ARTICLE IN PRESS

International Journal of Information Management xxx (xxxx) xxxx

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities

Marijana Zekić-Sušac[a,*], Saša Mitrović[a], Adela Has[b]

[a] Faculty of Economics in Osijek, University of Josip Juraj Strossmayer in Osijek, Gajev trg 7, 31000, Osijek, Croatia

[b] Faculty of Economics in Osijek, University of Josip Juraj Strossmayer in Osijek, Trg Lj. Gaja 7, 31000, Osijek, Croatia

ARTICLE INFO

JEL classification:
O21
O38
P18

Keywords:
Planning models
Energy efficiency
Machine learning
Public sector
Smart cities

ABSTRACT

Energy efficiency of public sector is an important issue in the context of smart cities due to the fact that buildings are the largest energy consumers, especially public buildings such as educational, health, government and other public institutions that have a large usage frequency. However, recent developments of machine learning within Big Data environment have not been exploited enough in this domain. This paper aims to answer the question of how to incorporate Big Data platform and machine learning into an intelligent system for managing energy efficiency of public sector as a substantial part of the smart city concept. Deep neural networks, Rpart regression tree and Random forest with variable reduction procedures were used to create prediction models of specific energy consumption of Croatian public sector buildings. The most accurate model was produced by Random forest method, and a comparison of important predictors extracted by all three methods has been conducted. The models could be implemented in the suggested intelligent system named MERIDA which integrates Big Data collection and predictive models of energy consumption for each energy source in public buildings, and enables their synergy into a managing platform for improving energy efficiency of the public sector within Big Data environment. The paper also discusses technological requirements for developing such a platform that could be used by public administration to plan reconstruction measures of public buildings, to reduce energy consumption and cost, as well as to connect such smart public buildings as part of smart cities. Such digital transformation of energy management can increase energy efficiency of public administration, its higher quality of service and healthier environment.

1. Introduction

The importance of energy efficiency of buildings is emphasized in directives of European Parliament and Council (particularly in the directives 2012/27/EU and 2010/31/EU) which state that 40 % of all energy consumption in the European Union belongs to the building sector which is itself expanding. Therefore, the directive sets the goal of reducing energy consumption by 20 % and increasing the natural in the European Union by 2020 and requires actions that will enable cost-effective energy savings. As a result, a number of national action plans for increasing energy efficiency have been enacted. Croatia is among the highest ten energy intensity countries in EU (Odyssee-Mure, 2016). The government of Croatia has established the central information system for managing energy - Croatian energy management information system (EMIS). The system gathers data about public sector buildings – their constructional and energetic characteristics, as well as their energy consumption data and CO2 emission in a centralized database with a web application accessible by all the managers of all public buildings, local and national government. The situation is similar in other EU countries, although most of those systems rely on standard statistical methodology and lack intelligent models based on machine learning as well as Big Data platforms that enable processing large amounts of data. There is a need of such intelligent systems that will be able to make predictions and feature extraction with the aim to assist decision makers in public sector.

Rare papers suggest the architecture of smart building information systems. One of the recent one is proposed by Marinakis and Doukas (2018) who suggested an advanced Internet of Things (IoT) based system for intelligent energy management in buildings. Our research refers to the work of Marinakis and Doukas (2018) and suggests a modified model which focuses on predictive analytics that could be integrated into the smart building concept. Thus, this paper aims to fulfill the gap by suggesting an architecture of the intelligent (smart) energy management system specifically designed for public sector that

* Corresponding author.
E-mail addresses: marijana@efos.hr (M. Zekić-Sušac), smitrovi@efos.hr (S. Mitrović), adela.has@efos.hr (A. Has).

https://doi.org/10.1016/j.ijinfomgt.2020.102074
Received 4 March 2019; Received in revised form 1 January 2020; Accepted 10 January 2020

Please cite this article as: Marijana Zekić-Sušac, Saša Mitrović and Adela Has, International Journal of Information Management, https://doi.org/10.1016/j.ijinfomgt.2020.102074

can support investment decisions in the public sector on the local as well as on the national level. The architecture of the MERIDA system suggested in this paper relies on our previous research in which machine learning models have been developed to predict the energy efficiency level of public buildings (Has & Zekić-Sušac, 2017), as well as energy consumption of natural gas (Tonković, Mitrović, & Zekić-Sušac, 2018), and electricity (Zekić-Sušac, Scitovski, & Has, 2018). It also incorporates the abilities of IoT to facilitate data gathering as well as Big Data technology to store and process data. The machine learning methods used to create predictive models were artificial neural networks (ANN), support vector machines (SVM) and recursive partitioning such as classification and regression trees (CART), conditional inference trees (CTREE), random forest (RF), and gradient boosted trees (GBT). The models were tested on real data from EMIS system, and the most successful ones were included in the design of the machine learning based information system architecture that could be used in public sector.

The paper provides an overview of related literature in the next section, after which the description of data and methods is given with the focus on preprocessing methods and machine learning methods used to create predictive models. Section 4 presents the results i.e. the proposed architecture of the intelligent system and discuss its possible implementation, followed by the conclusion.

2. Theoretical background and overview of literature

2.1. Theoretical background on smart cities, Big Data and intelligent models

A systematic literature review of the development of smart city concept is provided by Cocchia (2014). Calvillo, Sánchez-Miralles, and Villar (2016) define a smart city as a sustainable and efficient urban center that provides a high quality of life to its inhabitants through optimal management of its resources. They describe some of the existing information systems on the market that aim to manage and optimize energy consumption in buildings, and state that these tools are mostly used by energy/facility managers, and specialists, who make decisions on reconstructions. Papa, Gargiulo, and Zucaro (2014) emphasize that the majority of the research efforts are focused on the performance and energy efficiency of buildings, renewable energy plants and transport systems, while there is a lack of quantitative and holistic studies that deal with energy issue at urban scale (Papa et al., 2014). Regarding the variables found important for a city energy consumption, the two groups were identified in previous research: physical and environmental (Papa et al., 2014). They urge researchers to develop a comprehensive theoretical framework with suggested methods, techniques and strategies for the reduction of energy consumption in cities that will allow to generalize individual results.

Due to the production of large amounts of high frequency multimedia data within smart city networks, it is necessary to ensure that the data are stored and processed in a Big Data platform. Gandomi and Haider (2015) analyzed various definitions of Big Data concept and agreed that one of the most common is that it represents new technologies designed to address the volume, variety, and velocity challenges of Big Data, although additional characteristics such as veracity and variability (complexity) are also added recently. The authors emphasize the importance of including broader efficient analytical methods that are able to leverage massive volumes of heterogeneous data stored within Big Data as unstructured text, audio, and video formats (Gandomi & Haider, 2015). The role of big data in terms of connection with the feasibility of smart city initiatives, sustainability and improvement on the living standards was explored by Hashem et al. (2016). Authors suggested a business model for big data and smart city. However, they also distinguished two types of challenges of big data in smart cities: challenges from the business perspective such as planning, sustainability, cost of acquiring smart city, as well as challenges from the technological perspective such as privacy, data analytics, data integration, computational intelligence algorithms for smart city big data analytics. The development of intelligent models is thoroughly analyzed by López-Robles, Otegi-Olaso, Porto Gómez, and Cobo (2019). They state that the intelligence is understood as a process of gathering, analyzing, interpreting, and disseminating high-value data and information at the right time for use in the decision-making process, and that the usage of intelligent systems is an imperative in today's organizations. Marinakis and Doukas (2018) stress out that in energy sector it is important to use Internet of Things (IoT) networks in conjunction with the methods using "intelligent" energy management that will lead to efficient energy and environmental management of the "smart" building. Simonofski, Vallé, Serral, and Wautelet (2019) emphasized that the design of a smart city has to involve citizen participation and the respective context of the city. They have conducted a research at two cities aimed to be smart, one in Belgium and another one in Sweden. The research has identified five context factors important for citizen participation strategies in smart cities, such as the smart city consideration, the drivers for participation, the degree of centralization, the legal requirements, and the citizens' characteristics.

