Detaching the design, development and execution of big data analysis processes:

A case study based on energy and behavioral analytics

Anastasios Zafeiropoulos, Eleni Fotopoulou
Network Softwarization and Internet of Things Group
Ubitech
Athens, Greece
{azafeiropoulos, efotopoulou}@ubitech.eu

Aurora González-Vidal, Antonio Skarmeta
Departamento de Ingeniería de la Información y las Comunicaciones, Facultad de Informática
Universidad de Murcia
Murcia, Spain
{aurora.gonzalez2, skarmeta}@um.es

***Abstract*—Numerous tools and approaches are evolving towards the support of data mining and analysis processes, focusing on part or the overall lifecycle of such processes. In parallel, penetration of data analytics tools in the market is continuously increasing, along with their adoption by various stakeholders, including data scientists, decision and policy makers and business analysts. However, given the wide diversity in the needs for realizing an analysis and the level of expertise of the various stakeholders, there is a need for design and implementation of analysis toolkits that can support part or the overall lifecycle of an analysis process, without imposing dependencies on the type of tool or technology to be used. In the current manuscript, an approach for detaching the design, development and execution of big data analysis processes is detailed, focusing on the realization of energy and behavioral analytics, targeted to supporting the increase of energy efficiency in smart buildings through behavioral change of the citizens. The overall architectural approach as well as the set of energy and behavioral analysis processes integrated are detailed.**

***Index Terms*—Data analytics, Open APIs, energy analytics, behavioral analytics, OpenCPU.**

I. Introduction and Motivation

The volume, velocity and variety of data is rapidly growing the latest years, making inherent the need for powerful and innovative data techniques and tools, able to allow collecting, storing, analyzing, processing, and visualizing vast amounts of data, as stated at the Strategic Research and Innovation Agenda of the European Big Data Value Partnership [1]. The application of simple or complex data analysis techniques over such data can lead to advanced insights, leading decision making in various application domains.

Such techniques can be designed, developed and applied over the multitude of data analytics tools currently available or under development, including open-source and commercial tools. Such tools address analysis needs for a wide range of algorithms (e.g. R Project, Weka) or big data analysis needs based on big data computing frameworks (e.g. Apache Spark, Apache Fling, Apache Storm). Furthermore, tools supporting the design and realization of analysis workflows -consisted of a series of analytic processes that have to be realized- are available (e.g. Pentaho). The level and type of usage of such tools varies among the stakeholders involved in the data analysis process, an ecosystem that is also rapidly evolving.

These stakeholders include -among others- data scientists, data science researchers, business analysts, policy makers and big data computing engineers. Each type of stakeholder targets different type of usage and interpretation of analysis results, that may be associated with the design, development of execution of an analysis process. Under this perspective and considering the associated level of expertise, the learning curve for adopting a tool or a programming language for developing data analysis processes also ranges.

It can be claimed that there is inherent a need for the design, development and validation of toolkits that can support the independent realization of data analysis steps, while in parallel facilitating as much as possible the usage and re-usability of the available tools and developed processes. In this way, the penetration of data analysis toolkits in various application domains and the optimal exploitation of the provided functionalities can be realized. A clear separation of concerns among the realization of each step of an analytic process by a relevant stakeholder has to be done.

In the current manuscript, an approach for supporting the detaching of the design, development and execution of big data analysis processes is provided, targeting to the realization of energy and behavioral analytics in the energy efficiency domain in smart buildings. The proposed approach is realized over the ENTROPY platform [2]. ENTROPY regards an innovative energy-aware information technology (IT) ecosystem, aiming to support the design and development of novel personalized energy management and awareness services that can lead to occupants' behavioral change towards actions that can have a positive impact on energy efficiency. A set of data analytic processes are designed, developed and supported through the implemented analysis toolkit. Separation of concerns among the related stakeholders is implemented based on the adoption and integration of the OpenCPU open source tool for supporting embedded scientific computing and reproducible research [3].

OpenCPU provides a reliable and interoperable HTTP API for data analysis. Based on the provided API, appropriate

customization is realized for supporting the design, development and execution of energy and behavioral analytics over the ENTROPY platform in an independent way. Software developers are able to develop their analysis scripts without any restriction in the programming language (e.g. R, Python, Java) and make it available in the platform. Data scientists are able to design analytic workflows, consisting of set of processes and related input and output parameters in a user friendly and intuitive way. Decision makers are able to define the timeline for execution of the analysis processes and acquire access to the provided results. The overall analytics toolkit is applied in the energy domain, however it is designed and implemented in a generic fashion, making it suitable for various application domains.

