# PREDICTIVE ANALYTICS FOR ENERGY CONSUMPTION IN SMART HOMES WITH FOG AND CLOUD COMPUTING USING SUPPORT VECTOR REGRESSION

*Sofiene Haboubi*, Oussama Ben Salem*

Signals Images and Information Technologies Lab. University of Tunis El Manar,
National Engineering School of Tunis, BP 37 Le Belvedere, 1002 Tunis
Tunisia.

* Corresponding Author, e-mail: sofiene.haboubi@enit.utm.tn

**Abstract:** Predict energy consumption in smart homes is our objective in this work. We defined the two predictive analytics techniques, and we will be tested in this study. We presented the Linear Regression and the Support Vector Regression data mining techniques. Implemented those two machine learning models with the appropriate techniques, starting by cleaning and preparing the data then we visualize it so that we can uncover hidden information about the behavior of the smart home appliances using the energy consumption feature. Afterwards, implementing the two regression models to predict the whole house energy consumption. Compared the performances of the two techniques. The results achieved are promising and proves the reliability of the IoT smart home platform.

**Key words:** Smart Home, Energy consumption, Cloud Computing, Fog Computing, Support Vector Regression, Support Vector Machine

## 1. INTRODUCTION

Advances in Wireless Sensor Networks technologies have contributed a lot in realizing the modern smart home. With a noticeable computing power and adapted communication protocols for these small devices, data is being generated at an alarming rate.

The need to capture and analyze this data has become vital to understand and uncover hidden valuable information which can have a major impact on our economy, our lifestyle and even our safety. For example, IoT Big Data Analytics is when utility companies use data generated from smart cities to recommend electricity bill reduction plans based on a personalized profile for each home. Thus, it leads to reduce costs for both utility companies and homeowners and it also saves energy [1].

It is basically a giant network of connected devices that gather data and communicate through well adapted protocols such as MQTT [-4] and Zigbee [5]. It describes the next generation of the internet where the physical objects are identifiable through Wireless Sensor Networks (WSN) [6] and are able to exchange and process data accordingly to predefined situations. WSN technology became popular over the last years due to its unique characteristics; low cost, energy efficient, small physical size, computational power, communication capabilities, distributed sensing etc. [6]. Thus, it is the perfect fit to build a smart home where all appliances are connected

IoT Big Data Analytics have a lot of advantages as it helps understand the behavior of smart homes and their occupants, but it is quite challenging as it deals with huge amounts of data that are being generated each second. Hence, a system able to manage the growing volume of data and to analyze it in near real-time to provide actionable insights has to be set.

A platform for IoT smart home big data analytics has been proposed by many researchers [2, 3].

The purpose of our study is to take a different approach from other proposals by:

- including historical data, weather and whether or not the day is a day off to predict electricity consumption;
- hourly prediction of energy consumption.

## 2. MATERIALS

In this case study, we perform IoT data analytics on two years of data generated by appliances of a smart home in Vancouver, British Columbia, Canada. This analysis has a goal of predicting energy consumption in a defined period of time based on historical data of energy consumption, the day, the hour of the day and the weather.

This analysis is performed in the fog node as long as dealing with time series and continuous streams of IoT data that need to be accomplished in near real-time in order to meet the prerequisites of applications like electricity bill reduction plans.

Carried out analytics on energy consumption of the smart home appliances with time series where each smart home is attached to a fog node, by applying a machine learning supervised model on a dataset that consists of one hour interval measurements of multiple smart home appliances over the span of two years to predict energy consumption of each smart home.

### 2.1. The Almanac of Minutely Power dataset (*AMPds2*)

The data of the AMPDs2 was collected from a smart home in Vancouver, British Columbia, Canada. It has 730 days of captured data per meter [4]. It is publicly available for download from Harvard Dataverse in different formats, this dataset has 1,051,200 reading per each meter and contains twenty one power meters, two water meters and two natural gas meters datasets [5]. It also contains hourly weather data for the same period of time.