Recent research also emphasizes a need for prescriptive analytics which, relying on intelligent data analytics, provides actions that create business values (Lepenioti, Bousdekis, Apostolou, & Mentzas, 2020).

2.2. Overview of previous research

The rapid development of artificial intelligence (AI) brings a significant impact, opportunities, challenges, as well as new research agendas in various domains, shaping the future of industry and society, including smart cities (Dwivedi et al., 2019; Pérez-Chacón et al., 2018). Duan, Edwards, and Dwivedi (2019) discuss a growing role of AI related with the appearance of Big Data technologies and super computing power. The new generation of AI is rapidly expanding and has again become an attractive topic for research. This paper aims to identify the challenges associated with the use and impact of revitalized AI based systems for decision making and offer a set of research propositions for information systems (IS) researchers. The paper first provides a view of the history of AI through the relevant papers published in the International Journal of Information Management (IJIM). It then discusses AI for decision making in general and the specific issues regarding the interaction and integration of AI to support or replace human decision makers in particular.

A systematic review of literature about smart city research is provided by Israilidis, Odusanya, and Mazhar (2019) who analyzed a number of scientific publications and identified five most frequent topics such as (1) strategy and vision, (2) frameworks, (3) enablers and inhibitors, (4) citizen participation, and (5) benefits. They have introduced knowledge management perspective into smart city shaping, emphasizing the importance of knowledge sharing and co-learning among cities as key factors of success. Ismagilova, Hughes, Rana, and Dwivedi (2019) have conducted a comprehensive review of the role of smart cities in creating sustainable cities and communities, and their key findings can serve as an informative framework for research on smart cities. Related to technologies connected to smart city development, Ismagilova, Hughes, Dwivedi, and Raman (2019) emphasize that the previous authors have mostly mentioned IoT, cloud computing, and Bluetooth as important. In addition, there is also a review of barriers for smart city development analyzed by Rana et al. (2019) who used analytic hierarchy process (AHP) method to select the most important barriers. Their research conducted in India has shown that the barriers can be categorized into 'Governance', 'Economic', 'Technology', 'Social', 'Environmental', and 'Legal and Ethical'.

Regarding methodology used to construct smart city architecture, it is logical to conduct the policy selection first, followed by the technology-focused layers of the architecture. Wu and Chen (2019) proposed a policy selection structural model, which uses three different methods for multi-criteria decision making. The first phase is based on a modified Delphi method used by panel members who served as experts for selecting the decision elements. In order to determine the priority of alternatives in relation to the goal, the analytic hierarchy process (AHP) method is used, while the last phase was conducted by the zero-one goal programming models with the aim to select a feasible portfolio.

When technology-focused structure of smart city is concerned, IoT is one of the most important and also a challenging element. Janssen, Luthra, Mangla, Rana, and Dwivedi (2019) identified and prioritized the challenges for adopting and implementing IoT in smart cities. They used the Interpretive Structural Modelling (ISM) method to identify the contextual interactions between the challenges, while their dynamic interactions were analyzed by an integrated MICMAC-ISM approach. This integrated approach is able to classify challenges according to their driving and dependence power (MICMAC) and also takes into account their contextual relationships among elements (ISM). Jangili and Bikshalu (2017) identified three layers of an IoT system: (1) the perception layer (devices such as sensors, GPS, RFID, that can percept, detect objects, collect and exchange information with other devices through communication networks, (2) the network layer (combination of Internet and short-range networks to carry information from the perception to the application layer), and (3) the application layer (applications that process information from previous layers and are able to suggest better power's distribution and management strategies). The applications can include smart homes and smart cities (smart grid) applications, online monitoring of power lines, demand-side energy management, integration of distributed (renewable) energy sources, integration of electric vehicles. They emphasize Supervisory Control and Data Acquisition (SCADA) systems as the core of decision making in smart grid since it can provide demand-energy efficiency. Since the IoT network produces large amounts of multimedia data captured with sensors, cameras, and other devices, the issue of searching, browsing, and visualizing large amounts of multimedia data is analyzed by Grierson, Corney, and Hatcher (2015) who suggest the SIZL (Searching for Information in a Zoom Landscape) system for testing efficiency of graphical representations of multimedia data.

The architecture of an advanced Internet of Things (IoT) based system focused to intelligent energy management in buildings is suggested by Marinakis and Doukas (2018). Their architecture consists of three layers: (1) data integration layer, (2) prediction models / rules, and (3) action plans suggestions. They describe five pillars in the data integration layer such as building's data, energy production, energy prices, weather data, and end-users' behavior. Besides the conceptual architecture design, they also bring a semantic framework of the communication system for data integration based on Ztreamy system, a Python-based semantic service and create Optimus ontology. It was also emphasized in their paper that available systems lack interoperability with other applications, expansion, and semantics. While their aim was to integrate cross-domain data to produce daily and weekly action plans for the end-users with actionable personalized information for each building, our focus is on the intelligent support for decision makers on the local and national level to indicate buildings that have priority in reconstruction measures and create more efficient reconstruction plans in public sector that result with higher savings.

Galicia, Talavera-Llames, Troncoso, Koprinska, and Martínez-Álvarez (2019) presented ensemble models based on machine learning methods for forecasting big data time series. Through ensemble models decision tree, gradient boosted trees and the random forest were combined. They used a weighted average combination for predictions of each ensemble member, where different weights for each algorithm are computed based on its previous performance. The suggested model showed good performance in prediction on Spanish electricity consumption data for 10 years measured with a 10-minute frequency.

A deep feed forward neural network provided by the H2O big data analysis framework was used by Torres, Galicia, Troncoso, and Martínez-Álvarez (2018). Their platform was the Apache Spark for distributed computing. Electricity consumption in Spain from 2007 to 2016, with a ten-minute frequency sampling rate was the target variable. Authors compared the accuracy and runtimes vs. computing resources and the size of the dataset. They concluded that in terms of model accuracy deep learning stands out along with decision trees as best methods in terms of scalability in processing big data time series.

There were previous efforts on constructing intelligent management support for public sector buildings in Croatia. Tomšić, Gašić, and Čačić (2015) introduced Intelligent Information System for Monitoring and Verification of Energy Management in Cities (ISEMIC) as an upgrade of EMIS system. They suggested a three-layered architecture that gathers data on buildings and their energy and water consumption, monitors consumption indicators, detects anomalies, while in addition analyzes linear relationships among attributes by multiple regression analysis, sets energy efficiency targets and reports energy and water consumption savings by cumulative sum technique. Our approach could be considered as an additional upgrade of the EMIS/ISEMIC system by introducing Big Data collection with IoT, and machine learning prediction level that were not considered before, as well by more extensive data visualization and decision making levels.

3. Material and methods

3.1. Data

A real dataset from Croatian energy management information system EMIS was used which initially consisted of more than 17,000 public buildings with an extensive number of variables indicating their physical, environmental attributes as well their energy consumption. The physical group of attributes included construction, heating, cooling, and energy data, while meteorological, geospatial, and occupational attributes described environmental factors. The input space consisted of 82 attributes plus the energy consumption data for the following energy sources: electricity, natural gas, and heat, water, as well as CO_2 emission in the period from 2006 to 2017. The output variable in this research was the energy consumption named Q1HNDREF which represents the specific energy consumption (SEC) (expressed in kWh/(m^2.a)) as the energy consumed per m^2 of heated floor area of a building (in accordance with Annex VIII of the EC Directive (European Commission, 2014). It also incorporates the average difference in indoor (19 °C) and outdoor temperature over a heating season, as well as the duration of heating season. The complete formula for calculating this energy consumption is given in Annex VIII (European Commission, 2014). Due to a large number of inputs, only the inputs that are selected by variable reduction procedures are shown in Table 1 as well their descriptive statistics.