The structure of the paper is as follows: in section II we detail the overall architectural approach focusing on the openness of the overall solution and the way that part or the overall workflow of an analysis process can be realized; in section III, we provide information regarding a set of data mining and analysis processes already implemented and integrated, focusing on energy and behavioral analytics, while in section IV we denote a set of conclusions and plans for future work.

II. Architectural Approach

In this section, we present the proposed architectural approach for detaching the design, development and execution of big data analysis processes. The proposed approach consists a generic and stand-alone framework that can be applied in any domain, however, in the scope of this manuscript it is presented within the energy efficiency in smart building domain, given the instantiation of the architecture in the aforementioned ENTROPY platform [2].

Towards the design of the overall architecture, the following requirements and architectural choices were considered. A basic requirement regarded the need to provide simple and homogeneous access to a variety of algorithm execution packages, without the necessity to have a deep knowledge of the execution requirements of each algorithm. Such a requirement is supported through the provision of access to set of registered algorithms along with the provision of user-friendly interfaces for specification of main execution parameters (default options are also provided). In order to support the flexibility on realizing part or an overall analysis workflow, the design and implementation of the solution is based on a microservices-based architecture enabling a modular way of both registering new algorithms at the analysis engine as well as invoking the execution of the analytic processes. Openness and extensibility is supported through the specification of a set of open APIs for accessing the various analysis mechanisms by the various stakeholders and mainly software developers and data scientists. In this way, interconnection of the provided solution with external data analysis and management toolkits can take place.

Such toolkits may regard simple or big data analysis toolkits. Depending on the needs of the analysis (e.g. type and size of input data streams) and the associated toolkit, different execution mode may be selected and applied. For instance, in the realized implementation over the ENTROPY platform, for small data streams plain, R algorithms are used in sequential or parallel execution mode, while for heavy data streams, algorithms are executed on clustering mode by well-known big data frameworks (e.g. Apache Spark, Apache Mesos) over managed or stand-alone clusters.

The proposed architectural approach for detaching the design, development and execution of big data analysis processes is depicted at Figure 1. Data is made available through a big data repository, where energy and end users related data is stored. In the provided solution within ENTROPY, a Query Builder is developed for designing queries that fetch specific data views that are fed as input for an analysis process. Interconnection of the ENTROPY components with the analysis toolkits is based on the OpenCPU system. In the case of large-scale data processing and the need for a big data analysis framework, the Apache Spark engine is used, where the analysis process is realized in a set of worker nodes, each one of which is hosting an Apache Spark OpenCPU Executor [4]. The set of worker nodes are formulating a cluster orchestrated by a cluster manager.

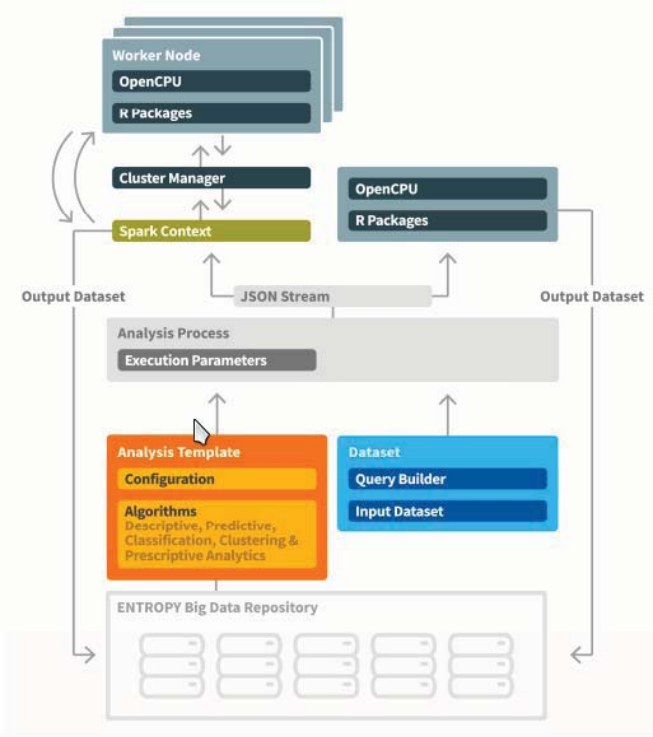

Fig. 1. Architectural approach.