For this research purpose, we are only interested in the electricity datasets and the weather dataset.

The first column represents the Unix timestamp which is the amount of seconds since 01-01-1970 12:00:00am (UTC) so the first timestamp is equivalent to 04-01-2012 07:00:00am (UTC) incremented by one minute until 04-01-2014 06:59:00am (UTC). The second column represents the voltage in volts, the third column represents the current in amperes, the fourth column represents the frequency in hertz, the fifth column represents the power factor which ranges between minus one and one, the sixth column represents the apparent power factor which also ranges between minus one and one, the seventh column represents the real power measured in VAR, the eighth column represents the real energy in VAR per Hour, the ninth column represents the reactive power in VAR, the tenth column represents the reactive energy in VAR per Hour, the eleventh column represents the apparent power in VA and the twelfth columns represents the apparent energy in VA per Hour.

The columns are the same for the twenty-one datasets of the different appliances.

### 2.2. Regression Analysis Techniques

Smart homes generate continuous stream of data, and since wanted to predict energy consumption based on historical data, the hour of the day, the month and the weather. Nothing better than a regression model to get the job done. A regression analysis is used to predict a variable by investigating the relationship between dependent and independent variable. It is most used for predicting prices based of given features or energy forecasting.

Regression models are a supervised machine learning technique, in work we will be interested only in two of those techniques: Linear Regression and Support Vector Regression.

***Linear Regression:***

Linear Regression is the most common and popular regression technique. In this technique, we use an input variable $X$ to predict an output variable $Y$. The relationship between those two variables is linear and can be written like: $Y = a+bX$. We want to predict energy consumption of HVAC system based on the temperature values. We can see that there's a correlation between these two variables. An increase in temperature means an increase in HVAC system energy consumption.

$Y$ represents the HVAC system energy consumption and $X$ represents the temperature. If wanted to predict $Y$ for a value of $X$ we need to calculate the coefficients $a$ and $b$.

Measured the error by a cost function in most is Mean Squared Error (MSE). So, for a Linear Regression model we try to find the coefficients when the cost function is at its minimal.

Once we get the regression line with a minimal error, we have to be careful in choosing a value *x* in order to avoid prediction of a *y* value for a value of *x* that is outside the range of the data.

Once the regression line is obtained, caution should also be used to avoid prediction of a *y* value for any value of *x* that is outside the range of the data. [6].

***Support Vector Regression:***

Support Vector Machine (SVM) is a supervised learning model that is quite popular for its use in classification problems. Support Vector Regression (SVR) [7] is almost similar to SVM but it is used for regression problems. Even although it is not as popular as SVM, SVR is an effective algorithm in real-value datasets predictions. Unlike Linear regression where tried to minimize the cost function, in SVR we try to fit the error within an optimal threshold. With this margin of error, noisy values of real datasets like AMPds2 are better fitted with this.

## 3. IMPLEMENTATION

In this work, we assume that the data generated by the smart home is acquired at the fog node. In the fog node, an analytical engine will perform all the data processing.

### 3.1. Data cleaning and preparation

Data cleaning and preparation is a vital step in any type of analytics since it ensures a good quality of data that is ready to mine by a specific algorithm.

In this study, going to predict energy consumption and to accomplish that only needed the apparent power column so we only keep the date/time column and the apparent power column of each dataset. We merge the datasets of each appliance into one final dataset where each column represents the apparent power of a certain appliance. Identified the standby power threshold of each appliance by plotting the graph of its energy consumption then we set the standby power threshold to zero in order to be able to identify when an appliance is switched on.

Figure 1 shows an example of the kitchen oven energy consumption, set y-axis limit to 50 VA to be able to identify the standby power threshold. Set the kitchen oven standby power threshold to 12 VA and repeat the same process for each appliance.

Once the standby power thresholds is set to zeros, imported the hourly weather dataset from the AMPDs2 [14]. From the weather dataset, we grab only the temperature column and then replace the missing value with the median of the measured temperatures.