For the purpose of this research, the initial prediction models based on machine learning methods and data from Table 1 were created. Furthermore, an architecture for intelligent system of energy management is proposed that will more intensively use IoT, Big Data platform, data preprocessing, and machine learning based modeling to assist in decision making process on energy savings and reconstruction measures. The modeling procedures by machine learning methods required a phase of data preprocessing including outlier removal and missing values replacement using the suggested algorithms (Scitovski, Zekić-Sušac, & Has, 2018). Data were also normalized by the distance between minimum and maximum values of each attribute. After the preprocessing phase, a sample of public 575 buildings was selected to create machine learning values. The dataset was divided randomly into the train and test subset, such that 70 % of data was used for training, and 30 % was a hold-out sample used to validate the results of all three tested machine learning methods. The training subset was additionally divided to train data used to train DNNs (80 % of training set), while the rest of the training set (20 %) was used to optimize the network architecture, parameters, and the learning time in a cross-validation procedure. The sampling is presented in Table 2.

Table 1
Descriptive statistics of input predictors selected by variable reduction procedures.

Group of attributes	Attribute name	Descriptive statistics
Geospatial attributes	County (21 counties)	Categorical, min = 1, max = 21
	Type of object (1 = administrative, 2 = cultural, 3 = educational, 4 = general, 5 = business, 6 = lightning, 7 = residential, 8 = social, 9 = hospitality, 10 = military, 11 = healthcare)	Categorical, 1 = 24.69 %, 2 = 1.04 %, 3 = 42.95 %, 4 = 4.69 %, 5 = 1.22 %, 7 = 0.17 %, 8 = 4.52 %, 9 = 0.52 %, 11 = 20.17 %
Construction attributes	F0 (Shape factor of the building)	Continuous, min = 0.1820, max = 1.5470, mean = 0.5803
	Year of completion of construction (1 = < 1919, 2 = 1919–1945, 3 = 1946–1970, 4 = 1971–1980, 5 = 1981–1990, 6 = 1990–2000, 7 = > 2000)	Categorical, 1 = 16.87 %, 2 = 5.04 %, 3 = 11.48 %, 4 = 16.87 %, 5 = 25.91 %, 6 = 8.69 %, 7 = 15.13 %
Heating attributes	Annual thermal energy needed for heat	Continuous, min = 6300, max = 4650534, mean = 269362
	Heated volume area of building	Continuous, min = 156.2, max = 132080, mean = 8353.9
	H1TRND (max. allowed coefficient of transmission heat loss per surface),	Continuous, min = 0.4430, max = 1.5500, mean = 0.7472
	Q1HNDDOP (allowed yearly energy needed for heat [kWh/m^{3}a])	Continuous, min = 16.42, max = 30.40, mean = 22.20
	Electric power of heaters	Continuous, min = 0.1820, max = 1.5470, mean = 0.5803
	Thermal power of heaters	Continuous, min = 0, max = 599119.0, mean = 1361.7
	Total heating power	Continuous, min = 0, max = 6000.0, mean = 397.8
Temperature attributes	Internal project temperature	Continuous, min = 16, max = 28, mean = 20.35
	Internal project temperature in cool season	Continuous, min = 18, max = 22, mean = 20.52
	Internal project temperature in heat season	Continuous, min = 22, max = 24, mean = 22.38
Occupational data	No. of working hours per workday	Continuous, min = 8, max = 24, mean = 12
	No. of working days per week	Continuous, min = 2, max = 7, mean = 5.802
	No. of employees per m^2	Continuous, min = 0.07, max = 0.5, mean = 0.03422
	No. of users per m^2	Continuous, min = 0.1820, max = 6.64, mean = 0.1994
Output variable	Q1HNDREF - specific energy consumption	Continuous, min = 15.02, max = 415.95, mean = 39.13

Table 2
Sampling procedures used for modeling.

Subsample	% and no. of cases	
Learning subsample	70 % of total sample: 402 cases	
	80 % for training: 322 cases	20 % of cross-validation: 80 cases
Test subsample	30 % of total sample: 173 cases	
Total	100 %: 575 cases	

3.2. Methods

In order to design the architecture of the machine learning based system for managing energy efficiency of public sector, we have integrated the methods of data collection and preprocessing as well as machine learning methods.

3.2.1. Methods of data collection and preprocessing

The Big Data collection methods in the suggested architecture consist of three types of procedures: (a) transferring constructional, energetic, geospatial, static occupational attributes of each public building from EMIS information system to the Big Data collection, (b) collecting energy consumption data and dynamic occupancy data from IoT network - automatic readings on energy consumption using SCADA and IoT network of sensors and other smart devices, and (c) collecting environmental data from the web (air temperature, wind speed, air pressure etc.). The above procedures implemented in each building of the public sector at a national level enable creation of a large Big Data collection stored in a cloud with high volume, variety, and velocity.

The veracity attribute of this collection need to be achieved by further data preprocessing methods that consist of the following steps:

(1) Algorithm for detecting anomalies or errors in data entry and their corrections – see Algorithm for constructional characteristics data cleansing of large-scale public buildings database (Krstić & Teni, 2018).
(2) Algorithm for data preprocessing of public buildings (Scitovski et al., 2018), including:
  a outlier elimination
  b missing values replacement
  c variable reduction

All preprocessing steps were performed in conjunction with clustering procedure. Missing values in each input attribute were replaced by the mean value of the remaining values in that attribute. Variable reduction was conducted by using correlation coefficients for continuous variables and the chi-square test for categorical input variables. Preprocessed in that way, the data were ready to be used in the next level for predictive modeling. For processing multimedia data generated by IoT sensors, cameras, and other devices, there are the methods for searching, browsing, and visualizing multimedia data (see Grierson et al., 2015).

3.2.2. Methods of predictive modeling

The machine learning methods were selected to be used in this paper due to their nonlinearity and ability to learn from historical data. Zekić-Sušac, Knežević, and Scitovski (2019) has shown that energy consumption in public buildings does not fulfill the assumptions of linearity. Three machine learning methods were used for creating predictive models of energy consumption and efficiency of public buildings: artificial neural networks (deep ANNs) and recursive partitioning methods such as CART decision trees, and random forest (RF).

Classical artificial neural networks (ANNs) as well as deep ANNs, or DNNs have shown a high accuracy in solving prediction, classification and association problems (Prieto et al., 2016; Torres et al., 2018). The multilayer perceptron (MLP) suggested by Werbos in 1974 and improved by Rumelhart et al. (Masters, 1995) is the most frequently used ANN with a multilayer structure, typically with one input layer, a hidden layer and an output layer. DNN enables to add a number of hidden layers, and is more appropriate for complex data with a large number of input variables and cases. In the case of one hidden layer, the basic computation consists of a summation function which sums weighted inputs from the input layer units, and an activation function which computes the output of the hidden layer by using a linear or a nonlinear function. The computation can be summed into:

$$y_c = f\left(\sum_{i=1}^{n} w_i x_i\right) \tag{1}$$

where y_c is the computed output, x_i are the elements of the input vector X, w_i are the elements of the weight vector W (values of the weights are initially randomly determined from the interval [-1,1] and later adjusted by the error term) and n is the number of hidden nodes in the layer. In this paper, DNN was used with more hidden layers (2, 3, and

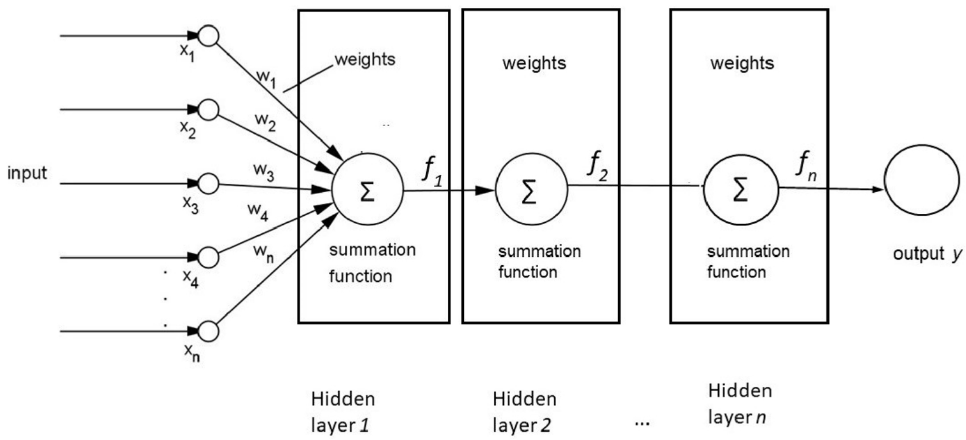

Fig. 1. DNN architecture with three hidden layers (Zekić-Sušac et al., 2019).