Data mining and analysis processes can be realized by associating an analysis template with a set of input and output data (see Figure 2). An analysis template regards an algorithm with the associated software and execution endpoint, along with a set of parameters that have to be defined by the end user for its execution. Such parameters include input parameters for the algorithm along with their description and their default

value, as well as output parameters along with their type (text, image, data, html). Four types of potential parameters are supported: plain values, queries to be executed for providing the input datasets (training and/or evaluation datasets), data files and code snippets. As already mentioned, queries are used for fetching data collected by sensor data streams (e.g., energy consumption, humidity, and indoor temperature data per hour for a specific room) and queries for fetching data related to the set of users participating at the energy efficiency campaign (e.g., a set of users with an educational level relevant to a Master's degree) [2]. Upon the execution of the queries, streams of the input training or evaluation datasets are provided to the analysis engine. An analysis process is also associated with a set of execution parameters that denote whether an analysis should be realized in a manual or automated way, as well as the periodicity factor for the latter case.

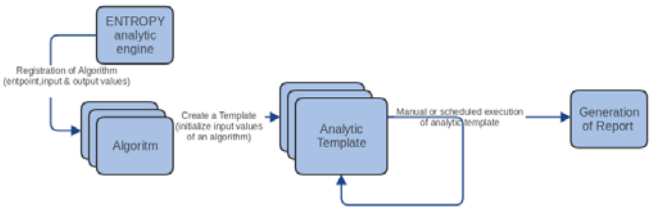

Fig. 2. Relationship between algorithm, analysis template and report.

Going one step further, the proposed approach supports the design and implementation of data analysis workflows, consisted of a series of data mining and analysis processes, interconnected among each other in terms of input/output data streams/objects. As depicted at Figure 3, upon the execution of an analysis template, the outcome can be visualized in the form of a report and/or constitute the input for another analysis template. In this way, complex analysis processes can be broken down in smaller processes interlinked in the form of a workflow.

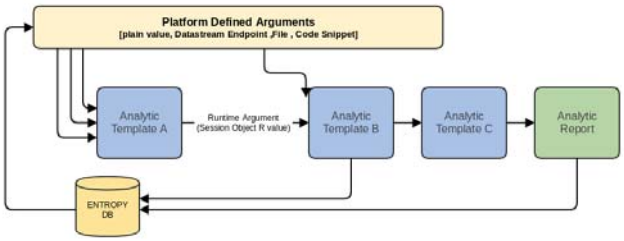

Fig. 3. Analysis chaining workflow.

According to the OpenCPU specification, the output of the analysis template execution is a session object that contains on memory all output values. Session object values returned by a template call can feed as arguments a subsequent template call, without ever retrieving the object. All state in OpenCPU is managed by controlling objects in sessions on the server. Analysis templates chaining becomes even more powerful taking under consideration that a code snippet can be considered as an input value parameter. This injects raw code into the function call and can be extremely useful at analysis workflows that explicitly let the user do some coding.

III. Energy and Behavioral Analytics

At this section, we present a set of algorithms supporting the extraction of energy and behavioral analytics that are integrated in the ENTROPY platform. For each algorithm, a short description of the main analysis process realized, and the algorithm execution context is provided. The objective is to present the capacity and the modularity of the presented approach, over a set of designed and implemented data analysis processes. It should be noted that the list with the set of algorithms is indicative, since the platform support the ease introduction of further algorithms, taking into account the end user needs.

A. Energy Analytics

In Table I, the set of supported algorithms for the extraction of energy analytics are presented.