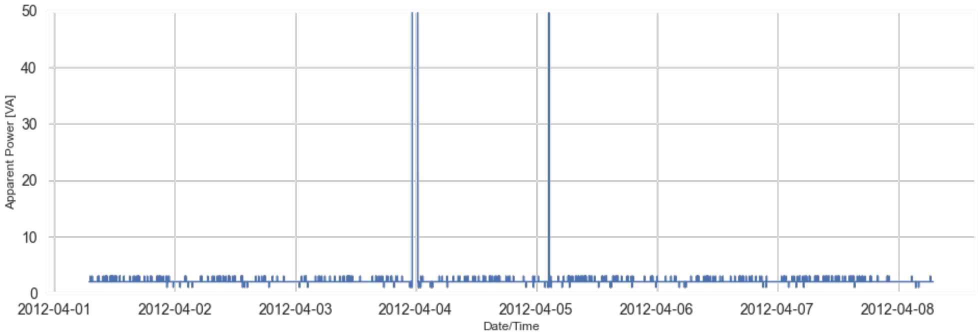

*Figure 1. The Kitchen oven energy consumption*

After that, merging the weather dataset with the appliances apparent power measurements dataset. By doing that, a lot of missing values are generated since the two datasets have different lengths (Minutely time series versus Hourly time series) so replacing each missing temperature value with the preceding value since the temperature is the same for the same hour. We add a month and an hour of the day columns because appliances and users behaviour in energy consumption changes from hour to hour and from month to month.

Another significant variable often neglected is whether or not a day is a day off (Weekends and Holidays)[1] because electricity usage during a day off is quite different than a typical day. Then create a categorical variable column where if the day is a weekend or a holiday it has a value of one otherwise zero. Resample our data by hourly mean.

Energy estimation falls under multi-step time Series estimation. Need to know the historical values of the target measure that we want to predict. For example, if needed to predict the next hour of energy consumption for a certain appliance (in our case the whole house energy consumption), we have to provide the actual value of the apparent power, the values of the previous hours and the corresponding weather measurements. We will be using a historical data of one day (24 hours) to generate a new dataset.

### 3.2. Data visualisation

Data visualization is not of any less significance than other steps since it helps us uncover valuable information about the behaviour of the smart home appliances and the occupants. We will also visualize temperature variations and their correlations with energy consumption. Started by verifying the veracity of our dataset. Figure 2 shows the energy consumption of the furnace fan. The graph shows that the furnace fan energy consumption is at its peak from October until April which makes total sense because during winter a furnace fan circulates hot air inside the house which will not be needed as much in the summer. Clothes washer is usually

[1] Statutory holidays in Canada both national and provincial

used when the homeowner is at home, which means out of work hours. As the graph shows it is in mostly used in the evenings and nights which proves the veracity of our data. Figure 3 shows the energy consumption of the clothes washer in relation to the hour of the day.

