PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHMS IN ELECTRIC LOAD FORECASTING

Mohan Ch, AVS Pavan Kumar

[1]*M Tech Final Year Student,*[2] *Assistant Professor*
Department of Computer Science and Engineering
BABA Institute of Technology And Sciences,Vishakapatnam

ABSTRACT:--*Nowadays, we have data in abundant from numerous sources. Retrieving useful information from this data is very tedious task. Hence, to address the complex nature of various real world data problems, specialized machine learning algorithms have been developed that solve these problems perfectly. Machine learning is a type of artificial intelligence that allows software applications to become more accurate in predicting outcomes without being explicitly programmed and has gained much popularity in recent years. This paper explores evaluation performance of Naïve Bayes, ID3, and Random Forest on electricity consumption or load datasets. Naïve Bayes algorithm is depending upon likelihood and probability; it is fast and stable to data changes. In decision tree learning, ID3 (Iterative Dichotomiser 3) algorithm is used to generate a decision tree from a dataset and is typically used in the machine learning and natural language processing domains. Random forest algorithm is an ensemble algorithm that fits multiple trees with subset of data and averages tree result to improve performance and control over-fitting. This paper concludes with a comparative evaluation of Naïve Bayes, ID3 and Random Forest in the context of Electric Load Dataset in order to forecast the electricity load or consumption.*

Keywords: Classification, Load Forecasting, Naïve Bayes, Random Forest, ID3, Machine Learning, Performance

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information [1][2]. Data mining software allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Specifically, data mining tool includes Time series, Clustering, Neural Network, Fuzzy Logic, etc. [3]- [5].

Electric load forecasting predicts the load demand in time intervals of one day to several weeks. It plays an important role in the operation of power systems and has been a benefit to several controls, such as energy transactions, security analysis, economic dispatch, hydro-thermal coordination, load management, and generator maintenance scheduling.

There are different methods to deal with load forecasting. Traditional short-term load forecasting can be classified in general regression models or time series models or artificial neural networks. [6]- [11] Before forecast, it is necessary for load data to be categorized from their data pool. Normally, most of the load data follow their day-type, such as this Saturday, nearly same as last Saturday. However, if this Tuesday is the day after a holiday, its load data will not follow last Tuesday but rather last Monday. If forecast is performed while containing such improperly categorized load data, such could lead to incorrect forecast results. Consequently, such must be conducted previously in order to conduct a correct forecast. Here in this paper, we are predicting the load consumption using simple but powerful algorithms like Naïve Bayes, IID3 and Random Forest.

Table 1: Datasets used with different load instances for 24 hours during the week of two different Companies