TABLE I. Energy Analytics Algorithms

Energy analytics algorithms		
name	***Description***	***Execution***
EntArima	Basic forecasting technique that can be used as a foundation for more complex models.	R package
EnergyClust	Time series clustering algorithm for grouping areas with similar energy demands. This algorithm can be chained with any timeseries forecasting algorithm to predict energy use.	R package & Distributed execution
GBDT4consumption	Offers insights on the way that energy consumption is related to other environmental variables.	Distributed execution
StreamQ	Checks data stream quality via time series data analysis, including outliers and constant values detection.	R package (sequential & parallel mode)
C&HDegree Days	Calculation of Heating degree days (hdd) and Cooling degree days (cdd) per area or subarea of the registered buildings.	R package
EnergyBaseline	Needs input from "C&HDegree Days" package. It examines the correlation between degree days with energy consumption.	R package
ECompPeriod	Realizes set of comparisons with regards to energy consumption per main area or subarea, taking into account the area characteristics.	R package
EntPastForecasting	Measures the impact of a campaign intervention on energy consumption of the registered building areas.	Distributed execution

Concerting EntArima, it supports mechanisms for the forecasting of future values as well as mechanisms for time series decomposition aiming at identifying trends and seasonality aspects. With regards to forecasting, an ARIMA model is a popular and flexible class of forecasting model that utilizes historical information to make predictions. This type of model is a basic forecasting technique that can be used as a foundation for more complex models. Within ENTROPY, the specific package focuses on examining time series data for any sensor enabled attribute (e.g. temperature, CO2, energy consumption), aiming to fit it to an ARIMA model and support forecasting of upcoming values. The execution of the specific package is a general purpose analytic process with no need of extra configuration that takes as input an hourly time series sensor data set and returns predictions for the following 24 hours. With regards to time series decomposition, EntArima is used for identifying historical patterns depending on the kind of the analyzed time series data. Some of the insights upon the execution of the algorithm are the identification of seasonal patterns in energy consumption, leading to better predictions (e.g. predicting the expected number room occupancy as well as the regulated temperature of each HVAC in each building area and estimating the effect of a new campaign on energy consumption). The results of the specific package are enabled in the form of a plot so as to be easily interpretable on behalf of the platform administrator and campaign managers.

GBDT4consumption is a customized implementation of the gradient boosting algorithm aiming at gaining insights at how energy consumption is related to other environmental variables. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. GBDT4consumption is adaptable, easy to interpret and produces highly accurate models. However, it is computationally expensive and requires all training data to be in main memory. As training data becomes ever larger, ENTROPY project makes use of a distributed implementation of the GBDT algorithm by parallelizing decision trees training, leading to execution of the algorithm over Spark.

EnergyCLUST improves load forecasting accuracy which is a challenging problem due to the variety of factors that influence the load. Traditional load forecasting methods include linear models for time-series load forecasting, such as autoregressive with moving average (ARMA) models. Typical methods for improving further load forecasts include the clustering of the data generated from smart meters [5]. That way, the knowledge about consumption behavior of users is used to improve the accuracy of forecasting. Instead of developing a single forecasting model for the accumulated load consumption of all rooms, clustering can be used to group the ones having similar demand profiles. The choice of the time series clustering algorithm is determinant in the accuracy of the models. EnergyCLUST can use both k-shapes and k-means algorithms. Previous works have compared the advantages of using the k-shapes time series clustering model against k-means in an energy consumption prediction scenario [6]. k-shapes captures the shape/based similarity between time series by using a normalized version of the cross/correlations measure and claims to be the only scalable method that significantly outperforms k-means. Once rooms have been clustered, the administrator can chain EnergyCLUST with another algorithm package (e.g. EntArima) that implements one of the several machine learning algorithms to predict energy consumption.

Data quality assessment is the crucial process in engineering systems where sensor data are examined in terms of their overall quality, characteristics that is very important for meaningful interpretation of analysis results over them [7]. Defining the quality of the sensor data is important because it has impact on the selection of the model, the estimation of parameters and consequently, on forecasts. To do so, we have followed the approach described in [8] for automatic detection of outliers in time series data, leading to the development of the StreamQ R package [9]. StreamQ package may be applied previous to general seasonal and non-seasonal ARMA process so as to filter bad quality data streams. The different types of outliers are counted and then, the proportion of outliers with respect to the length of the stream is calculated as the quality measure. The result of the algorithm process is a data quality rate from 0 - 1 that directly feeds the streams quality tags within the ENTROPY platform. This algorithm can be sequentially or parallel executed for a list of data streams.

C&HDegreeDays data mining and analysis process regards the calculation of Heating degree days (hdd) and Cooling degree days (cdd) per area or subarea of the registered buildings. For each area, it is calculated the average temperature per hour for indoor and outdoor conditions. The base temperature is denoted separately for winter and summer. Based on the calculated temperature difference, the associated indicative energy waste is estimated. Such energy waste is calculated based on an indicator for the rate of transfer of heat per square meter. Indicators regarding the estimated values and the overall energy consumption (e.g. hdd/energy consumption (days/KWh) and cdd/energy consumption indicators) are calculated in a daily/weekly fashion, leading to comparisons among similar buildings (in terms of size, floors etc.).