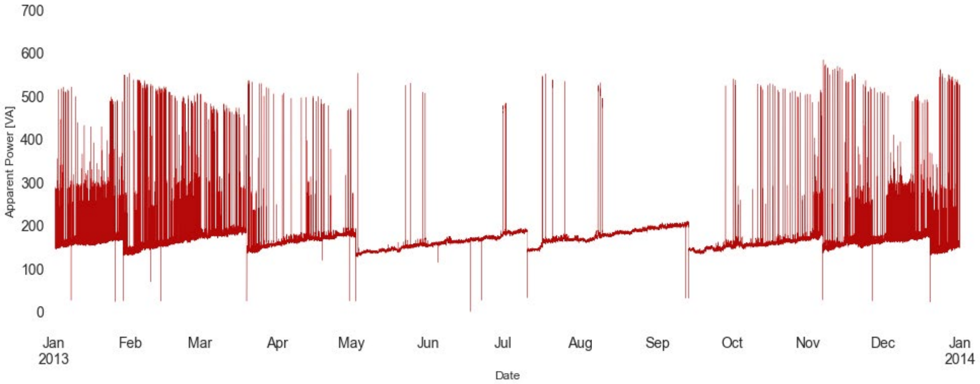

*Figure 2. The Furnace Fan energy consumption.*

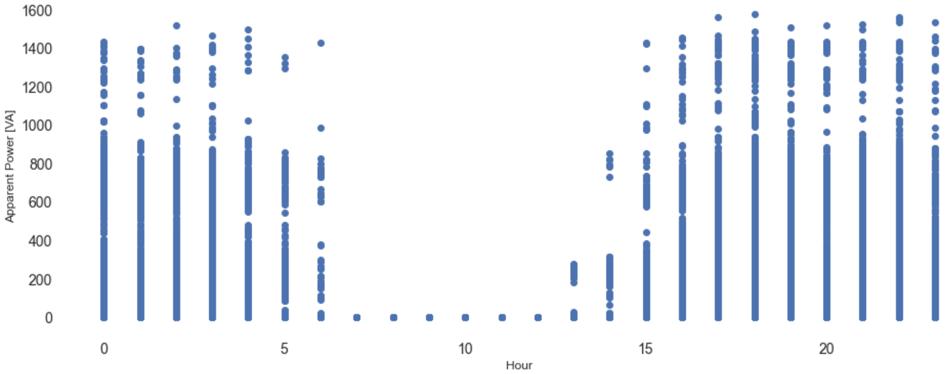

*Figure 3. The Clothes Washer energy consumption in relation to the hour of the day.*

Another important feature that affects energy consumption is the temperature. Figure 4 shows the temperature variations in Vancouver, British Columbia, Canada over the span of two years.

One of the steps we did in the data preparation phase, is that we resampled the data by hourly mean. We want to verify if the resampled data keeps the same information about the appliances and users behaviour as the original one.

Figure 5 shows the minutely energy consumption of the kitchen oven versus its hourly mean. As shown in the graph, hourly resampled data shows the same behaviour of the appliance as the minutely measures. Thus, working with the hourly resampled dataset would be a better alternative since it speeds up calculations. Instant analytics are a crucial characteristic in the fog nodes.

