Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: 0005-1144 (Print) 1848-3380 (Online) Journal homepage: https://www.tandfonline.com/loi/taut20

Semantic web service-based messaging framework for prediction of fitness data using Hadoop distributed file system

R. Sethuraman & T Sasiprabha

To cite this article: R. Sethuraman & T Sasiprabha (2019) Semantic web service-based messaging framework for prediction of fitness data using Hadoop distributed file system, Automatika, 60:3, 349-359, DOI: 10.1080/00051144.2019.1637175

To link to this article: https://doi.org/10.1080/00051144.2019.1637175

Published online: 03 Jul 2019.

AUTOMATIKA
2019, VOL. 60, NO. 3, 349–359
https://doi.org/10.1080/00051144.2019.1637175

Taylor & Francis
Taylor & Francis Group

SPECIAL ISSUE: COMPUTATIONAL INTELLIGENCE AND CAPSULE NETWORKS

OPEN ACCESS

Check for updates

Semantic web service-based messaging framework for prediction of fitness data using Hadoop distributed file system

R. Sethuraman and T Sasiprabha

School of Computing, Sathyabama Institute of Science and Technology, Chennai, India

ABSTRACT

Big data is coined as word of mouth in this era due to the generation of huge volume of data every second from multiple sources like logs, web sources, and sensors, electrical and electronic devices. The manipulation is performed over the injected data and is termed as Data Processing Segment. In this proposed paper the data are obtained from the wearable devices with attributes like calories, weight, fat, step count, sleep, BMI and so on. The obtained data is stored in HDFS in a persistence manner. The component Kafka acts as a queue for the real time data and regulates the data before storing in HDFS. Now Apache Spark does the streaming of data. Here the data are cleaned, applied the Machine Learning Algorithms (KNN Classifier) to obtain the model, by splitting the cleaned data into Training data and Testing Data. Now the obtained predicted result is sent to Web service Telephony ontology, which in turns communicates with ontology service repository consisting of cloud telephony services ontology and fitness activity ontology through OWL API. The classified and predicted value is analysed and intimated to the users through visualization graphs, SMS, IVR and e-mail.

ARTICLE HISTORY
Received 31 January 2019
Accepted 29 April 2019

KEYWORDS
Wearable device; big data; HDFS; spark; prediction; semantic web service ontology

1. Introduction

Advancements and Innovations in the field of electronic and wireless communications resulted in the use of wearable devices. Mostly these devices are in compact size and easy to carry. They can be easily mounted in human body mostly worn in the wrist. These wearable's [1] track the fitness activities of the humans like walking, jogging, steps count, weight, sleep time, calories, BMI and so on. These wearable along with their Application Programming Interface (API) tracks and stores the data of these activities and is stored in the HDFS. This helps the people to ensure the fitness of the individuals. Based on the final outcome of this data the fitness activities of the individuals are viewed and improvements are made accordingly. The recordings of the data includes distance covered, number of floors moved, elevation carried out, calories, sedentary minutes, Lightly active minutes, Fairly active minutes, very active minutes, sleep time, weight, BMI, fat and calories [2]. These data received continuously from the wearable's are crawled and stored in HDFS.

Big data refers to huge volume of data which grows continuously in real time systems. These data support and fall into any one of the knowledge system commonly known as 10 V's of Big Data, which includes volume, velocity, variety, value, veracity, validity, variability, viscosity, and visualization. The activities on Big data depends on the need of the application. The major classification is Data Collection, Data Ingestion, Data Processing, Data Storage and Data Visualization. Various frameworks and components like Flume, Nifi, Kafka, Sqoop, Flink, Spark, and Storm are used for performing corresponding activities. After data ingestion, processing of these data plays a vital role which determines the final prediction and provides classification for the ingested data.

Apache Spark is used for processing and cleaning the data since it is designed in such a way to support the computational speed of the data. Spark is advanced cluster computing engine than Hadoop Map Reduce because Map Reduce processes structured and unstructured data available in clusters in batch mode only whereas the Spark does from multiple ranges like interactive, iterative, batch and streaming. Also it is easy to manage and do real time analysis. Spark process real time data using Spark Streaming.it supports cross platform and is developed in language Scala. Since Spark is an analytic tool, most of the data scientist prefers this for Analytics.

Data processing layer is the core activity in the big data Analytics. The cleaning of data takes place in this layer. The accuracy of the model is directly propositional to the accuracy in the cleaning data. After data cleaning, applied the Machine Learning Algorithms

CONTACT R. Sethuraman sethuramanr218@gmail.com

and Models to this data to obtain the predictive value. The machine learning algorithms are classified into three major categories as Supervised Learning Algorithms, Unsupervised and Reinforcement Algorithms. Each of these categories has their own functionality. The most preferred categories of Machine Learning Algorithm is Supervised Learning. In this category the input data is processed and obtained desired output. A Function is created to accept the input and processed till the desired output is obtained. The outcome of this algorithm yields to two categories as Regression and Classification. Most popular Regression categories are Decision Tree, Random Forest and KNN. Suitable Algorithm is applied and the model is inferred to evolve at the predictive results. Also consider few parameters before concluding the algorithm. Below is the table which suggests the selection of algorithm based on the parameters like Size, Accuracy and speed for Supervised Learning. Similarly unsupervised learning can also be classified with its parameter. This paper provides its elaboration with respect to Supervised Learning Algorithm. There are no restrictions with the selection of parameters. The entire process is carried out with the training and testing data, thus obtaining the Predicted value. After data processing the flow control goes to on demand cloud Telephony Web Services.

Web service Registry is an intermediate repository that connects cloud telephony web services and ontology service Repository through SOAP/XML and OWL API respectively. Ontology is defined as set of concepts, descriptions, properties and relations. There are many tools available to create the ontology. This paper proposes two categories of Ontology namely Cloud telephony services and Fitness Activity ontology [3]. The cloud telephony ontology decides upon the selection of mode of communication of the predicted results like SMS, IVR, E-mail, Mobile App and visualization. Fitness Activity ontology provides the data on the predicted value, which will be communicated through the HTTP Request and Response API to the end user.