Hour/Date	01-01-17	02-01-17	03-01-17	04-01-17	05-01-17	06-01-17	07-01-17
1	0.4115	0.4115	0.0000	0.4268	0.4268	0.0000	0
2	0.4115	0.4115	0.0000	0.4268	0.4268	0.0000	0
3	0.4115	0.4115	0.0000	0.4268	0.4268	0.0000	0
4	0.4115	0.4268	0.0000	0.4268	0	0	0
5	0.4115	0.4115	0.0000	0.4115	0	0	0
6	0.4115	0.4115	0.0000	0.4115	0	0	0
7	0.4115	0.3963	0.0000	0.3963	0	0	0
8	0.4115	0.4115	0.0000	0.4115	0	0	0
9	0.4268	0.4115	0.0000	0.4268	0	0	0
10	0.4268	0.4115	0.0000	0.4115	0	0	0
11	0.4115	0.4268	0.0000	0.4268	0	0	0
12	0.4268	0.4268	0.0000	0.4268	0	0	0
13	0.4268	0.442	0.000	0.4268	0	0	0
14	0.4268	0.4268	0.0000	0.4268	0	0	0
15	0.4268	0.4268	0.0000	0.4268	0	0	0
16	0.4268	0.4268	0.0000	0.4268	0	0	0
17	0.4268	0.4268	0.0000	0.4268	0	0	0
18	0.4268	0.4115	0.0000	0.4268	0	0	0
19	0.4268	0.4268	0.0000	0.4268	0	0	0
20	0.4268	0.4268	0.0000	0.4268	0	0	0
21	0.4268	0.442	0.000	0.442	0	0	0
22	0.4268	0	0	0.4268	0	0	0
23	0.4268	0	0	0.4268	0	0	0
24	0.4115	0	0	0.4268	0	0	0
Max	**0.4268**	**0.442**	**0.000**	**0.442**	**0.4268**	**0.0000**	**0**
Min	**0.4115**	**0.3963**	**0.0000**	**0.3963**	**0.4268**	**0.0000**	**0**
Avg	**0.4204**	**0.4202**	**0.0000**	**0.4236**	**0.4268**	**0.0000**	**0**
1	0.788	0.7164	0	0.3109	0.2328	0	0
2	0.7522	0.6447	0	0.3109	0.2149	0	0
3	0.6805	0.6447	0	0.3109	0.2241	0	0.2896
4	0.6805	0.5373	0	0.3326	0.3224	0	0.2896
5	0.6805	0.5373	0	0.3292	0.3734	0	0.2896
6	0.7522	0.5194	0	0.3142	0.4294	0	0.362
7	0.8059	0.4656	0	0.3109	0.4481	0	0.3982
8	0.1433	0.3582	0.2774	0.2987	0.4298	0	0.4706
9	1.0208	0.6805	0.2774	0	0.4835	0.4161	0.9231
10	0	1.0029	0	0.7468	0.9492	0	0
11	0	1.0349	0.4477	0.9896	1.251	0	1.0749
12	1.1283	0	0.5373	0.9492	1.2136	0	1.1443
13	1.182	0.4294	0.9313	0.9313	1.046	0	1.1616

14	1.2357	0.4481	0.9313	0.9134	1.2012	0	1.1963
15	1.2056	0	0.9671	0.8775	1.2696	0	1.2309
16	1.2715	0.4668	1.0387	0	0.9896	0	1.283
17	0.8954	0.439	1.0924	0.9134	1.0642	0	1.2136
18	0.7343	0	0.9877	0	0.3814	0.5415	0.3294
19	0.7343	0.9709	0.9694	0.2801	0	0.3921	0.3294
20	0.7164	0	0.9328	0.2614	0	0.3696	0.3294
21	0.9134	0	0.8029	0.2614	0	0.3547	0.2947
22	0.8238	0	0.3547	0.2614	0	0.3326	0.2947
23	0.788	0.7975	0.3292	0.2987	0	0.3142	0
24	0.7628	0.7343	0.7802	0.3109	0.2427	0	0.2896
Max	**1.2715**	**1.0349**	**1.0924**	**0.9896**	**1.2696**	**0.5415**	**1.283**
Min	**0.1433**	**0.3582**	**0.2774**	**0.2614**	**0.2149**	**0.3142**	**0.2896**
Avg	**0.7074**	**0.69655**	**0.6849**	**0.6255**	**0.7422**	**0.4278**	**0.7863**

The above datasets also shows the minimum, maximum and average consumption of electricity during the weekdays in two different companies considered.