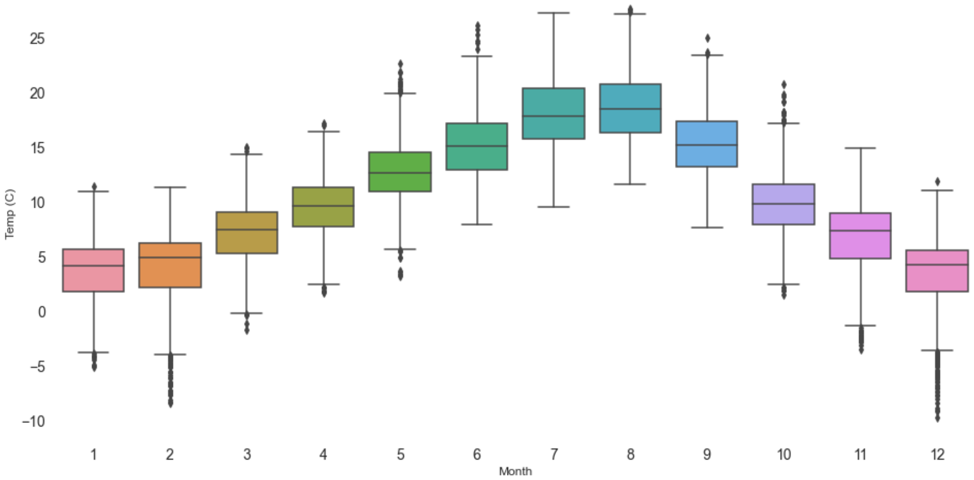

*Figure 4. The temperature variations in Vancouver, British Columbia, Canada.*

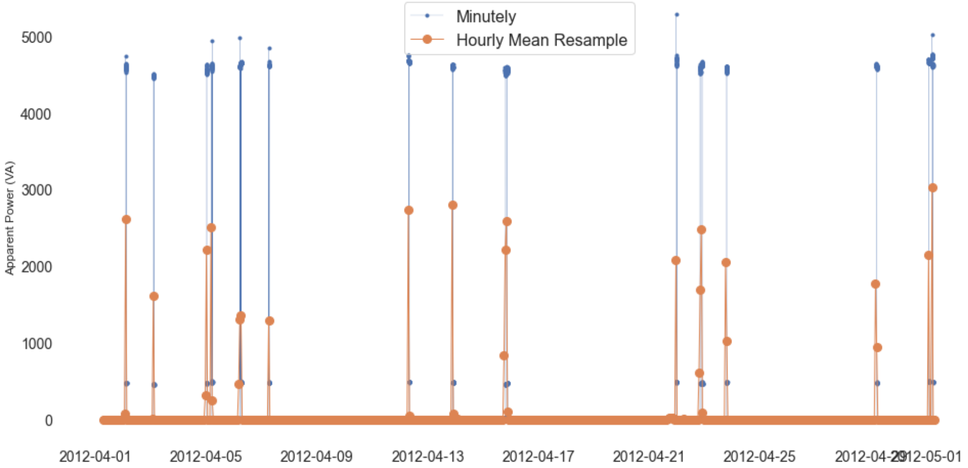

*Figure 5. The minutely energy consumption of the kitchen oven versus its hourly mean.*

On the next step, Resampled the smart home dataset in a monthly mean to verify once again the veracity of the data and to estimate the kitchen fridge energy consumption compared to the whole house consumption. Figure 6 shows the kitchen fridge share of the whole house energy consumption for a one year span. The kitchen fridge takes a higher share out of the whole house energy consumption during summer. Hence, the results confirms that the temperature has significant influence on the energy consumption.

Another way to clearly see this correlation is when visualize the temperature variations versus the energy consumption in the same graph. Figure 7 shows the variations of the temperature values versus the variations of the heat pump energy consumption for the year 2013. As it is shown in the graph when temperature increases, the heat pump energy decreases, which makes total sense.