The key contributions in this paper are as follows:

- Data obtained from various sources are ingested on HDFS available in the cloud Environment, which is enhanced by the component Apache Kafka shown in Table 1.
- Data Analytics is done in Data processing layer by Speed analytical tool Apache Spark along with the Supervised Machine Learning Algorithm.
- KNN Predicting Model is applied which yields the maximum accuracy with other Models and predicted the values.
- Ontology knowledge representation of Cloud Telephony ontology and Fitness Activity Ontology for effective selection of modes of communication.
- On demand cloud Telephony web services to represent and implement the decisions of Ontology knowledge representation to the end users.

Table 1. Model suggestion for supervised category based on parameters.

Supervised type	Parameter	Suggested model
Classification	Accuracy	SVM, Neural Network, Gradient Boosting Tree, Random Forest
Classification	Speed and Elaborated	Decision Tree, Logistic Regression
Classification	Speed, Non-Elaborated and Large data	Naïve Bayes
Classification	Speed, Non-Elaborated and Moderate data	Linear SVM, Naïve Bayes
Regression	Accuracy	Neural Network, Gradient Boosting Tree, Random Forest
Regression	Speed	Decision Tree, Linear Regression

2. Fitbit data on big data ecosystem in cloud

Input data is obtained from the wearable devices which the customers will wear on their body mostly likely as a watch for tracking the daily activities towards their fitness. These data are injected in the Hadoop Distribution File System available in the Hadoop Ecosystem, which resides on the cloud Environment. These data are processed in Data Processing layer where the splitting of data into training data and testing data happens. These are then computed with multiple Machine learning Algorithms and predicted with the model inferred from the Algorithm. Based on the Algorithm and the Model predicted value is obtained in this paper. Figure 1 shows entire flow of data from wearable devices till the visualization and communication platforms. The novelty of this proposed paper is explained below in various sections. This paper provides you the analysis, prediction and appropriate communication channel selection for the real time data obtained from the wearable devices. The benchmark Data Set is available in web in the location https://zenodo.org/record/14996#.WfxTJFuCzZ4

2.1. Data ingestion

In the concept of big data, the sources of data are multiple and the data to be handled are in surplus volume. Generally the data are obtained from sensor devices, production logs, Events, sources of web like Facebook, Twitter, and Linked In and so on. The selection of data source depends on the type of analysis and prediction to be carried out. This paper performs the Analysis and prediction for the data obtained from wearable devices like Fit Bit. The available data are either in structured, Unstructured or Semi structured formats like in HTML, Pdf, Image and Text. These types of data

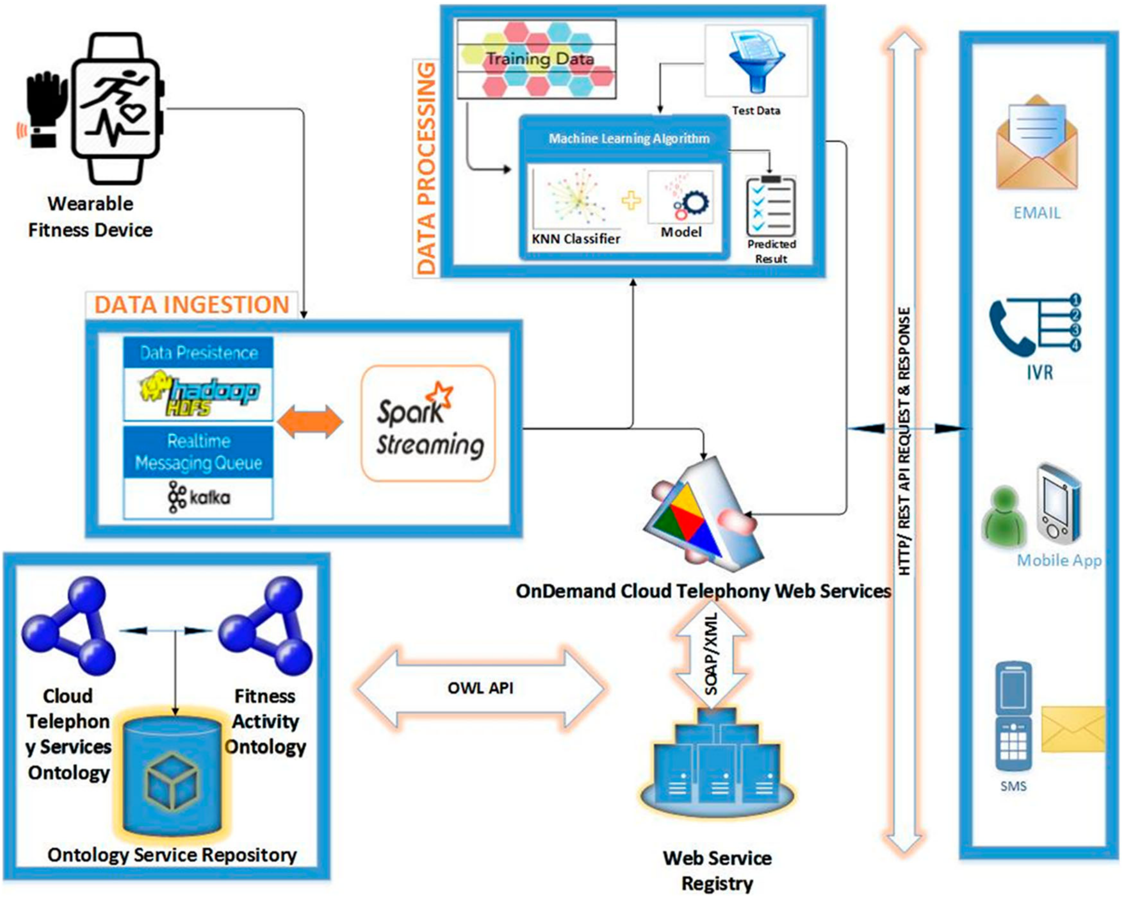

Figure 1. Overall architecture.

are crawled using the crawler service like web crawler which is built using the Framework named Akka. This framework supports HTTP and HTTPS protocols. It also supports proxy server for performing this activity. Here the data are obtained from the wearable devices and stored in the Hadoop Distributed File System. HDFS is a component available in the.