4) in R-software tool, with Keras library. Weight regularization was conducted by a dropout method (Srivastava, 2014) which enables better generalization, the dropout rate was 0.1. The number of hidden units in each hidden layer was optimized by a cross-validation procedure, where min. no. of hidden units was selected as the random number from the interval (2,100). The sigmoid activation function was applied in each hidden layer (Masters, 1995), and linear function in the output layer. The optimization algorithm in each hidden layer was the Adam algorithm with learning rate of 0.001 and maximum 200 training epochs with an early stopping Kingma and Ba (2014). A graphical representation of the DNN model with three hidden layers is given in Fig. 1.

The accuracy of all DNN models was evaluated by using the symmetric mean average percentage error (SMAPE) (3) according to Tofallis (2015):

$$SMAPE = 100\frac{1}{n}\sum_{i=1}^{n}\frac{|y_t - y_c|}{|y_t| + |y_c|} \tag{2}$$

where y_t is the target (real) output, while y_c is the computed output and n is the number of observations in the sample. In order to enable the comparison of different models' results, the same training set is used for the learning phase of all DNN architectures and decision trees, while the same hold-out test sample was used to finally evaluate the model performance. The training sample is furtherly divided into train and test subsample in order to determine the training time of DNNs by a cross-validation procedure. During that procedure, the neural network was trained on the training sample for a k number of iterations, and tested on the test subsample. The process is repeated if the error on the test subsample decreases, or the maximum number of iterations 100,000 is reached, and the process is stopped if the error starts to increase. By using the above parameters, two DNN models were created in order to select the most accurate one. DNN Model 1 used all available input variables, while DNN Model 2 was created on the reduced set of input predictors extracted by the feature selector FSelector library in R. In both models, the architectures with 2 and 3 hidden layers were tested. The results were compared by the statistical t-test, also conducted in R software tool. All the models used normalized data in the training and test phase, while the data were de-normalized before computing the SMAPE.

Besides DNN, Recursive Partitioning and Regression Tree (Rpart) and Random forest (RF) methods were additionally used to create predictive models of energy consumption. In the Rpart model, the ANOVA method was used to assess the importance of each input variable during the modelling phase. The basic algorithm was suggested by Breiman, Friedman, Ohlsen, and Stone (1984) named Classification and Regression Tree (CART). The algorithm builds a tree by splitting the input vectors at each node according to a function, and the parent node is divided into child nodes by separating the objects with values lower and higher than the split point with the highest reduction of impurity. Since the output variable in this research is continuous, the regression variant of the CART was used. The response for any observation in the regression tree is computed by following the path from the root node down to the appropriate terminal node of the tree. The values for the splitting variables are observed, and the predicted response value is calculated by averaging response in that terminal node (Hothorn, Hornik, & Zeileis, 2006). The splitting process is repeated for all input variables until the tree reaches its maximum size. Then the tree is pruned backwards using the cross-validation procedure to select the right-sized tree. All possible tree splits were evaluated by the Gini index according to Apté and Weiss (1997):

$$Gini(t) = 1 - \sum\nolimits_i p_i^2 \tag{3}$$

where t is a current node and p_i is the probability of class i in t. Since previous research has shown that Rpart has some limitations, such as the biasness regarding the variable types, (Grömping, 2009), another type of recursive partitioning, the random forest was additionally used.

The random forest (RF) is a tree partitioning method that generates a set of decision trees by randomly dividing the data into subsets, and randomly generating sets of candidate variables. The splits within each tree are created based on a random subset of candidate variables (Hartshorn, 2016). The final response is obtained by averaging responses of the individual trees. The main advantage of the RF method is in its stability comparing to single-tree techniques, while the main limitations are in its complexity and computing time (Grömping, 2009). In its basic form it uses the CART algorithm in each tree, which was used in our experiments with maximum depth 30, the complexity parameter cp = 0.01, and ANOVA method. Splits that do not improve the fit (R-squared) by cp will be pruned off by cross-validation.

Using the above three machine learning methods, the models for predicting the energy consumption were obtained and their accuracy is compared on the hold-out test sample.

4. Results and discussion

4.1. Results of the machine learning models for predicting energy consumption

The results of machine learning models are presented in Table 3. The DNN models were developed with all available 82 input variables (Model 1), and with a reduced set of variables obtained by rank correlation prior modeling (Model 2), and with a reduced set of variables selected by the RFE method (Model 3). The reduced set of input variables included the following attributes: Annual thermal energy needed for heat, Heated volume area of the building, Maximal allowed annual

Table 3
Machine learning prediction models of energy consumption in public buildings.

Machine learning model	Inputs	Machine learning method and parameters	NRMSE and SMAPE on the test sample
DNN Model 1	82 predictors described in Table 1	2 hidden layers, number of hidden nodes: 75-59, sigmoid function, Adam optimization algorithm	NRMSE = 0.1329, SMAPE = 15.4873 %
		3 hidden layers, number of hidden nodes: 87-99-94, sigmoid function, Adam optimization algorithm	NRMSE = 0.1345,SMAPE = 15.9967 %
DNN Model 2 DNN – reduced by correlation	10 input predictors selected by rank correlation	DNN 2 hidden layers, number of hidden nodes: 71–87, sigmoid function, Adam optimization algorithm	NRMSE = 0.1391, SMAPE = 15.5725 %
		DNN 3 hidden layers, number of hidden nodes: 14-31-28, sigmoid function, Adam optimization algorithm	NRMSE = 0.1888, SMAPE = 21.1026 %
DNN Model 3 – reduced input by RFE function	5 input predictors selected by RFE algorithm	DNN 2 hidden layers, number of hidden nodes: 57–63, sigmoid function, Adam optimization algorithm	NRMSE = 0.1529, SMAPE = 17.1482 %
		DNN 3 hidden layers, number of hidden nodes: 61-70-69, sigmoid function, Adam optimization algorithm	NRMSE = 0.1579, SMAPE = 18.1593 %
Decision trees, all available variables used, selection by tree pruning	7 predictors used in pruned tree construction	RPart decision tree, method = ANOVA, *cp* = 0.001, pruned tree with 9 terminal nodes	NRMSE = 0.1604, SMAPE = 17.4028 %
	15 most important variables identified by RF	Random forest tree-partitioning, number of trees: 500, no. of variables tried at each split: 27	NRMSE = 0.0989, SMAPE = 13.5875 %[a]

[a] The most accurate prediction model.

thermal energy needed for heat, F0 factor of building shape, and Max. allowed coefficient of transmission heat loss per surface. The decision tree models, such as Rpart, CTREE and Random forest have also used all available variables, but each method has selected only a subset of predictors during the process of tree pruning. When observing the prediction accuracy of the models, it can be seen from Table 3 that the most successfull DNN model is Model 1 with all avaialble inputs, producing NRMSE of 0.1329, and SMAPE of 15.4873 %. This accuracy was obtained by the architecture of 2 hidden layers, 75 hidden nodes in the first, and 59 in the second hidden layer, with sigmoid function, and Adam optimization algorithm. The DNN Model 2 has a slightly higher SMAPE of 15.5725 %, while DNN Model 3 has SMAPE of 17.1482 %. It reveals that DNN have sufficient complexity to deal with a high input dimension. If the results of the decision trees are observed, it can be seen that the most accurate model is the random forest which produces the highest overall accuracy of all tested models yielding NRMSE of 0.0989, and SMAPE of 13.5875 %.

A graphical representation of the Rpart decision tree is given in Fig. 2 showing the predictor variables used to split the tree nodes and their interval values used for splitting from the root node to the final terminal nodes, where *n* determines the number of cases in the training sample that belong to each node.