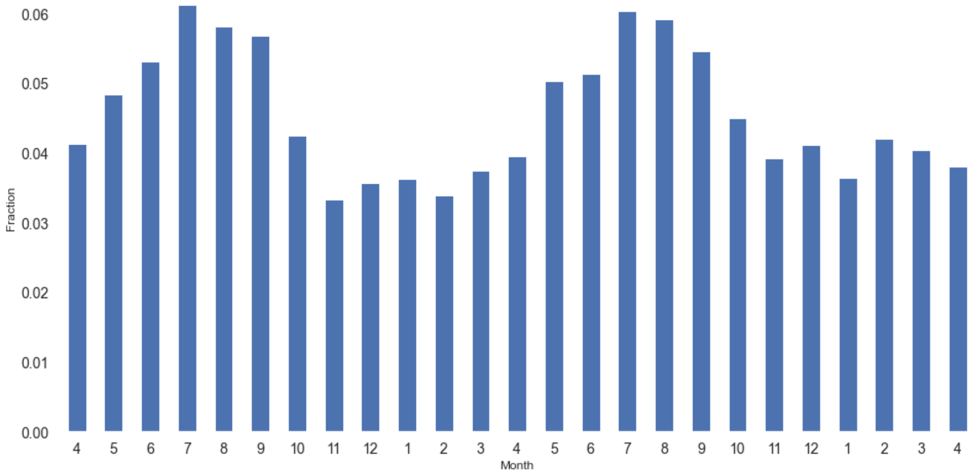

*Figure 6. The Kitchen Fridge share of the whole house energy consumption.*

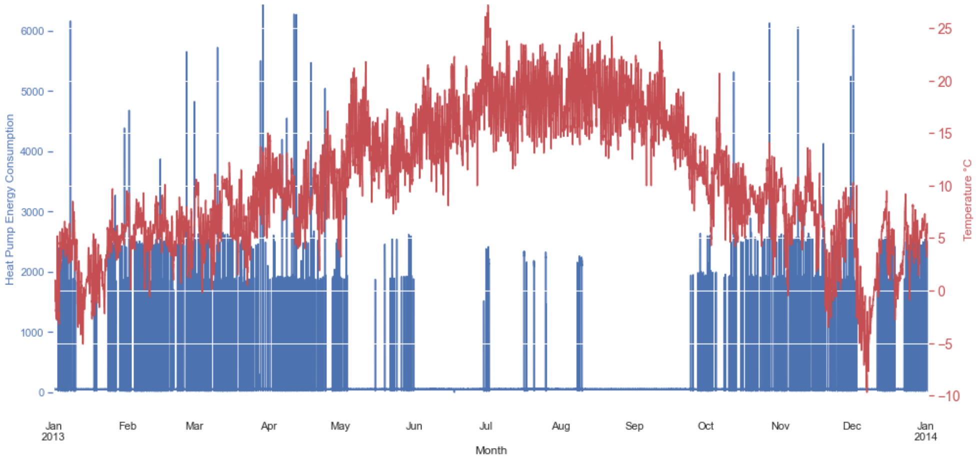

*Figure 7. Heat Pump energy consumption and Temperature.*

The regression models will do the predictive analytics in order to predict the whole house energy consumption based on the weather, the hour of the day, whether or not the day is day off, the month and the historical data of the energy consumption of the smart home. The probability distribution of the whole house energy consumption is checked. Figure 8 represents the probability distribution of the whole house energy consumption.

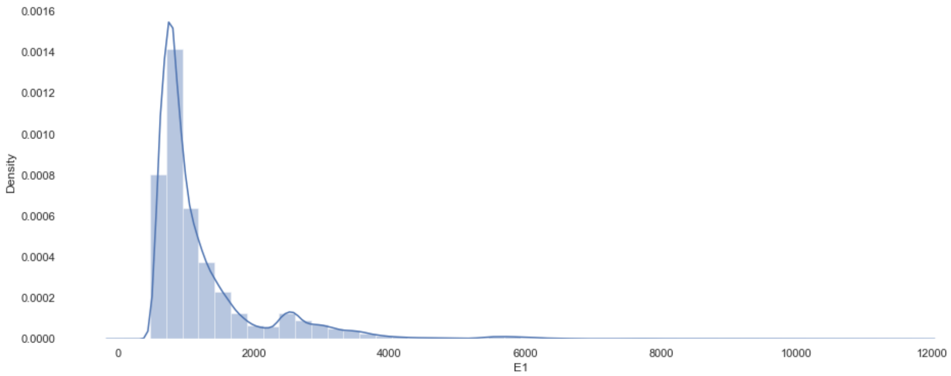

*Figure 8. Probability distribution of the whole house energy consumption.*

The average of the whole house energy consumption is around 1000 VA, and from the average it's not perfectly distributed thus it is a challenging case to test out the machine learning models we'd perform later.

One final step to reveal more valuable knowledge about our dataset. The correlation matrix between all the columns is calculated. The correlation matrix representation shows low correlation amounts in most cases which is another challenge to test the reliability of the regression models we'll use to predict the whole house energy consumption.

### 3.3. Regression models implementation

Starting first with Linear Regression after that the Support Vector Regression will be performed. Beginning by splitting our dataset into features that the model will train on and a target variable that the model will predict. Then, we split it into a training set and a test set.

An additional step is to scale the variables. Feature scaling is used to normalize the data, which in certain algorithms like Support Vector Machines it speeds up the calculations. Even though it is not necessary for Linear Regression since it the outcome is the same whether the data is scaled or not. We will use it for both regression models so we can objectively compare the results.