EnergyBaseline: For examination of energy consumption and energy waste, it is applied a linear regression model for examining the correlation between degree days with kWh. Once the formula of the regression line is made available, it is used to calculate the baseline, or the expected energy consumption, given the degree days. Hence, it is possible the comparison of these energy indicators with the actual energy consumption for that period and is possible to determine whether more energy was used than expected.

ECompPeriod supports the realization of set of comparisons with regards to energy consumption per main area or subarea, taking into account the area characteristics. Comparison of energy consumption is realized in an hourly, daily, weekly or monthly level. Comparison may regard the average/min/max values obtained, while comparisons considering the overall surface of the area, the number of occupants etc. are also supported. Modularity in terms of the definition of the comparison periods is supported through the design and

execution of appropriate queries through the implemented ENTROPY Query Builder.

EntPastForecasting instead of doing a typical forecasting, aims to measure the impact of a campaign intervention on energy consumption of the registered building areas. EntPastForecasting takes as input data of all the historic sensor data streams related with energy consumption in a place. Then it splits the pool of data in training and testing dataset. The date of initialization of a campaign is taken as point of data splitting into training and testing data. A forecasting model is defined with the train dataset. Then the predicted values that come out of the forecasting model are compared with the real values. The difference between the predicted values and the real ones is interpreted as the proofed impact of the campaign intervention. Given the high computational needs for generating the forecasting model, we make use of the random forest model available in Spark.

B. Behavioral Analytics

In addition to the energy analytics, a set of behavioral analytics are designed and implemented, aiming at motivating end users to adopt more energy efficient lifestyles. Considered user's energy-consuming behavior, the use of lights, air-conditioning, computers and other office equipment directly affects the operation of buildings and consequently their energy use. It should be noted that in ENTROPY, a recommender system is implemented that creates personalized recommendations based on the results of the behavioral analysis (e.g. user profiling). In Table II, an indicative set of supported behavioral analytics processes is detailed.

TABLE II. BEHAVIORAL ANALYTIC ALGORITHMS

Behavioral analytics algorithms		
name	***Description***	***Execution***
Behavioral Heatmap	Graphical representation depicting the summary of the behavioral characteristics of end users per campaign.	Embedded within the platform (Java)
FeedbackStatistics	Graphical representation depicting the responsiveness of the end users to the generated content.	Embedded within the platform (Java)
BehaviouralContentDistribution	Selection of personalized content depending on the behavioral profile of the consumer.	Embedded within the platform (Java)
ConsumersCLUST	Bottom up hierarchical agglomerative clustering algorithm to categorize consumers depending on their responsiveness to recommendations.	R package

BehavioralHeatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. The data mining and analysis process supported regards the creation of heatmaps, depicting the summary of the behavioral characteristics of end users per campaign. Such characteristics are produced through the processing of online questionnaires filled in by campaign participants. The processing of the questionnaires regards the execution of a script realizing profiling analysis per end user. Comparison of initial and final profiling analysis can lead to meaningful insights with regards to the behavioral change achieved per end user. For instance, Figure 4 depicts the personal motives towards energy saving at work for one of the ENTROPY pilots placed at University of Murcia.

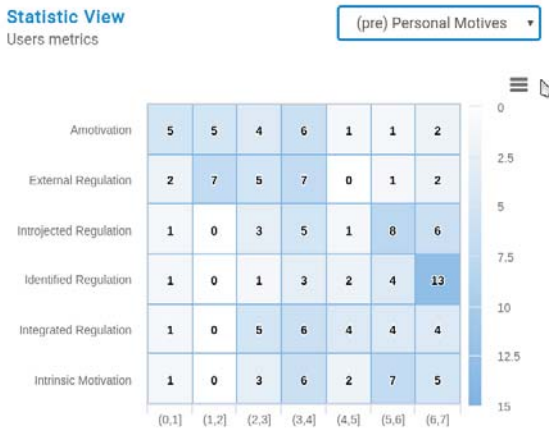

Fig. 4. Personal motives towards energy saving at work.