Hadoop ecosystem which resides in the form of a cluster in the cloud environment. The fit bit data is stored in a better way in HDFS with the implementation of a light-weight library known as Kafka. This library is capable of integrating with any application and is meant for perfect and efficient storage of incoming data over HDFS. Kafka is a Queuing system which maintains all the input data in a pipeline and sends the data in a Queue to store on HDFS. The stored data is sent to Apache Spark Streaming which performs the process of data cleaning. Spark streaming receives real time events from Kafka. This is used in the Data Processing layer to fetch the data and then split into training and test data. In the Spark streaming live data streams are received from Kafka and are split into batches. These batches are processed by the Spark Engine and the final data processed are obtained in batches. The working flow of Spark Streaming is depicted below. Spark does not have its own file system. It works along with Hadoop making use of Hadoop Distributed File System(HDFS) in processing the streaming data into Batches of Input Data. this processing is made faster and efficient using various features of Spark like Speed, multi-language acceptance with interactive querying and Analytics. The work of Spark Engine is accepting the input batches and applying Resilient Distributed Datasets(RDD) transformations on it. RDD stores memory state as objects across.these objects are shared between the jobs. Here the intermediate results are stored in Distributed memory there by making the process faster and efficient shown in Figure 2.

2.2. Data processing

Batches of processed data received from Spark streaming are fed into the Data Processing layer. Here these data are split into two data sets namely Training data sets and testing data sets. The ratio of splitting the training data and testing data is 80:20. Now the Machine Learning Algorithms are acted upon these data and required information's are extracted. Supervised Learning Techniques are used for these scenarios of running the models. The selection of appropriate Machine Learning Algorithm is the heart of Data Analytics [4]. The selection of improper Algorithm for a scenario yields to the drastic variations in the outcome of the application Results. The selection of Supervised Machine Learning Algorithms is made in an accurate way by considering the factors like size of data,

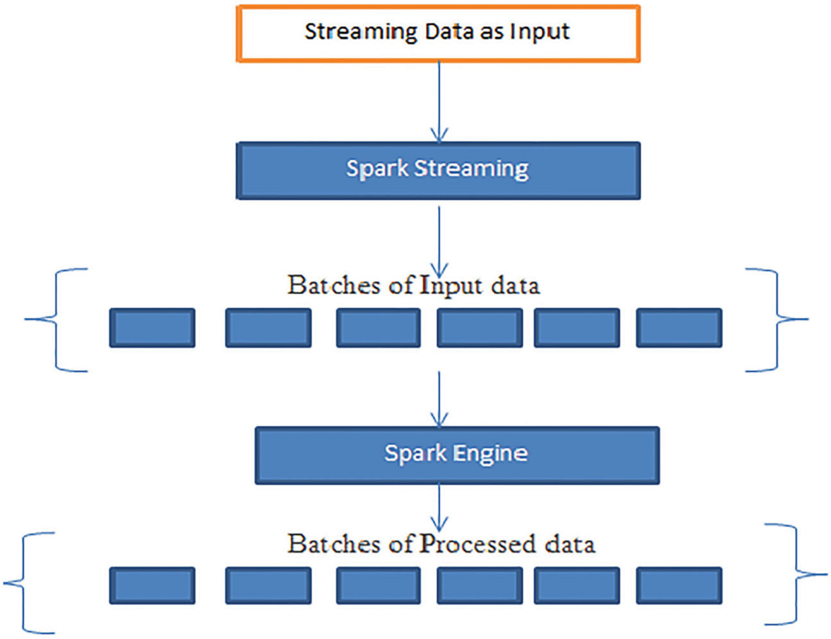

Figure 2. Spark streaming working flow.

data for processing is continuous or not, the expected output is of form classification or Regression, the predefined variables in the data set are labelled or unlabelled, the goal is to predict the output or to rank the output, what is the level of interpretation of the result – easy or hard and so on. Most of the practical machine learning implementations uses Supervised Learning Algorithms. The most commonly used MLL's are Regression, Logistic Regression, Decision Tree, Random Forest and KNN.

2.3. Predictive modelling using KNN

In this proposed paper, the classification technique K-nearest neighbours (KNN) Machine Learning Algorithm to predict the output for the considered input parameters. KNN handles both the classification and regression Problem Predictions. However the industry people use this Model for classification problems. The main reason to prefer KNN is its Predictive power is too high and very high Accuracy level is achieved. It also takes less time for computation [5]. Besides all it is very flexible with the data to interpret the output. In this model the key controller is the value of "K" selected to run this Model. With its simplicity it provides high accurate results. Rather to perform exact match with the stored data, it provides the way to recognize patterns of the data. Similar cases are close to each other and are known as neighbours, more the distance between the cases implies they are more dissimilar. The number of nearest neighbours is controlled by the "K" value. The main challenge in the predictive modelling is Training the data. The training of data is performed based on the following models.

2.3.1. Distance metric

This metric is used to measure the similarity of the cases and their nearest neighbours. The methods are available to calculate this Distance Metrics are Euclidean Distance and City Block Distance. In Cartesian coordinates,

If a = $(a_1,a_2,a_3, \ldots .a_n)$ and b = $(b_1,b_2,b_3 \ldots ..b_n)$ are two points, then the distance (dis) is calculated between a to b (or) b to a by the Pythagorean formula as in Eq.1 and the generalized formula is given in Eq.2

$$dis(a,b) = dis(b,a) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2} \quad (1)$$

$$dis(a,b) = dis(b,a) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \quad (2)$$

City Block Distance measures, the distance between cases is the sum, overall dimensions, of the weighted absolute differences between the values for the cases.

