe that the accuracy of the decision tree was marginally less than random forest (96.96%). In comparison to this naive Bayes accuracy was about 5% less.

XI. CONCLUSION AND FUTURE SCOPE

In this paper, based on analysis of ensemble learning, the random forest model generated relatively less error on the dataset in comparison to other models towards energy consumption and prediction. Hence, a random forest model was employed. The dependent variables include temperature, pressure and humidity. The most prominent variable in energy consumption is temperature with the importance rate of 42.26 percent.

The matching percentage of the predicted and actual values influences the efficiency of the classifier. Forecasting energy consumption is important to overcome the wastage of energy while producing and also helps to monitor the commercial energy system. This system is beneficial to the producer as it allows to foresee the future forecast of energy consumption. Due to this approach, the structured management of energy consumption can be done by the producers, in such a way that there is neither shortage nor the wastage of energy.

ISSN: 2233-7857 IJFGCN

REFERENCES

[1] C. J. Chang and C. D. Li and C. C. Chen and W. C. Chen, Forecasting short term electricity consumption using the adaptive grey-based approach—An Asian case, OMEGA journal, 40, 2012, 767-772.

[2] E. Mocanu, P. H. Nguyen, M. Gibescu, and W. L. Kling, "Deep learning for estimating building energy consumption," Sustainable Energy, Grids and Networks, vol. 6, pp. 91–99, Jun. 2016.

[3] Deqing Zhai, Tanaya Chaudhuri, Yeng Chai Soh,"Energy Efficiency Improvement with k-means Approach to Thermal Comfort for ACMV Systems of Smart Buildings",IEEE,2017 Asian Conference on Energy, Power and Transportation Electrification (ACEPT)

[4] Ramazan Bayindir, Mehmet Yesilbudak, Medine Colak, Naci Genc, "A Novel Application of Naïve Bayes Classifier in Photovoltaic Energy Prediction", 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)

[5] Samuel Bimenyimana, "Traditional Vs Smart Electricity Metering Systems: A Brief Overview", Journal of Marketing and Consumer Research, Vol.46, 2018

[6] Le, T. N., Chin, W. L., Truong, D. K., & Nguyen, T. H. (2016). Advanced Metering Infrastructure Based on Smart Meters in Smart Grid. In Smart Metering Technology and Services-Inspirations for Energy Utilities. InTech.

[7] S.S. Nikam, "A Comparative study of classification techniques in data mining algorithms", Oriental Journal of Computer Science & Technology, vol. 8, No. 1, pp. 13-19, 2015

[8] Siddalingeshwar Patil, Umakant Kulkarni, "Accuracy Prediction for Distributed Decision Tree using Machine Learning approach ",Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019), IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8

[9] Yu-Tung Chen, Eduardo Piedad Jr., Cheng-Chien Kuo ,"Energy Consumption Load Forecasting Using a Level-Based Random Forest Classifier",Creative Commons Attribution(CC BY) license (http://creativecommons.org/licenses/by/4.0/).,29 July 2019

[10] Krishna Prakash N, Prasanna Vadana D, "Machine learning based Residential Energy Management System", 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)

[11] Noorollah Karimtabar, Sadegh Pasban, Siavash Alipour,"Analysis and predicting electricity energy consumption using data mining techniques- A case study I.R. Iran - Mazandaran province ", 2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA 2015) March 11-12, 2015

[12]Sohum R. Dhavale, Sakshi D. Mane Deshmukh, Khyati P. Shah and Pradyot J. Itkurkar, "Survey on the most efficient machine learning algorithm to predict the future electricity consumption and production",International Journal of Innovative Research in Science, Engineering and Technology, Vol. 8, Issue 12, December 2019

ISSN: 2233-7857 IJFGCN