FeedbackStatistics regards the extraction of descriptive statistics regarding the way that end users (energy consumers) respond to the generated content (e.g. statistics regarding their feedback to the provided tips, tasks, questions, quizzes are responded). Based on such statistics, and considering the behavioral characteristics of each end user, the BehavioralContentDistribution package provides personalized recommendations to end users. In this way, the frequency and type of the recommendation is adapted to the personality and engagement level of each user.

ConsumersCLUST takes as input a set of predefined synthetic metrics such as number of clicks per user, percentage of tasks/tips/questions/quizzes performed per user and savings compared to the predicted consumption per room (each user carries out activities in one room, but each room is an entity that belongs to several users). According to such variables ConsumersCLUST performs a clustering, by using a bottom up hierarchical agglomerative clustering algorithm. Initially, the algorithm places each user in a cluster of his own. Then, at each step, it merges the two most similar clusters. The similarity between two clusters is defined as the minimum similarity between any two users that belong to these clusters (max linkage). Three final groups are defined: successfully classified, medium classified, badly classified. The parameters of successfully classified users remain untouched. Those who are medium or badly classified are sent more tips in order to collect more information about their behavior and also their

classification will be adjusted according to the percentage of successful tasks.

IV. Conclusions and Future Work

Following the evolving trend in the development and adoption of data mining and analysis toolkits by various application domains and stakeholders, in the current manuscript an open and modular approach able to support various parts of an analysis process in an independent and extensible way is presented. The presented approach is built upon the OpenCPU system for embedded scientific computing, while the overall solution is integrated in the ENTROPY platform, aiming at supporting set of energy and behavioral analytics targeting at increasing energy awareness and motivation for energy efficiency of citizens in smart homes. An indicative set of implemented energy and behavioral analysis algorithms by exploiting the open APIs provided by the presented approach is detailed.

Considering the work presented in this manuscript, open issues for extensions are identified, including the design and deployment of simple user interfaces for supporting the analysis workflow processes, reducing even more the complexity of designing, developing and executing analysis processes, as well as the extensive experimentation with regards to the performance aspects of supporting workflows associated with real time big data streams.

Acknowledgment

This work is supported by the European Commission Research Programs through the Entropy Project under Contract H2020-649849.

References

[1] European Big Data Value Partnership - Strategic Research and Innovation Agenda, Available Online: http://www.bdva.eu/sites/default/files/EuropeanBigDataValuePartnership_SRIA__v3.pdf

[2] Fotopoulou, E.; Zafeiropoulos, A.; Terroso-Sáenz, F.; Şimşek, U.; González-Vidal, A.; Tsiolis, G.; Gouvas, P.; Liapis, P.; Fensel, A.; Skarmeta, A. Providing Personalized Energy Management and Awareness Services for Energy Efficiency in Smart Buildings. Sensors 2017, 17, 2054.

[3] OpenCPU system for embedded scientific computing, Available Online: https://www.opencpu.org/

[4] Apache Spark OpenCPU Executor (ROSE). Available Online: https://github.com/onetapbeyond/opencpu-spark-executor.

[5] Ilić, D., da Silva, P. G., Karnouskos, S., & Jacobi, M. (2013, March). Impact assessment of smart meter grouping on the accuracy of forecasting algorithms. In Proceedings of the 28th Annual ACM Symposium on Applied Computing (pp. 673-679). ACM

[6] Fahiman, F., Erfani, S. M., Rajasegarar, S., Palaniswami, M., & Leckie, C. (2017, May). Improving load forecasting based on deep learning and K-shape clustering. In Neural Networks (IJCNN), 2017 International Joint Conference on (pp. 4134-4141). IEEE.

[7] Mogles, N., Walker, I., Ramallo-González, A. P., Lee, J., Natarajan, S., Padget, J., ... & O'Neill, E. (2017). How smart do smart meters need to be?. Building and Environment, 125, 439-450.

[8] De Groot, J. I., & Steg, L. (2008). Value orientations to explain beliefs related to environmental significant behavior: How to measure egoistic, altruistic, and biospheric value orientations. Environment and Behavior, 40(3), 330-354.

[9] González-Vidal, A., Ramallo-González, A. P., Terroso-Sáenz, F., & Skarmeta, A. (2017, December). Data driven modeling for energy consumption prediction in smart buildings. In Big Data (Big Data), 2017 IEEE International Conference on (pp. 4562-4569). IEEE.