- Linear Regression: fit our training set inside the model. Once the Linear Regression model is trained, grab the coefficients to see how each the features relate to the target. To evaluate our model, calculate the R squared which is the proportion of the variance in the dependent variable that is predictable from the independent variable. R squared is between 0 and 1 and can be negative which means that the model is way below the mean value. When the *R squared* is equal or superior to 0.29 [8], it is considered substantial. R squared of the Linear Regression model on this dataset equals 0.296. It is acceptable.

- Support Vector Regression: fit our training set inside the model. We use the radial basis function kernel which is in the form of a Gaussian function. Tuned the model parameters (Gamma and C) to optimize the results. The Gamma parameter is the inverse of the radius of influence of samples selected by the model as support vectors. The C parameter gives the model freedom to select more samples as support vectors. R squared for the Support Vector Regression model on our dataset equals 0.308.

## 4. RESULTS

After implementing the two regression models, the *R squared* for both is calculated. *R squared* (0.296) for the Linear Regression model on this dataset is substantial. Plotted the predicted values of the smart home energy consumption by Linear Regression model and the actual values in the same graph to be able to identify how accurate the results are. Figure 9 shows the predicted values by the linear regression model versus the real values of the whole house energy consumption.

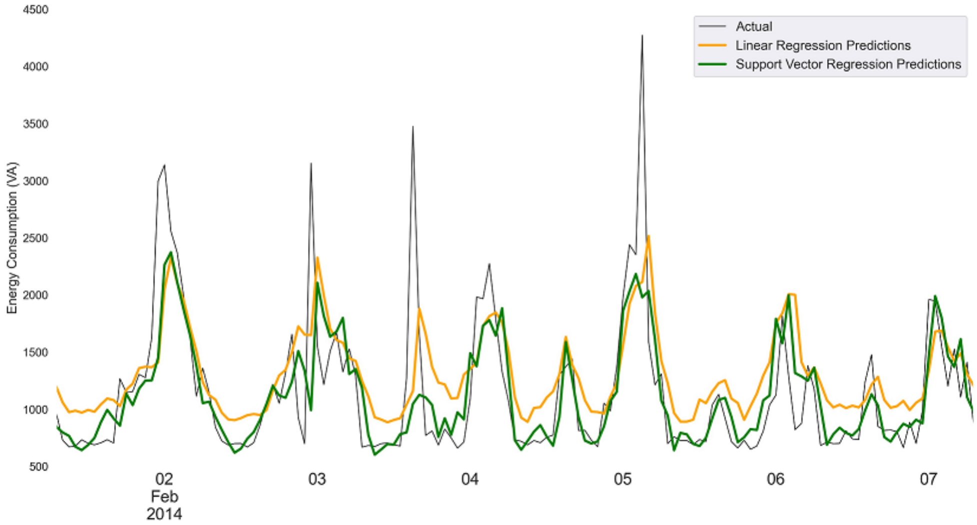

*Figure 9. The whole house energy consumption predicted values by linear regression and support vector regression.*

As shown in the linear regression graph, the prediction was acceptable as it clearly captures the shape of the curve. It didn't predict perfectly the values, but it shows clearly the variations between the minimums and the maximums, and it almost encapsulates all the peaks.

On the other hand, R squared for the Support Vector Regression (0.308) is calculated. Then, plotted the values of the whole house energy consumption predicted by the SVR model and the actual values in the same graph in order to see how they relate to each other's. Fig 9 shows the predicted values by the support vector regression model versus the real values of the whole house energy consumption.

As shown in the support vector regression graph, the prediction was accurate enough in capturing all the variations. It's clearly the predicted values by both the linear regression and the support vector regression models versus the real values of the whole house energy consumption

## 5. DISCUSSION

Our approach in this paper differs from the proposed by Yassine [3], where we performed the regression data mining technique in the fog node in order to predict energy consumption. Those predictions can be used by utility companies to recommend electricity bill reduction plans. It can also be quite beneficial for companies that are in the advertisement business where they can use an appliance energy consumption prediction to pass on targeted commercials based on a specific profile.