2.3.2. Crossvalidation for selection of K

Cross validation is an important data modelling technique in which only 80 percent of training data are trained and remaining 20 percent is not trained but is reserved for testing. Testing of the model with this 20 percent of the data before finalizing the model. Cross validation (CV) helps in improving the performance of the model. In cross validation for each value of k, where $K \in [k_{\min}, k_{\max}]$, calculate the average error rate or sum of square error of k, denoted by e_v. The equation for cross validation is given in Eq.3. In Cross validation, the parameter k is selected based on the number of parts the data is split as train and test. In 10 Cross validation, training is done with 9 parts and tested with 1 part. this is repeated for all combinations of train and test splits. The accuracy of the model is evaluated as ratio of number of correctly predicted instances to the total number of instances in the dataset. This is used while building and evaluating each model. The CV method is great because it measures directly the parameter which are about to measure, mostly the predictive performance on unseen data. For the more complex models it is used to select the tuning parameters. In a data set with n observations, n fold CV is used and is known as leave-one-out cross validation (LOOCV).

$$\mathrm{CVk} = \sum_{v=1}^{V} \frac{e_v}{V} \tag{3}$$

2.3.3. Feature selection

Feature selection is derived from feature engineering. Here more information is extracted from the existing data. Good hypothesis gives good results. Future selection is the process of finding the best subset of attributes, which better explains the relation of independent variable with the target variable. Feature selection is based on forward selection approach which is implemented for the features entered into the model. Other features are entered in a sequential manner. At each step, the selected feature reduces the error rate or sum of squares errors. Feature selection makes the ML Algorithms to train faster by reducing the complexity of the model and makes it easy to interpret. Based on the appropriate selection of subset accuracy of the model is attained. Pearson's correlation, a measure for quantifying two continuous variable X and Y in a Linear Manner. The range is between ± 1. The Pearson equation is described in Eq.4 as PA,B. The Pearson's correlation coefficient is the ratio of covariance of the two variables X and Y to the product of their standard deviations, where X and Y are random variables. The equation is represented by letter ρ.

$$\mathrm{PA,B} = \frac{\mathrm{cov(A,b)}}{\sigma_A \sigma_B} \tag{4}$$

The three classes of feature selection are filter methods, wrapper methods and embedded methods. Few examples of filter methods are Chi Squared Test and information gain. Wrapper methods are used in algorithms like hill climbing and forward – backward pass. The embedded method does the optimization of the Algorithm like regression. Feature selection is the method of selecting the most appropriate parameters available in the dataset for prediction of target variables and building better model. FS trains the model faster and reduces the model complexity thereby easy to interpret. It also provides the wide possibilities of selecting the subsets. As the appropriate model is selected, high accuracy is attained. The flow starts with set of all features to selection of best subset and then building best model to attain the accuracy.

3. Ondemand cloud telephony web services

After the prediction of results in data processing layer using the ML Algorithms and Models, the result is sent to the Telephony Web services which reside in the Cloud Environment. When the results are recommended by the customers it is given to them through this on demand telephony services. Cloud Telephony (CT) is a communication technology in which the hosting takes place at the service provider's premises. The application does not need any software or hardware for the activation and use of this service. There are two networks available to support and carry over these services namely IP and PSTN. The services through IP enable the delivery of telephony services via Internet. VoIP enabled devices provides these services. The availability of services without Internet is PSTN. Cloud Telephony Services provides the scalable solutions. This acts as a complete solution to track the customers. The efficiency of cloud telephony has the following characteristics for better services. These services includes on demand self-service, Broad Network Access, Resource Pooling, Rapid Elasticity and Measured Service. The on demand cloud telephony service includes Missed call, SMS, (IVR). These services are communicated with the web service Registry through SOAP/XML [6]. All the web services are available in this repository. This communicates with the ontology services and desired output is obtained. This cloud Telephony system communicates with the ontology services using OWL API. The Business logics written in Web services has multiple methods with common insights. These insights are delivered as services to the customers based on their Demand request. The unique parameter is User Insights in all services. The other parameter varies based on the services as Mobile number in IVR/SMS, Email Id in Mail Communication and Application id in Mobile App.

4. Ontology knowledge representation

The knowledge representation is the next model of active after cloud telephony web services. The representation of the obtained knowledge is done by Ontology service Repository. Two ontologies are available in this repository namely cloud telephony services ontology and Fitness Activity ontology. The fitness activity ontology maintains the predictive result about calories burnt based on various attributes of comparisons classified in its ontology structure. The modes of communication like SMS, e-mail, IVR, mobile app and visualization are stored in telephony services ontology. Based on

Table 2. Metric values for the selected feature variables.

Variable	Min	Max	Mean	Correlation	Covariance	Standard deviation	Skewness
Step	0	34954	4012.656	0.644	998413.594	4300.212	1.334
Calories	985	6092	2807.43	1	130026.142	360.591	2.679
Weight	80.309	97.523	89.267	0.157	252.444	4.462	−0.35
BMI	22.13	26.873	24.598	0.157	69.562	1.229	−0.35
Fat	14.892	24.443	20.93	0.158	138.964	2.441	−0.53

the user request the appropriate mode of communication is selected and intimated to the user [7]. This communication takes place via OWL API to the cloud telephony web services and is communicated via HTTP Request and Response to the end user. To the selected communication mode the details or analytics on the value of calories burnt against each of the attributes used in the data set is provided. Semantic service communication is established along with Ontology. Based on the selection of web services corresponding ontology is chosen. It helps to take right decision to select appropriate semantic-based services.

5. Related work

Big Data and the analysis done or prediction made from the available data is the trend that has been following in all the verticals that exists. The predicted results help them to improve their verticals to reach the maximum out of their business and progress. In the recent days it has also been done in Animal research [8] also. Here data collected from multiple sources has been maintained and processed with Machine Learning Models to identify the new species, behaviour and movements. It also provides better improved classification. The handling of data plays a vital role. The cloud environment helps us in handling and storing these multiple data in the cloud. This storage is supported and the functionality is enhanced by NoSql [9] databases. The better processing is supported here, since more comfortable the data for storage, then more better the data are taken and processed for efficient usage. This cloud supports in providing the good quality of service for the handling data. This involves the generation of data set for processing. The crucial point is the selection of model. Based on the model the evaluation is performed. The other trend is EHR [10], the Electronic Health Record. For the generation of EHR, the data are available from wearable devices and then stored for processing. The processed data are made available for the customers as a report which they can make available from their mobile devices. These reports are more like statistical form which the user can view in a comfortable way. They use Kafka framework for the ingestion of data from multiple devices as input. The data available from wearable and other health domains are of streaming data. These streaming data are processed by the Apache Spark. These reports help the doctors in monitoring the patient and Guides in the suggestion of Fitness activities, food habits and Disease Management for the patients. Irrespective of the data that are available, processed and applied with Machine Learning Models the security plays a vital role for the data [11]. Parameters like consumption of power on the wearable devices, transmission rate of the data and other security issues. To attain security, the devices are controlled by the corresponding mobile devices [12]. Upon the authentication from the devices the data are stored continuously and then processed. Similarly multiple analyses are performed and the prediction is made and used in an efficient manner.