A comparison of predictors selected by each of the most successful machine learning models is given in Table 4. It reveals that DNN Model 2 as well as the Rpart and Random forest model have selected the Type of object as an important geospatial variable, while RPART has additionally selected the *County*. Among the construction group of variables, all four model have extracted *Heated volume area of building*, while the DNN Model 3 and Random forest model have extracted also *H1TRND* (max. allowed coefficient of transmission heat loss per surface), *Q1HNDDOP* (allowed yearly energy needed for heat [kWh/m^{3}a]). The *Thermal power of heaters*, *Electric power of heaters* as well as the *Total heating power* were additionally selected by the DNN Model 2 and Random forest. The DNN Model 2 has also selected variables related to internal temperature. The occupational variable *No. of working hours per week* was selected by the three models, while additional predictors of occupation, such as *No. of employees per m*^{2}, and *No. of users per m*^{2} were extracted by the most accurate Random forest model.

The developed machine learning models are incorporated into the design of the architecture of the suggested intelligent system for energy management MERIDA.

4.2. Proposed architecture of the intelligent system for energy management

The suggested architecture of the intelligent system for energy management MERIDA has its foundation on the previous research of Tomšiö et al. (2015) and Marinakis and Doukas (2018). Due to the fact that our aim was to create an intelligent energy management system that will include all buildings in the public sector, and can be used on both local and national levels, we have modified and extended their models such that all system components are designed at micro and macro level with the aim to increase the efficiency of the individual buildings, as well as of the whole public sector. Also, the IoT smart building network of public buildings is used as a foundation for creating a Big Data collection, which can serve as a platform for predictive modeling and decision making when the existing system is extended

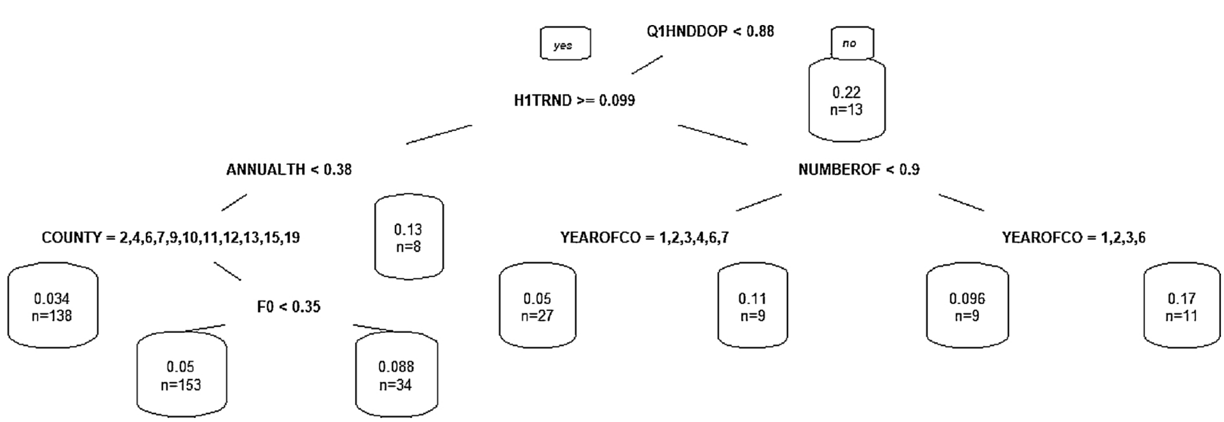

Fig. 2. Plot of pruned Rpart tree for predicting energy consumption.

Table 4
Comparison of predictors selected by machine learning models.

Group of attributes	DNN Model 2	DNN Model 3	RPART model	Random forest model
Geospatial attributes			County, Type of object	Type of object
Construction attributes	F0 (shape factor of the building)	F0 (shape factor of the building)	F0 (shape factor of the building), Year of completion of construction	F0 (shape factor of the building), Year of completion of construction
Energy attributes	Heated volume area of building, H1TRND (max. allowed coefficient of transmission heat loss per surface), Q1HNDDOP (allowed yearly energy needed for heat [kWh/m^{3}a]), Thermal power of heaters	Annual thermal energy needed for heat, Heated volume area of building, H1TRND (max. allowed coefficient of transmission heat loss per surface), Q1HNDDOP (allowed yearly energy needed for heat [kWh/m^{3}a])	Annual thermal energy needed for heat	Annual thermal energy needed for heat, Heated volume area of building, H1TRND (max. allowed coefficient of transmission heat loss per surface), Q1HNDDOP (allowed yearly energy needed for heat [kWh/m^{3}a]), Electric power of heaters, Total heating power
Temperature attributes	Internal project temperature, Internal project temperature in cool season, Internal project temperature in heat season			
Occupational data	No. of working hours per workday, No. of working days per week		No. of working days per week	No. of working days per week, No. of employees per m^2, No. of users per m^2

with IoT devices which substantially increase the amount of generated data. The proposed architecture of the intelligent system for energy management of public buildings MERIDA is presented in Fig. 3.

It can be seen from Fig. 3 that the architecture consists of six main levels of data processing which can be derived after data gathering from IoT Smart Building network in the public sector. The main components of the architecture are:

(1) Big Data collection level – based on data from EMIS database and IoT Smart Building network in the public sector. The EMIS database provides data presented in Table 1 that can be transferred regularly to a Big Data collection in the cloud. The IoT Smart Building network consists of sensors, SCADA devices, and other smart devices located in the public buildings to collect information on real-time energy consumption and dynamic occupational data. The Hadoop Spark platform is suggested for that purpose.
(2) Data preprocessing level - prepares data for prediction models using the methods described in previous section. The statistical packages R or Python can be used in this level.
(3) Machine learning prediction level – uses machine learning methods, such as ANNs, RF, SVM and others to predict energy efficiency, the consumption of each energy resource (natural gas, electricity, water, etc.) and CO2 emission. The models also extract important predictors that could be used for decision making. The TensorFlow platform with deep learning libraries for Python and R is suggested as a tool in this level.
(4) Data interpretation and visualization level – In this level, the results from machine learning models are presented to users using dynamic charts and tables, incorporated by a number of libraries in R and Python (or additionally in Tableau or other visualization tools)
(5) Decision making level – suggests decisions to be made by users (building managers, city government, and sector-level managers), and consists of: (A) decisions on energy power regulations, (B) decisions on investments in reconstruction measures.

Those decisions can be supported by the system using the rule-based decision tree.

(6) Benefits level – measures the benefits from the system usage, which can be expressed on micro level for each building (such as improved energy efficiency, reduced energy cost, improved energy conservation, reduced CO2 emission), as well at the public sector level in assisting decision makers in allocating resources to reconstruction measures. Decisions on investments in reconstruction imply short and long term benefits related to energy efficiency of the whole sector, cost savings, and healthier environment. On the strategic level such system can contribute the creation of national energy policy, and also decisions on EU and global level.

The IoT Smart Building network of public sector can be designed using ontology modeling. In Fig. 4 the ontology-modeling approach has been used to address information gathering from IoT sensors. Model in this paper is based on Building Topology Ontology (bot), DogOnt - Ontology Modeling for Intelligent Domotic Environments (dog), Sensor, Observation, Sample, and Actuator (sosa) ontology and OpenSmartHomeData repository. Building Topology Ontology is ontology for defining sub-components of a building (Rasmussen, Pauwels, Hviid, & Karlshoj, 2017).