II. LOAD FORECASTING

Initially, to forecast load data, it is necessary to classify the load data. The characteristic of each day of load is a different feature. Mostly, the day from Tuesday to Friday have nearly the same characteristic. But Weekends and Monday have their distinctive characteristic. For example, Monday has relatively low morning load compare to week day one and Saturday has low afternoon load. Figure shows the load pattern during a week in relation to holidays that may not be always applicable. If a regular holiday is Thursday, some companies may give a holiday to Employees. Hence, that Friday has low load as a whole compare to the normal Friday load. Through analysis, these distinctive data was sorted out.
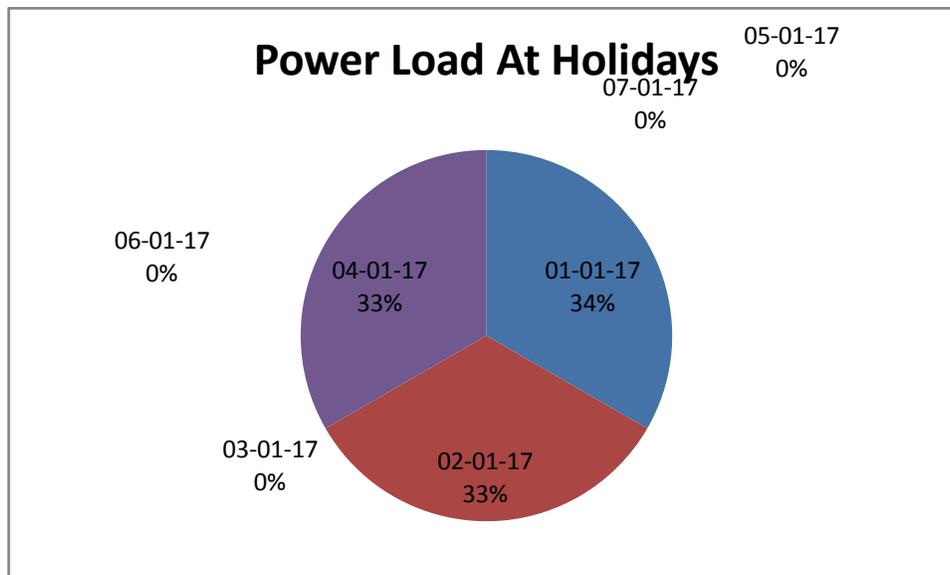
Figure1: Pattern of Load Data

II.   LOAD FORECASTING ALGORITHMS

In this work, Load Forecasting is done using, Naïve Bayes, ID3 and Random Forest and their performance will be evaluated [12].

A.   Naïve Bayes

The naïve Bayes algorithm is a simple probabilistic classifier that calculates a collection of probabilities by investigating frequency and combination of values in a given data set. The algorithm is based on applying Bayes theorem with the "naïve" assumption of independence between every pair of features.

Due to simple structure of Naive Bayes, construction of it is very simple and also has several advantages. Moreover, the inference (classification) is achieved in a linear time (while the inference in Bayes networks with a general structure is known to be NP-complete). Also, it does not require much larger data set smaller data set can also be used. Finally, the construction of naive Bayes is incremental, in the sense that it can be easily updated (namely, it is always easy to consider and take into account new cases in hand). [8][9]

B.   ID3

ID3, a classification algorithm builds a decision tree from a fixed set of examples to classify future samples. The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch being a possible value of the attribute. ID3 uses information gain to help it decide which attribute goes into a decision node. The advantage of learning a decision tree is that a program, rather than a knowledge engineer, elicits knowledge from an expert. [13]

ID3 improves on CLS by adding a feature selection heuristic. ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; otherwise it recursively operates on the number of possible values of an attribute partitioned subsets to get their "best" attribute. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices. ID3 is a non incremental algorithm that derives its classes from a fixed set of training instances.

C.   Random Forest

A random forest is a collection of randomized decision trees. Each decision tree in the forest is learned from a random subset of training examples and a random subset of features. To classify a test example, the outputs from each decision tree are averaged to determine the overall output. Specifically, each tree is traversed until reaching a leaf node. A probability score is assigned according to the ratio of training examples of each activity type that belong to the leaf node. These probability scores are averaged over each tree in the forest to obtain an overall probability score for the example. Finally, the activity type with highest probability is predicted for that example [14].

III.   CASE STUDY

**Preparing Data Set and Application**

Classification was performed using electricity load consumption datasets of two different companies. Data sets with load consumption during the initial week in the year 2017 of different organizations were considered as the training sets with around 1500 records to predict the average load consumption in future. Special days, such as a National holidays and employee's day, summer holidays etc, were eliminated to avoid interfering with ordinary day classification.

Results and Comparison Graphs

Three supervised machine learning algorithms, Naïve Bayes, ID3 and Random Forest were applied to predict the load consumption during the week in different organizations upon 1500 records. It was shown that accuracy of Naïve Bayes algorithm in predicting the load consumption is high in terms of Time and Space complexity compared to other two classifiers.