In this study is performed two regression data mining techniques: Linear regression which is a popular technique, it generates the best fitting line using ordinary least squares that minimizes the sum of the squared prediction error. And support vector regression, which is a less popular technique, it uses the same logic as support vector machine but instead of predicting classes it predicts values.

From the obtained results of the two models performances, it is quite clear that the SVR model generated better predicted values. Linear regression didn't capture the correct shape of the curve unlike support vector regression. Thus, performs the predictive analytics in the fog node using the SVR model since it is helpful in dealing with the limitations related to the probability distribution and the shape of the data.

## 6. CONCLUSION

The results obtained were promising as the two regression models outputs were sustainable, but the support vector regression model generated better predicted values. Hence, SVR is the suitable predictive data mining technique to be used in the fog node for smart home electricity usage measurements.

In this study, we confirmed that the platform proposed is reliable to perform IoT big data analytics and we brought a new predictive analytics technique that is quite helpful for in-time decision making process.

The predictive analytics performed in the fog node can manage the continuous incoming streams of data and sending the uncovered knowledge about the appliance's energy consumption to the cloud. Thus, with the distributed fog nodes technology, a smart city energy management can be easily handled by this platform.

## REFERENCES

[1] Singh, S. and Yassine, A. Mining energy consumption behaviour patterns for households in smart grid. *IEEE Transactions on Emerging Topics in Computing*. Vol. 7, No. 3, 2019, pp. 404-419, doi: 10.1109/TETC.2017.2692098.

[2] Islam, I., Rojek, L., Hartmann, M., Goran Rafajlovski. Artificial Intelligence in renewable energy systems based on smart energy house. *International Journal on Information Technologies and Security*, ISSN: 1313-8251, Vol. 12, No.4, 2020, pp. 3-12.

[3] Yassine, A., Singh, S., Hossain, M. S. and Muhammad, G. IoT big data analytics for smart homes with fog and cloud computing. *Future Generation Computer Systems*, ISSN: 0167-739X, Vol. 91, 2019, pp. 563-573.

[4] Makonin, S., Ellert, B., Bajic, I. V. and Popowich, F. Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014. Scientific Data, 2016, doi:10.1038/sdata.2016.37.

[5] Makonin, S. AMPds2: The Almanac of Minutely Power dataset (ver. 2), 2019, doi:10.7910/DVN/FIE0S4, Available at: https://dataverse.harvard.edu/dataset.xhtml

[6] Zou, K. H., Tuncali, K. and Silverman, S. G. Correlation and simple linear. *Radiology*, Vol. 227, No. 3, 2003, pp. 617-628

[7] Ali, U., Rauf, A., Iqbal, U., Shoukat, I.A., Abu ul Hassan. Big data analytics for a novel electrical load forecasting technique. *International Journal on Information Technologies and Security*, ISSN: 1313-8251, Vol. 11, No. 3, 2019, pp. 33-40.

[8] Khan, T., Tian, W.H., Ilager, S., Buyya, R., Workload forecasting and energy state estimation in cloud data centres: ML-centric approach. *Future Generation Computer Systems*. Vol. 128, 2022, pp. 320-332.

### *Information about the authors:*

**Sofiene Haboubi** – is Assistant professor of Computer Science at Higher Institute of Medical Technologies, where he has been since 2005. He received a B.S. from Faculty of Sciences of Tunis El Manar University in 2003, and an M.S. from the National Engineering School of Tunis in 2005. He received his Ph.D. from the University of Tunis El Manar in 2011. Since 2004 he worked at Signal, Images, and Information Technologies Research Laboratory. Her research interest includes: Pattern recognition, Optical character recognition, document processing, Medical informatics, Artificial Intelligence, Satellite image processing.

**Oussama Ben Salem** – He obtained his master’s degree in Technologies and Information Systems from National Engineering School of Tunis, Tunis El Manar University, Tunisia. Her research interests are in the fields of Data science and Machine learning.

**Manuscript received on 07 November 2021**