6. Results and discussions

6.1. Data exploration

The process that defines the quality and accuracy of outcome is Data Exploration. It is the process of cleaning and making the data ready for the Analysis using the Model. The process starts with the identification of input and Target variable. In the proposed model the data input variables includes calories, steps, weight, bmi and fat. The target variable is the Calories burnt. The data received from Fit Bit device is a continuous variable, so the Outliers are very less when compared to categorical variables. The strength of the Relationship depends on the correlation value which lies between −1 and +1. The metrics like Correlation, Covariance, Standard Deviation and Skewness are calculated and given in Table 2. Metric Values are statistical values available in the core Dataset, obtained from the wearable FitBit device. From the available parameters, few parameters like Step, Calories, Weight, BMI and Fat are considered as Feature Variables and statistical calculations such as Min, Max, Mean, Correlation, Covariance, SD and Skewness are calculated. Missing values are treated by removing incomplete data and by calculating mean value for the particular Variable. The presence of outliers if any is also removed. Since data was available from realtime dataset, many inconsistent and missing value data was found. Using statistical analysis and real time Data Exploration using open source tools, the outliers and incorrect data was rectified.

6.2. Visualization charts

The input of data from the wearable device fit bit is measured with multiple parameters like calories, steps,

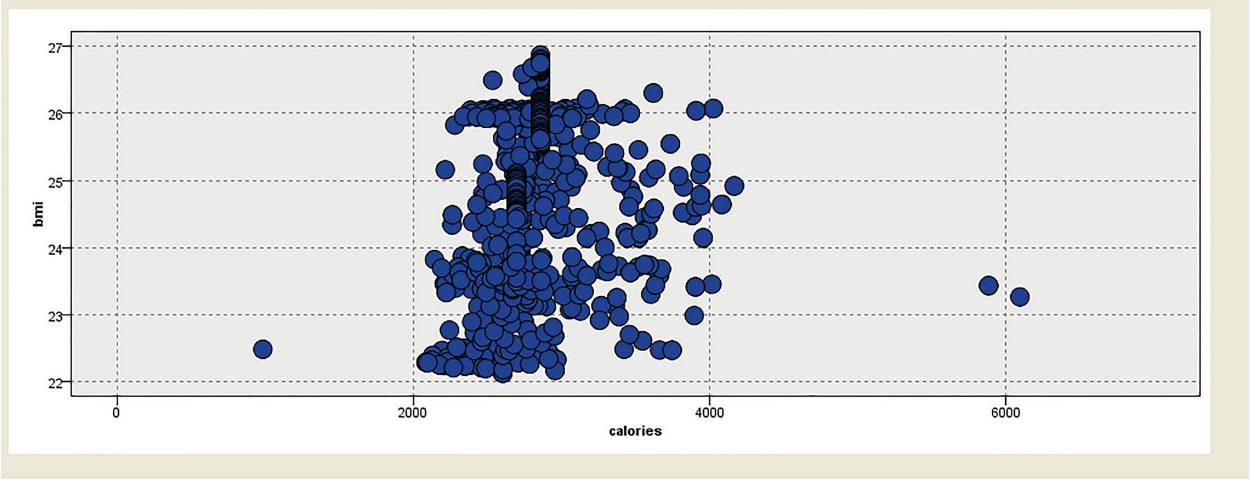

Figure 3. Visualization of calories against BMI.

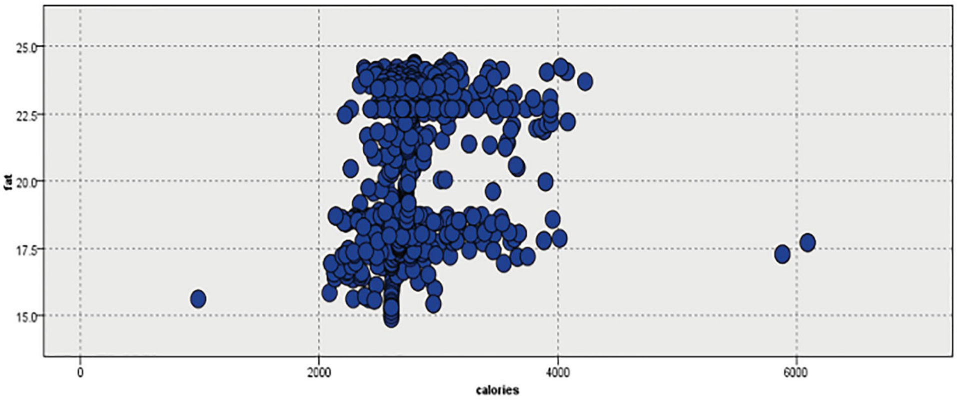

Figure 4. Visualization of input calories against the variable fat.

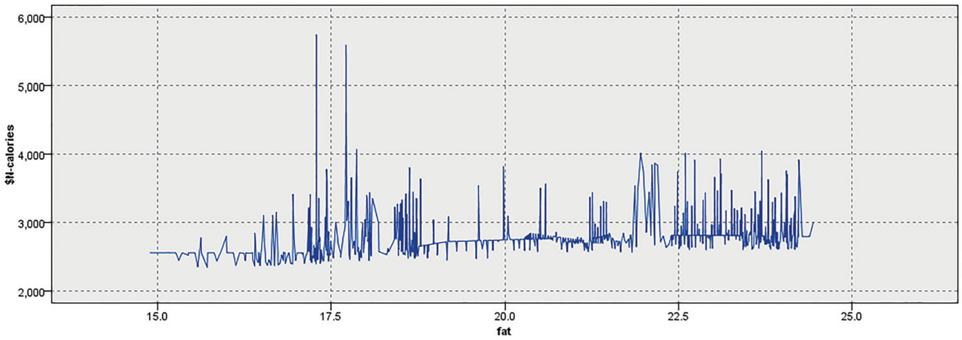

Figure 5. Predicted calories burnt against the variable fat using KNN.