1/01/2017 Work Load of Power is High

Work Load 02/01/2017
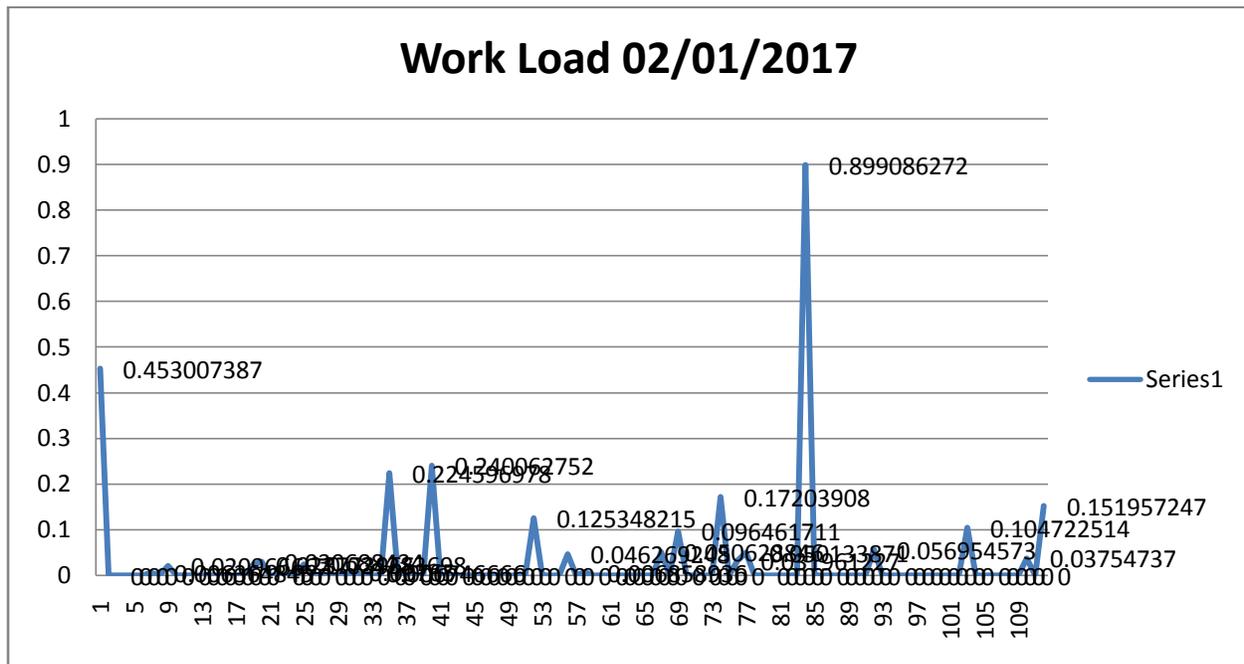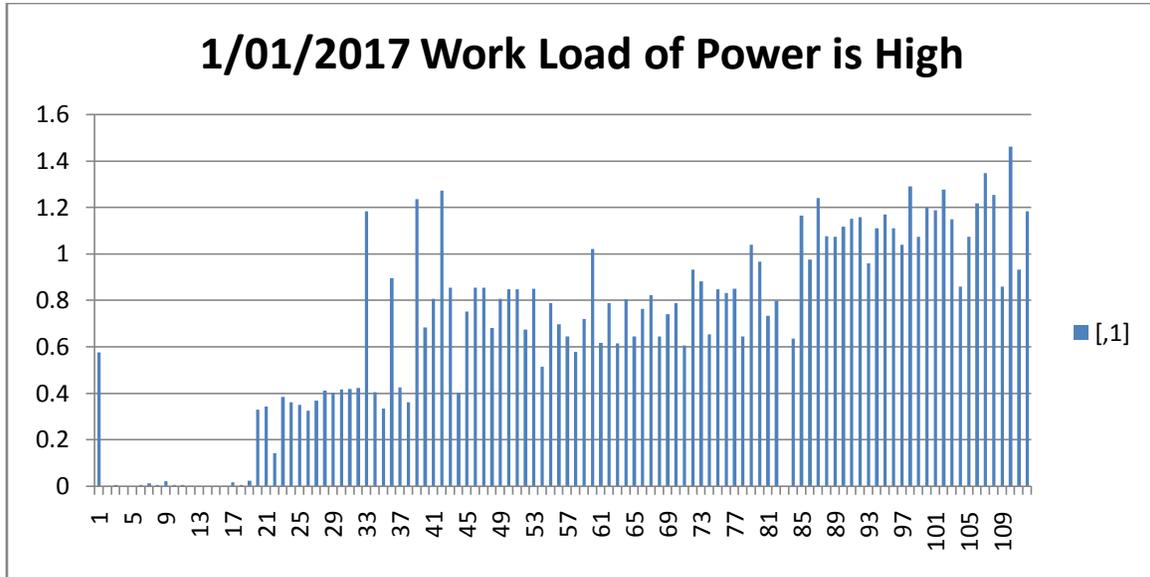
Algorithms	No of Records Processed	Time Complexity (ms)
Naive Bayesian	1500	22589
ID3	1500	38922
Random Forest	1500	26555
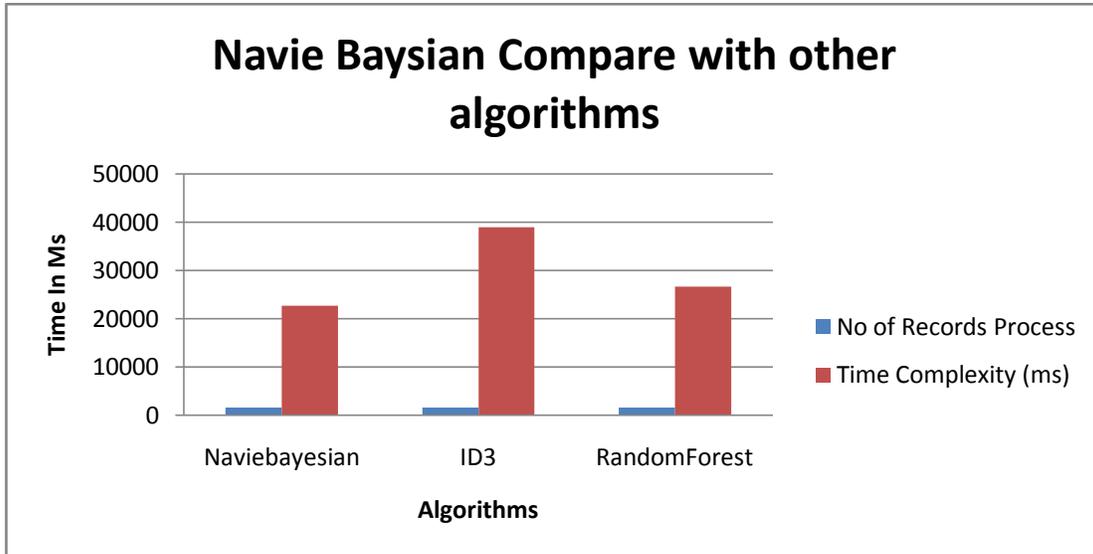
Graph Shows the Time Complexity of different algorithms used to evaluate performance