DogOnt ontology has been used for defining home automation system and related appliances, and is one of the first ontologies for smart homes (Bonino & Corno et al., 2008). Sensor, Observation, Sample, and Actuator ontology provides interaction between subcomponents of a building and their entities involved in gathering data Janowicz, Haller, Cox, Le Phuoc, & Lefrançois, 2018). OpenSmartHomeData represents data set repository of measured data from a smart home which was used in this research as a reference point for

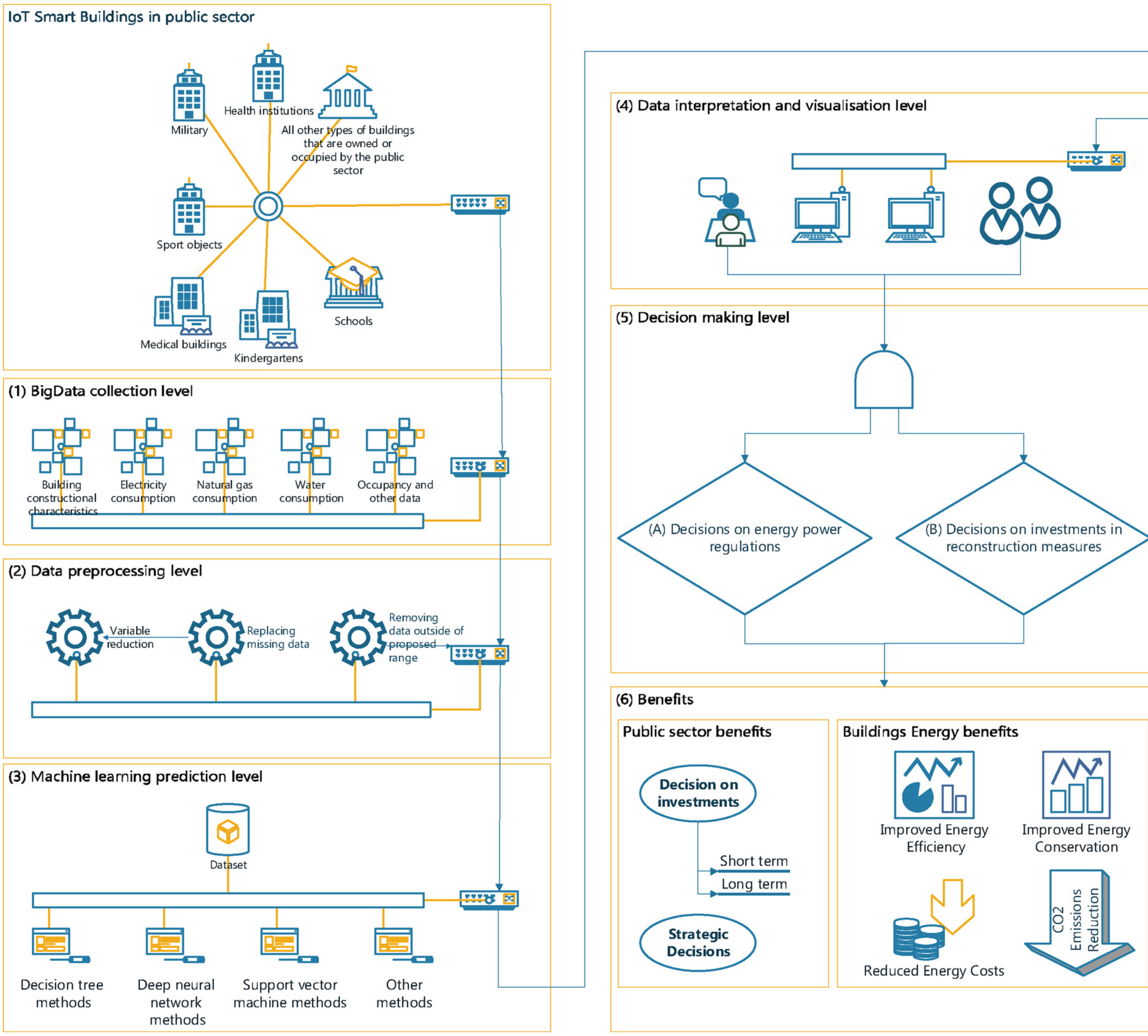

Fig. 3. Architecture of the machine learning based system for managing energy efficiency of public sector.

building ontology model framework with IoT sensors (Schneider & Rasmussen, 2018).

In this paper ontology model framework bot:Storey and bot:Space are used as sub-components for defining main relationships in building. Objects dog:Workspace, dog:Kitchen and dog:Bathroom are part of bot:Space and they provide interoperation mechanisms that are adding intelligence in the home automation system. Every object has implemented IoT sensors acting as the feature of interest (sosa:FeatureOfInterest) for observation, actuation and other data fetching policies. Presented model framework has the special combined ability of data distribution, remote access (over the internet) sensors with automation and machine learning intelligence. When applied on each building in the public sector, such model will enable automatic collection of dynamic occupancy information in each room in the Big Data collection and could be used to control devices of lighting, heating, and other energy devices in order to reduce the energy consumption. The IoT network in each building is to be connected with other public buildings in the city, therefore enabling the smart city concept, as well the concept of smart public sector within a state.

5. Discussion

The results of this research are two-fold: (1) in creating predictive models based on machine learning methods that show a potential of those methods in modelling energy consumption of public sector buildings, and (2) in suggesting an architecture of the machine learning based system for managing energy efficiency of public sector which can be incorporated into a larger system of a smart city. The previous research in the area of smart city were mostly focused on the structural determination of the system, by identifying important policies, challenges, barriers, or other elements of a smart city concept, such as the work of Simonofski et al. (2019), Wu and Chen (2019), Dwivedi et al. (2019), Israilidis et al. (2019), and Janssen et al. (2019). As majority of authors emphasize the role of AI in the smart city, rare papers discuss and suggest concrete AI models and their integration into an intelligent decision system of a smart city concept. Therefore, the focus of this paper was on the technology-based structure of a smart city system. Following the guidelines of previous authors, such as Jangili and Bikshalu (2017) who identified three layers of an IoT system, we have suggested a more complex architecture of the whole intelligent system for decision making in a smarty city containing six layers which include installing IoT in public buildings as a prerequisite, (1) Big Data collection level, (2) data processing level, (3) machine learning prediction level, (4) data interpretation and visualization level, (5) decision making level, and (6) benefits measuring level. The focus of this paper was to dive deeper into the machine learning prediction level, in order to provide concrete models for decision makers in the area of energy consumption and management. Furthermore, this paper is an extension

Fig. 4. Example of the ontology modeling of IoT Smart Building network in public sector.

of the work of Marinakis and Doukas (2018) whose focus was to produce daily and weekly action plans for the end users of each building. Our approach aimed to provide an intelligent support for decision makers on the local and sector level to facilitate decisions on investments in reconstruction measures that will yield more savings for the whole public sector. The paper also extends the research of Tomšić et al. (2015) who suggested and used only linear methods to make predictions on energy consumption of public buildings without presenting its accuracy of prediction. Our research uses more advanced machine learning methodology, and can be considered as an upgrade of previous authors' work in the sense of more advanced methodology and also in the sense of suggesting the architecture that includes IoT and Big Data as the technological basis. Therefore, the machine learning methods such as deep neural networks, Rpart regression tree, and Random forest are tested in this research to create an efficient prediction model of energy consumption in the public sector. It is important to emphasize that the models are created on real data collected from the Croatian energy management information system (EMIS) which manages energy data of public buildings in the country and gathers a large number of attributes including geospatial, constructional, energetic and occupational data.

The most accurate model was produced by the Random forest method yielding NRMSE of 0.0989, and SMAPE of 13.5875 %, although all three tested methods: DNN, Rpart tree and Random forest have produced SMAPE below 20 % showing a potential of all three machine learning methods in predicting energy consumption. Also, a comparison of predictors extracted by all three models has shown that they overlap in a substantial number of selected features, especially in selecting attributes related to power related to heat, internal temperature, and occupational attributes.

Besides creating prediction models, the paper also suggests the architecture of the intelligent energy management system named MERIDA, with six main layers which shows that the IoT network, Big Data platform, and machine learning predictive models can be integrated to provide support to decision makers on local and state level. The suggested six-level architecture is modular (each level can be developed and maintained individually, although it is connected to other levels. Due to the modularity, other modules can be added, such as the module of renewable energy sources and measurement of their energy production. The proposed architecture is also scalable, i.e. it can be used by a single building, or a number of buildings on the city-level or sector-level, and usable (the usability is assured by the data interpretation and visualization level that will take care of user experience). In addition, the paper provides an example of the DogOnt ontology modeling of IoT Smart Building network in public sector and OpenSmartHomeData repository based on the guidelines of Janowicz et al. (2018) and Schneider and Rasmussen (2018).