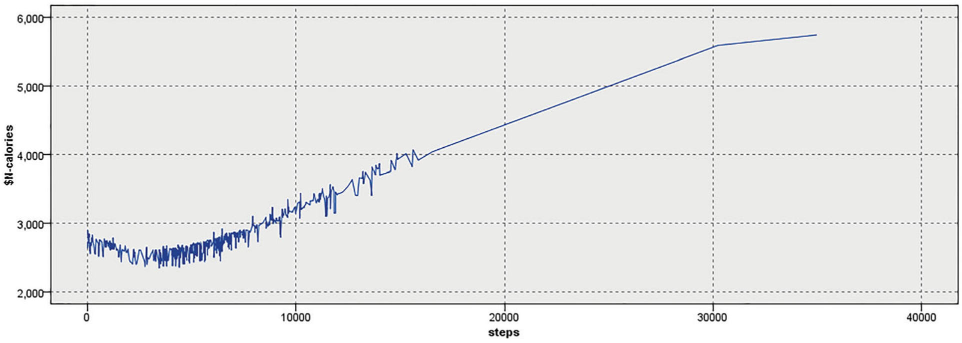

Figure 6. Predicted calories burnt against the variable steps using KNN.

weight, bmi, fat and so on. The data obtained is of continuous category because each and every second data is tracked and stored in the HDFS through Cloud API in the cloud Environment [13]. The data outliers are eliminated in the process of data Extraction and Data Ingestion using the library Kafka in our application. The quality of the data set is improvised by eliminating the outliers and Extremes. Since the measurement of data is continuous the probability of input missing is reduced. The amount of calories burnt against the recorded input of BMI value is depicted in a graph named Figure 3.

For the computation of data using any Machine Learning Algorithms, the data need to be available in the form of Training data and testing data. The algorithm is capable of taking all the features into account for the prediction. Doing so for the large data set leads to exhaust the computational power and the memory of the System. Selection of most appropriate feature is so mandatory. In the proposed paper the most appropriate features includes BMI, Fat and Steps for the prediction of calories burnt. The visualization of a selected feature fat against the calories is shown in Figure 4.

The most crucial step in Data processing is to suggest the next set of actions to be carried out after building an accurate and efficient model. In this proposal it refers to KNN. These set of actions are useful for enhancing the ratio of the goals. In this context, had the most important predictor calories burnt by the model. These are based on activities taken and analysed using data set. Figure 5 shows the plotting against the fat value and the KNN Model predicted calories burnt values [14]. This is based on the predicted value computed from training and testing data. Based on the hypothesis concluded, reports and information in graphics showcase the impact of these predictions. After obtaining the predicted value several input features are used to have new predictions of the same parameter. In this paper in all cases the amount of calories burnt is predicted based on the input features Fat, Steps and Weight. More comparison of predicted calories burnt against the other features leads to strong discussion and focus to have more practical conclusions.

This comparison yields the best of the model with more accuracy. This scenarios are tested, predicted calories burnt and visualized in Figures 6 and 7 as the feature variable Steps vs Calories burnt and

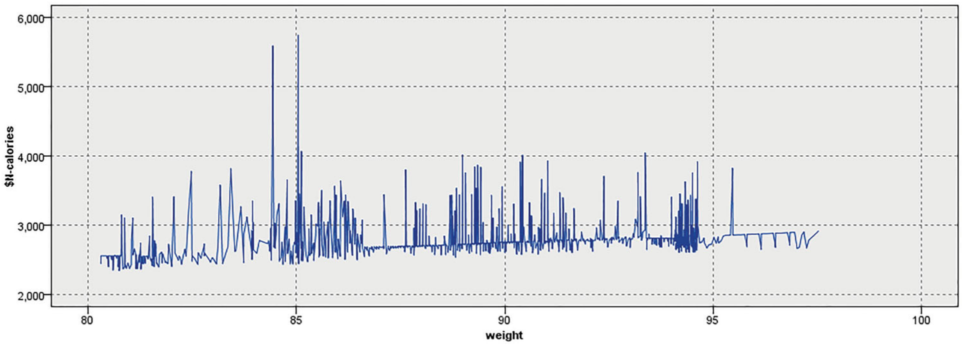

Figure 7. Predicted calories burnt against the variable weight using KNN.

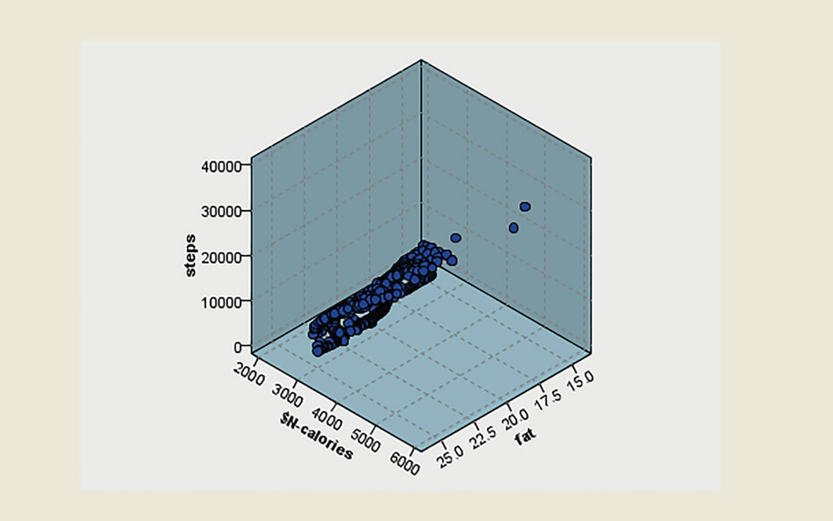

Figure 8. Three dimensional representation of prediction against calories, fat and steps.