Algorithms	No of Records Processed	Space Complexity (MB)
Naive Bayesian	1500	18256
ID3	1500	24568
Random Forest	1500	19652
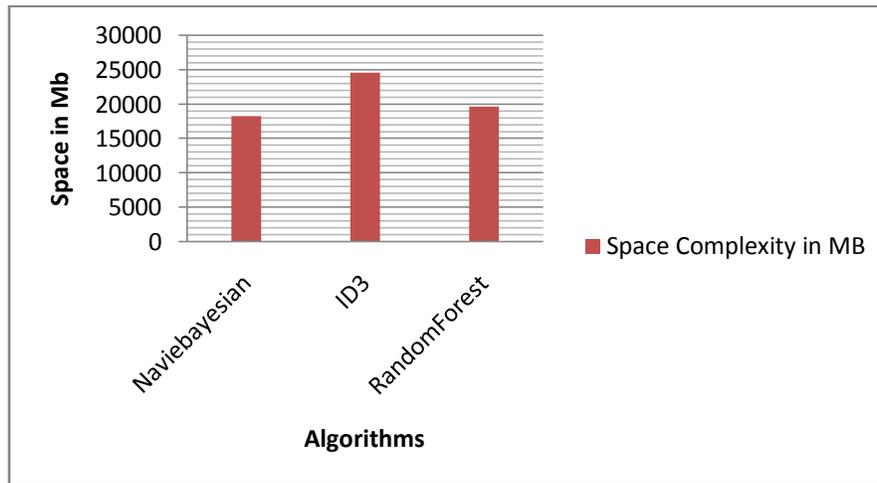
Graph Shows the Space Complexity of different algorithms used to evaluate performance