The MERIDA architecture differs from the architectures suggested in previous research in a way that it is more focused to the public sector buildings and it specifies the usage of Big Data platform and machine learning methods applied in predictive modeling that could be used for energy management.

5.1. Theoretical contributions

The research contributes the theory of smart city development by providing models that could be used in the data processing level to predict and manage energy consumption of public buildings with an acceptable accuracy. The most accurate model created in this research was obtained by leveraging the strengths of three different machine learning methods : deep neural networks, RPart decision trees, and random forest. The tests have shown that the random forest algorithm produces the highest accuracy and can be suggested as the most appropriate machine learning method for predicting energy consumption of public buildings.

Furthermore, the theoretical contribution is also in the proposed architecture of the intelligent machine learning based system for managing energy efficiency of public sector with the suggested six layers that include all steps from the installation of IoT in public buildings, Big data collection using cloud computing, data processing, machine learning prediction, data interpretation and visualization level, decision making level, and the benefits level which measures the benefits from the system usage, both on micro level for each building, and the macro level for the whole public sector. The suggested architecture is more detailed than the ones provided in previous research. For instance, the work of Jangili and Bikshalu (2017) identified three layers of an IoT system, but did not include the models and methods that could be used to process the data and provide a decision support based on an IoT system. Marinakis and Doukas (2018) suggested prediction models focused on daily and weekly action plans for the end users of each building, while our research creates models and their possible generalization on the macro level. Some authors have provided linear prediction models (Tomšić et al., 2015), and did not include IoT, Big Data collection, and more advanced machine learning methods. Another theoretical contribution of this paper is in extracting the features that are the most relevant for predicting energy consumption. Such selected set of features can be used by decision makers to put more focus on those attributes and plan reconstruction and other measures in order to reduce energy consumption in the public sector, and transform it into an energy-efficient system.

5.2. Implications for practice

The suggested models for predicting energy consumption show a potential for implementation in real systems for intelligent energy management on the level of individual buildings management, as well as on the macro level of managing energy consumption in the public sector. The saved weights of the models enable their usage in practice by decision makers who will be required to enter the values of input attributes of a new building and observe the produced output – a predicted value of a building's yearly energy consumption. The new buildings that were not included previously into the training set, can be later added into the trainig set and used for retraining the model. Besides their possible usage for predicting energy consumption, the models can be used for planning reconstruction measures of buildings by putting focus to the most important predictors of energy consumption. Investing in those predictor attributes of buildings can significantly decrease their energy consumption and cost. Furthermore, the analysis of extracted attributes of buildings can be used by decision makers to set priorities while investing in reconstruction measures by selecting buildings and their characteristics that will be most influential for reducing energy consumption. If practically used for the above purposes, the models can server as a support for creating national policies of reducing the consumption of non-renewable energy sources which implies lower emission of CO_2 and more healthier environment, which are the important guidelines in EU energy directives.

In addition, the paper suggest an architecture of the intelligent energy management system of a smart city, with created prediction models as an integral part of that architecture. The architects of a smart city could use that architecture as a guideline in creating an intelligent system that will be able to interconnect buildings into a network and process their energy data using advanced machine learning techniques. An example of ontology for interconnected rooms with IoT sensors in a smart building is given by using OpenSmartHomeData according to Janowicz et al. (2018) and Schneider and Rasmussen (2018).

5.3. Limitations and future research direction

There are several limitation of the conducted research that are planned to be overcomed in future research. The final sample size of buildings used to train and test the machine learning models was relatively small, which was unexpected due to the fact that the initial dataset contained more than 17,000 public buildings in EMIS system of Croatia. The reason for the necessary reduction of sample was in a large number of missing data and human errors in data entry. In order to overcome this issue, the automatic software warnings in case of out-of-range entries is needed, as well the automatization of data entry whenever it is possible. So far, the attributes that are entered automatically by a device are the energy consumptions of buildings, while other attributes are entered manually. The authors have established a collaboration with the agency that manages the EMIS system in order to increase the validity of data. Another limitation is in the accuracy of the prediction models that should be furtherly increased by testing other machine learning algorithms and enlarging the data sample.

Since the implementation of the suggested system is to be realized in the near future, the assessment of increased energy efficiency and savings as well other expected benefits will be possible in the further research. The future research will include the development of prediction models for water consumption and CO_2 emissions, and the measurement of implementation effects.

6. Conclusions

The paper deals with the issue of energy efficiency of the public sector, creates machine learning models for predicting energy consumption, and proposes the architecture of an intelligent machine learning based energy management system for public sector that could be used as a part of the smart city concept. The data are collected from two sources: central database of public sector buildings from EMIS system, and the IoT network of public buildings. Three machine learning methods were used for modelling energy consumption: DNN, Rpart and Random forest. The methods of outlier elimination, missing data replacement, and variable reduction are suggested based on previous research. The results show that the most accurate model on validation data was the Random forest, which has produced the SMAPE of 13.5875 % showing a potential of machine learning methods in energy management in the public sector. All three methods have extracted a similar set of important features for predicting energy consumption, most of them belonging to the group of energy power used for heating, internal temperature, and occupational data. In addition, the paper suggests an architecture of the intelligent system, which consists of six levels starting from the Big Data collection level which gathers, stores, and search the data, followed by the level where the data are preprocessed to be used by predictive modeling. Such system can include the created machine learning models, and extend them to predict energy efficiency level, consumption of natural gas, electricity, and other energy resources. After obtaining predictions, the level for data interpretation and visualization assists users in making decision on future actions that will enable benefits on the micro level of each building as well on the macro level of the whole public sector. The benefits level serves to measure benefits as improved energy efficiency, reduced energy consumption and reduced energy cost. The suggested system named MERIDA integrates all levels and enables their synergy into a managing platform for improving energy efficiency of the public sector within Big Data environment. The ontology-based model of IoT network of a public building is also suggested that could be used to determine the installation of sensors and other devices that will be used to capture dynamic occupancy of each room in the building, and send data to the Big Data collection. Such collection of buildings data can be connected with other parts of city and integrated within a smart city. The suggested model shows a potential for an intelligent energy management on both micro and macro level that embraces the possibilities of machine learning and Big Data platforms.

Declaration of Competing Interest

None.

Acknowledgments

This work was supported by Croatian Science Foundation through research grant IP-2016-06-8350 "Methodological framework for efficient energy management by intelligent data analytics" MERIDA.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ijinfomgt.2020.102074.

References

Apté, C., & Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer Systems, 13*(2-3), 197–210.

Bonino, D., Corno, F., et al. (2008). DogOnt – Ontology modeling for intelligent domotic environments. In A. Sheth (Vol. Ed.), *The semantic web – ISWC 2008. ISWC 2008. Lecture notes in computer science: 5318*, (pp. 790–803). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-88564-1_51.

Breiman, L., Friedman, J., Ohlsen, R., & Stone, C. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Calvillo, C. F., Sánchez-Miralles, A., & Villar, J. (2016). Energy management and planning in smart cities. *Renewable and Sustainable Energy Reviews, 55*, 273–287. https://doi.org/10.1016/j.rser.2015.10.133.

Cocchia, A. (2014). Smart and digital city: A systematic literature review. In R. Dameri, & C. Rosenthal-Sabroux (Eds.). *Smart city. How to create public and economic value with high technology in urban space. progress in IS* (pp. 13–43). Cham: Springer. https://doi.org/10.1007/978-3-319-06160-3_2.

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... Galanos, V. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002 In Press, Corrected Proof. Available online 27 August 2019.

European Commission (2014). *Annexes to the Commission regulation implementing directive 2009/125/EC of the European parliament and of the council with regard to ecodesign requirements for ventilation units.* Brussels, 7.7.2014, https://ec.europa.eu/transparency/regdoc/rep/3/2014/EN/3-2014-4517-EN-F1-1-ANNEX-1.Pdf (Accessed 1 September 2019).