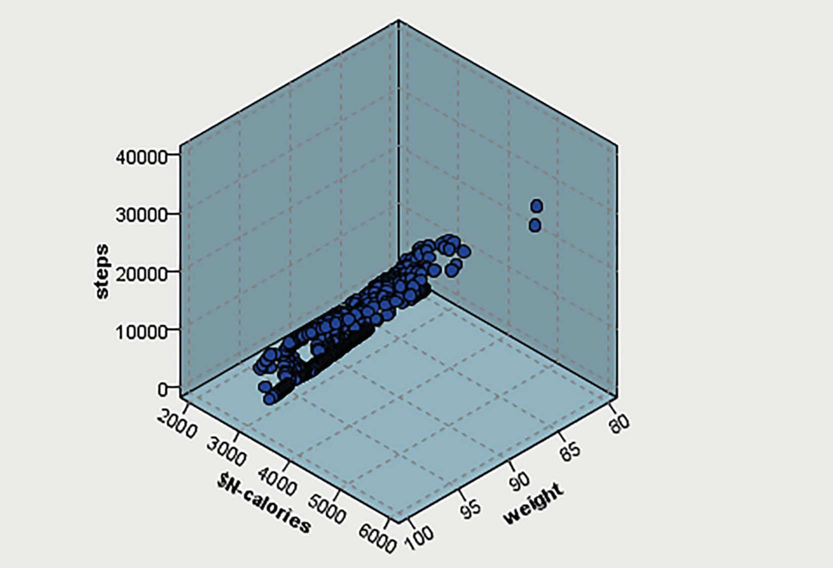

Figure 9. Three dimensional representation of prediction against calories steps and weight.

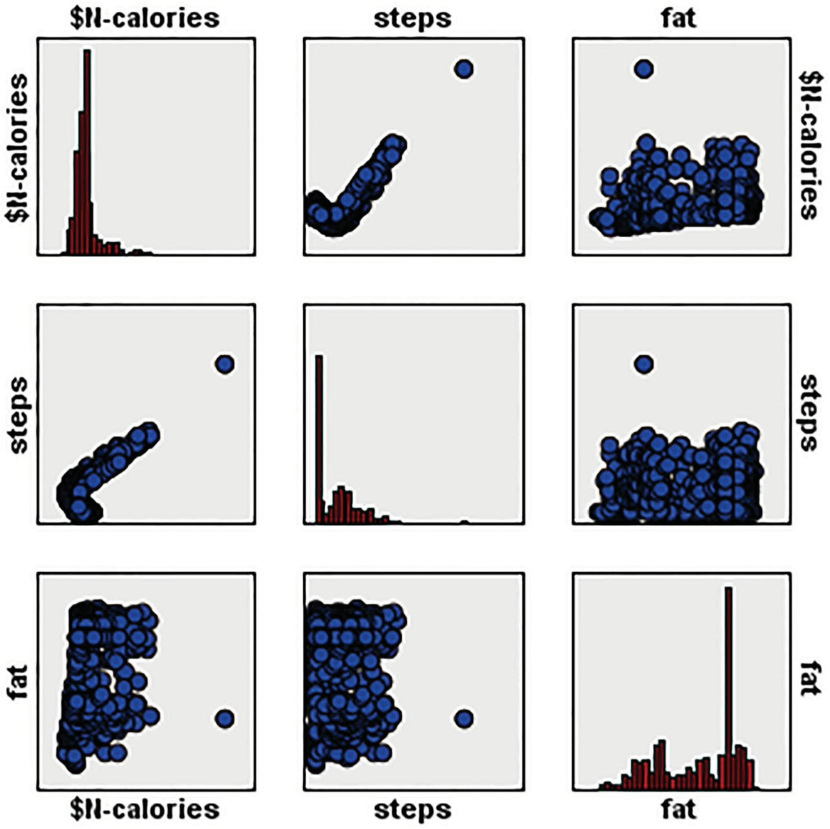

Figure 10. 3 cross 3 matrix representation of prediction against calories, steps and fat.

the feature variable weight vs Calories burnt respectively. From Figures 5–7, these visualized graph model has the Y-axis labelled as $N-calories. This notation stands for the predicted value obtained from the KNN Model.

In general Bar graph is the most commonly graph used irrespective of representing dependent and independent variables. Figure 8 represents the steps, predicted calories, and fat. From these plots the mappings are dragged and inferred if necessary. Based on these inferences conclusions are made as per the requirements. Other categories which are available are Scatter graph, Line graph and multiple line graphs. Similarly in Figure 9 the plotting of steps, weight and burnt calories are visualized.

To find out the linear correlation between multiple variables go for Scatterplot Matrices [15]. In Figures 10 and 11 multiple feature variables such as predicted calories, steps, weight and Fat are involved. Here each variable is plotted against each other.

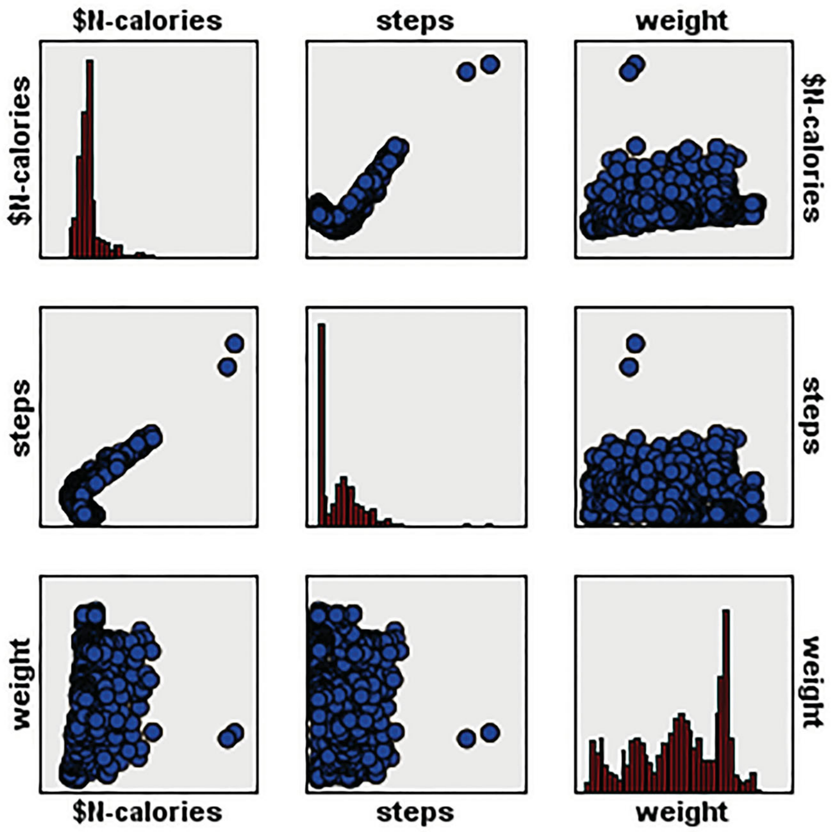

Figure 11. 3 cross 3 matrix representation of prediction against calories steps and weight.