Comparing the accuracy of the three models Naïve Bayes, ID3 and Random Forest, Naïve Bayes shows that the prediction using this algorithm is more accurate in terms of space and time complexities. It is concluded that Naïve Bayes performs better than the other classifiers such as ID3 and Random Forest model. Though the differences in results of these three models are minor, it is true that the Naïve Bayes classification model performs better and accurate than the ID3 and Random Forest classifiers for short-term load forecasting.

IV.   CONCLUSION

This paper suggested load classification by Naïve Bayes, ID3 and Random Forest and compared three algorithms in load forecasting. As preprocess of load forecasting, classification was successfully completed. By comparing these forecasting models, Naïve Bayes model shows superior performance in total prediction. The accurate prediction also sometimes depend on the other factors like weather, temperature, event of the day etc, that affects the forecasting accuracy.

V.   REFERENCES

[1] Koo, B. G., Kim, M. S., Kim, K. H., Lee, H. T., Park, J. H., & Kim, C. H. (2013, January). Short-term electric load forecasting using data mining technique. In *Intelligent Systems and Control (ISCO), 2013 7th International Conference on* (pp. 153-157). IEEE.
[2] Bill palace, "Data mining technology note", UCLA Anderson School of management, less angles, spring 1996.
[3] Paolo Giudici. Applied Data Mining-Statistical Methods for Business and Industry. Italy: John Wiley and Sons, Inc.
[4] Mehmed Kantardzic. Data Mining-Concepts, Models, Methods and Algorithms. USA: John Wiley and Sons, Inc.
[5] Hiroyuki Mori, Noriyuki Kosemura, Toru Kondo, Kazuyuki Numa (2002). Data Mining for Short-term Load Forecasting. IEEE.
[6] Dae-Yong Kim, Chan-Joo Lee, Yun-Won Jeong, JongBae Park, Joong-Rin Shin, Development of System Marginal Price Forecasting Method Using SARIMA Model, Proceeding of conference KIEE Nov. 2005, pp.148-150.
[7] Mian Xing, Wei Sun, "Data mining based neural network model for load forecasting", Machine learning and Cybernetics 2005, Proceedings of 2005 International Conference on Volwne 3, 18-21 Aug. 2005, pp. 1713-1715.
[8] Heon Gyu Lee, Bum Ju Lee, Jin-ho Shin, Long Jin, Cheng Hao Jin, Keun Ho Ryu, "Application of calendar based temporal classification to forecast customer load patterns from load demand data", Computer and Short-Term Electric Load Forecasting Using Data Mining Technique 157 Information Technology 2008, 8-11 July 2008, pp.149-154.
[9] K.Y. Lee, Y. T. Chao J. H. Park, "Short-Term Load Forecasting using an Artificial Neural Network", Transaction on Power System vol.7, No l, February 1992.
[10] Hyeonjoong Y 00, Russel L. Pimmel, "Short-Term Load Forecasting using a Self-Supervised Adaptive Neural Network", IEEE Transaction on Power System, Vol. 14, No. 2, May 1999.
[11] G. Juberias, R. Yunta, J. Garcia, Moreno C. Mendivil, "A New ARIMA Model for Hourly Load Forecasting", Transmission and distribution Conference, 1999 IEEE Vol. 1, 11-16 April 1999, pp. 314-319, Vol. l.
[12] XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda , Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, ZhiHua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg, "Top 10 algorithms in data mining", KnowlInf Syst, 2008 14, pp. 1-37.
[13] "Building Decision Trees with the ID3 Algorithm", by: Andrew Colin, Dr. Dobbs Journal, June 1996
[14] Ellis, K., Kerr, J., Godbole, S., Lanckriet, G., Wing, & Marshall, S. (2014). A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiological measurement, 35(11), 2191.