Galicia, A., Talavera-Llames, R., Troncoso, A., Koprinska, I., & Martínez-Álvarez, F. (2019). Multi-step forecasting for big data time series based on ensemble learning. *Knowledge-Based Systems, 163*, 830–841.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Grierson, H. J., Corney, J. R., & Hatcher, G. D. (2015). Using visual representations for the searching and browsing of large, complex, multimedia data sets. *International Journal of Information Management, 35*(2), 244–252. https://doi.org/10.1016/j.ijinfomgt.2014.12.003.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician, 63*(4), 308–319.

Has, A., & Zekić-Sušac, M. (2017). Modelling energy efficiency of public buildings by neural networks and its economic implications. In L. Zadnik Stirn, M. Kljajić Borštnar, J. Žerovnik, & S. Drobne (Eds.). *Proceedings of the 14th international symposium on operations research in Slovenia* (pp. 461–466). Slovenia: Bled.

Hartshorn, S. (2016). *Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners.* Amazon.com.

Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., & Chiroma, H. (2016). The role of big data in smart city. *International Journal of Information Management, 36*(5), 748–758.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics, 15*(3), 651–674.

Ismagilova, E., Hughes, L., Dwivedi, Y. K., & Raman, K. R. (2019). Smart cities: Advances in research—An information systems perspective. *International Journal of Information Management, 47*, 88–100. https://doi.org/10.1016/j.ijinfomgt.2019.01.004.

Ismagilova, E., Hughes, L., Rana, N., & Dwivedi, Y. (2019). Role of smart cities in creating sustainable cities and communities: A systematic literature review. *International Working Conference on Transfer and Diffusion of IT* (pp. 311–324).

Israilidis, J., Odusanya, K., & Mazhar, M. U. (2019). Exploring knowledge management perspectives in smart city research: A review and future research agenda. *International Journal of Information Management*101989. https://doi.org/10.1016/j.ijinfomgt.2019.07.015 In Press. Corrected Proof. Available online 19 August 2019.

Jangili, S. R., & Bikshalu, K. (2017). Smart grid administration using big data and wireless sensor networks. *International Journal of Advance Research in Science and Engineering, 6*(10), 629–636 2017.

Janowicz, K., Haller, A., Cox, S. J. D., Le Phuoc, D., & Lefrançois, M. (2018). SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics, 56*, 1–10. https://doi.org/10.1016/j.websem.2018.06.003.

Janssen, M., Luthra, S., Mangla, S., Rana, N. P., & Dwivedi, Y. K. (2019). Challenges for adopting and implementing IoT in smart cities. *Internet Research, 29*(6), 1589–1616. https://doi.org/10.1108/INTR-06-2018-0252.

Kingma, D. P., & Ba, M. (2014). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations*https://arxiv.org/abs/1412.6980.

Krstić, H., & Teni, M. (2018). Algorithm for constructional characteristics data cleansing of large-scale public buildings database, High Performance and Optimum Design of Structures and Materials III. *WIT Transactions on the Built Environment, 175*, 213–224.

Lepenioti, K., Bousdekis, A., Apostolou, D., & Mentzas, G. (2020). Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management, 50*(2020), 57–70. https://doi.org/10.1016/j.ijinfomgt.2019.04.003.

López-Robles, J. R., Otegi-Olaso, J. R., Porto Gómez, I., & Cobo, M. J. (2019). 30 years of intelligence models in management and business: A bibliometric review. *International Journal of Information Management, 48*, 22–38. https://doi.org/10.1016/j.ijinfomgt.2019.01.013.

Marinakis, V., & Doukas, H. (2018). An advanced IoT-based system for intelligent energy management in buildings. *Sensors, 18*(2), 610. https://doi.org/10.3390/s18020610 1-16.

Masters, T. (1995). *Advanced algorithms for neural networks, a C++ sourcebook.* New York: John Wiley & Sons, Inc.

Odyssee-Mure (2016). *Key indicators.* Accessed 3 January 2019 http://www.indicators.odyssee-mure.eu/online-indicators.html.

Papa, R., Gargiulo, C., & Zucaro, F. (2014). Urban systems and energy consumptions, a critical approach. *TeMA journal of land use, mobility and environment, eighth international conference INPUT*https://doi.org/10.6092/1970-9870/2552.

Pérez-Chacón, R., Luna-Romera, J., Troncoso, A., Martínez-Álvarez, F., & Riquelme, J. (2018). Big data analytics for discovering electricity consumption patterns in smart cities. *Energies, 11*(3), 683.

Prieto, A., Prieto, B., Martinez Ortigosa, E., Ros, E., Pelayo, F., Ortega, J., & Rojas, I. (2016). Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing, 204*, 242–268. https://doi.org/10.1016/j.neucom.2016.06.014.

Rana, N. P., Luthra, S., Mangla, S. K., Islam, R., Roderick, S., & Dwivedi, Y. K. (2019). Barriers to the development of smart cities in Indian context. *Information Systems Frontiers, 21*(3), 503–525. https://doi.org/10.1007/s10796-018-9873-4.

Rasmussen, M. H., Pauwels, P., Hviid, C. A., & Karlshoj, J. (2017). Proposing a central AEC ontology that allows for domain specific extensions. *Proceedings of the joint conference on computing in construction (JC3)* (pp. 237–244). . https://doi.org/10.24928/JC3- 2017/0153 Vol. 1.

Schneider, G. F., & Rasmussen, M. H. (2018). *Technical building systems/open smart home data: First release of open smart home data set (Version v1.0.0).* Zenodohttps://doi.org/10.5281/zenodo.1244602.

Scitovski, R., Zekić-Sušac, M., & Has, A. (2018). Searching for an optimal partition of incomplete data with application in modeling energy efficiency of public buildings. *Croatian Operational Research Review, 9*(2), 255–268. https://doi.org/10.17535/crorr.2018.0020.

Simonofski, A., Vallé, T., Serral, E., & Wautelet, Y. (2019). Investigating context factors in citizen participation strategies: A comparative analysis of Swedish and Belgian smart cities. *International Journal of Information Management*102011. https://doi.org/10.1016/j.ijinfomgt.2019.09.007 In Press. Corrected Proof. Available online 3 October 2019.

Srivastava, N. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*, 1929–1958.

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *The Journal of the Operational Research Society, 66*(8), 1352–1362. https://doi.org/10.1057/jors.2014.103.

Tomšić, Ž., Gašić, I., & Čačić, G. (2015). Energy management in the public building sector – ISGE/ISEMIC model. *Journal of Energy: Energija, 64*(1), 1–4.

Tonković, Z., Mitrović, S., & Zekić-Sušac, M. (2018). Business intelligence system for managing natural gas consumption of public buildings. In R. Veselica, G. Dukić, & H. Hammes (Eds.). *36th International scientific conference on economic and social development – "Building resilient society" – EBS 2018, building resilient society* (pp. 769–778).

Torres, J. F., Galicia, A., Troncoso, A., & Martínez-Álvarez, F. (2018). A scalable approach based on deep learning for big data time series forecasting. *Integrated Computer-Aided Engineering, 25*(5), 1–14.

Wu, Y. J., & Chen, J. C. (2019). A structured method for smart city project selection. *International Journal of Information Management*101981. https://doi.org/10.1016/j.ijinfomgt.2019.07.007 In Press. Corrected Proof. Available online 31 July 2019.

Zekić-Sušac, M., Scitovski, R., & Has, A. (2018). Cluster analysis and artificial neural networks in predicting energy efficiency of public buildings as a cost-saving approach. *Croatian Review of Economic, Business and Social Statistics, 4*(2), 57–66. https://doi.org/10.2478/crebss-2018-0013.

Zekić-Sušac, M., Knežević, M., & Scitovski, R. (2019). Deep learning in modeling energy cost of buildings in the public sector. In F. Martinez Álvarez, A. Tronosco Lora, J. Sáez Muñoz, H. Quintián, & E. Corchado (Vol. Eds.), *Advances in intelligent systems and computing: 950*, (pp. 101–110). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-20055-8 2019.