Since multiple feature variables are involved in scatter plot matrix, the individual reference is defined in terms of variable X_i and Y_i for Vertical axis and Horizontal axis respectively.

From this scatterplot matrix the existence of pairwise relationship among the variables are concluded. The outliers among the data are easily figured out. This Scatterplot Matrix is also coined as SPLOM. The representation of histogram in the above figures is against the same variables. For better insights from the dataset, various visualization methods and techniques are used for multiple combinations of features available in the dataset. Multiple Feature variables like Input Calories vs Fat, Calories burnt vs Step Count and Calories burnt vs Weight are shown in charts.

7. Conclusion and future work

The novelty of this paper is explained in terms of Data Ingestion and Data Processing. Data from wearable devices like Fit Bit is stored on HDFS using the library called Kafka in the application. Apache Spark handles the Streaming data in an efficient manner and the data is divided as Training and Test data. These data are acted upon with Machine Learning Supervised Algorithm K-Nearest Neighbour (KNN). Based on the value of “K”, neighbours are formed and the model is drawn. Thus obtained the predicted value from the KNN classifier and the Model. In this proposal the precious Input includes step count, fat, weight and calories. Various other Parameters are available in the Data set but not all the parameters are considered for the prediction by the Model. The predicted value here is the amount of calories burnt. This prediction is carried over against all the precious inputs mentioned above. This value is sent to the on demand cloud telephony ontology, which intimated the results to the end user through email, SMS, IVR, Mobile App and visualization Techniques. The selection of this intimation is decided by the Ontology service repository, where two ontologies are maintained. The interface between the telephony web services and ontology is carried over by SOAP/XML and OWL API. All these services are maintained in the web service registry. The end result reaches the user or the requestor through the HTTP request and Response API.

The future work of this paper lies on the creation of Ontology by the Machine Learning Model itself. Based on the type of data and the Analysis performed the semantic ontology can be created. Resource Description Framework (RDF) can be used for the creation of semantic ontology. In the Machine Learning model the memory size needs to be scalable, so that all the variables can be considered for computing the prediction parameter.

Acknowledgements

We wish to acknowledge the Department of Science and Technology, India and School of Computing, Sathyabama

University, Chennai for providing the facilities to do the research under the DST-FIST Grant Project No.SR/FST/ETI-364/2014.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] Takacs J, Pollock CL, Guenther JR, et al. Validation of the fitbit One activity monitor device during treadmill walking. J Sci Med Sp. 2014; doi:10.1016/j.jsams.2013.10.241.

[2] Diaz KM, Krupka DJ, Chang MJ, et al. An accurate and reliable device for wireless physical activity tracking. Int J Cardiol. 2015; doi:10.1016/j.ijcard.2015.03.038.

[3] Cadmus-Bertram LA, Marcus BH, Patterson RE, et al. Randomized trial of a fitbit-based physical activity intervention for women. Am J Prev Med. 2015; doi:10.1016/j.amepre.2015.01.020.

[4] Sasaki JE, Hickey A, Mavilia M, et al. Validation of the fitbit wireless activity tracker for prediction of energy expenditure. J Phys Activity Health. 2015; doi:10.1123/jpah.2012-0495.

[5] Slootmaker SM, Schuit AJ, Chinapaw MJ, et al. Disagreement in physical activity assessed by accelerometer and selfreport in subgroups of age, gender, education and weight status. Int J Behav Nutr Phys Act. 2009. doi:10.1186/1479-5868-6-17.

[6] Gusmer RJ, Bosch TA, Watkins AN, et al. Comparison of FitBit® ultra to ActiGraph™ GT1M for assessment of physical activity in young adults during treadmill walking. Open Sp Med J. 2014; doi:10.2174/1874387001408010011.

[7] Shameer* K, Badgeley* MA, Miotto R, et al. (2016). Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams. doi:10.1093/bib/bbv118.

[8] Valletta JJ, Torney C, Kings M, et al. Applications of machine learning in animal behaviour studies. Anim Behav. 2017; doi:10.1016/j.anbehav.2016.12.005.

[9] Farias VAE, Sousa FRC, Maia JGR, et al. Regression based performance modeling and provisioning for NoSQL cloud databases. Future Gener Comput Syst - Int J eSci. 2018; doi:10.1016/j.future.2017.08.061.

[10] Ilapakurti A, Kedari S, Vuppalapati C. (2016). The role of big data in creating sense EHR, an integrated approach to create next generation mobile sensor and wearable data driven electronic health record (EHR). IEEE Second International Conference on Big Data Computing Service and Applications. doi: https://doi.org/10.1016/j.future.2017.08.061.

[11] Manogaran G, Thota C, Lopez D, et al. Big data knowledge system in healthcare. In: Bhatt C, Dey N, Ashour A, editor. Internet of things and big data technologies for next generation healthcare (pp. 133–157, vol. 23). Cham: Springer; 2017.

[12] Shamim Hossain M, Muhammad G. Cloud-based collaborative media service framework for healthcare. Int J Distrib Sens Netw. 2014; doi:10.1155/2014/858712.

[13] Vigneshwari S, Aramudhan M. Personalized cross ontological framework for secured document retrieval in the cloud. Natl Acad Sci Lett. 2015; doi 10.1007/s40009-015-0391-3.

[14] Reichherzer T, Timm M, Earley N, et al. Using machine learning techniques to track individuals & their fitness activities. In: A Bossard, G Lee, L Miller, editor. Proceedings of 32nd international conference on computers and their applications, March, 20-22, 2017, Honolulu, Hawaii, USA. Winona, MN: ISCA; 2017. p. 119–124.

[15] Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis (Vol. 821). London: John Wiley & Sons; 2012.