The Design and Implementation of a Semantic-Based Proactive System for Raw Sensor Data: A Case Study for Laboratory Environments

Mehmet Milli[a], Özlem Aktaş[b], Musa Milli[c*], Sanaz Lakestani[d]

a Department of Computer Engineering, Faculty of Engineering, Bolu Abant Izzet Baysal University, Bolu, Turkey

b Department of Computer Engineering, Faculty of Engineering, Dokuz Eylul University, Izmir, Turkey

c Department of Computer Engineering, Turkish Naval Academy, National Defense University, Tuzla Istanbul, 34942, Turkey

d Scientific Industrial and Technological Application and Research Center, Bolu Abant Izzet Baysal University, Bolu, Turkey

Corresponding author: Assist..Prof. Dr. Musa MİLLİ
Address:
Department of Computer Engineering
Turkish Naval Academy
National Defense University
Tuzla, Istanbul, 34940
Turkey

E-mail: musamilli@gmail.com; mmilli@dho.edu.tr
Mobile Phone: +90 555 720 6744
Work Phone: +90 216 395 2630 / 3565

The Design and Implementation of a Semantic-Based Proactive System for Raw Sensor Data: A Case Study for Laboratory Environments

Abstract: In the last decade, raw sensor data from sensor-based systems, the area of use of which has increased considerably, pose a fundamentally new set of research challenges, including structuring, sharing, and management. Although many different academic studies have been conducted on the integration of sets of data emerging from different sensor-based systems until present, these studies have generally focused on the integration of data as syntax. Studies on the semantic integration of data are limited, and still, the area of study mentioned have problems that await solutions. In this article; parameters (CO_2, TVOC, CO, $PM_{2.5}$, PM_{10}, Temperature, Humidity, Light), affecting laboratory analysis results and threatening the analyst's health, were measured in laboratory environments selected as "use cases", and semantic-based information management framework was created for different sensor-based systems. Classical machine learning methods, and regression approaches which are frequently used for such sensor data, have been applied to the proposed sensor ontology and it has been measured that machine learning algorithm performs better on ontological sensor data. The most efficient algorithms in terms of accuracy and time were selected, through comparison of the obtained results, and integrated into the proposed proactive approach, in order to take the selected laboratory environment's condition under control.

Keywords: sensor ontology; semantic sensor web; machine learning; supervised learning; prediction on stream data.

1. Introduction

Although sensors are defined differently in many studies, the most common definition is that it is known as devices that detect phenomena in the physical environment in which it is located (Bermudez et al., 2009). In another definition, sensors are defined as devices that can convert chemical, physical, and biological values into digital values (Abd Hakim et al., 2018). Sensors have evolved continuously since the day they emerged and reached such a capacity that it can be utilized in almost every application, presenting efficiency in size, cost and adequacy. As a result of all these developments, sensor-based systems have become the heart of many electronic systems today. The use of such systems in many areas has caused an exponential increase in raw data on the Internet. A

demonstration of how the raw sensor data obtained from the sensor reaches consumers of data appears in Figure 1.

Most of the sensor data obtained from such systems on the Internet reach consumers without configuration. The unstructured presentation of sensor data causes a series of problems that include sharing, interpreting, and managing data. Moreover, the sensor data is heterogeneous in nature because it bears different syntaxes, structures, and meanings in different systems (Sheth, 1999). The heterogeneity of the sensor data causes these data to remain application-specific, and hinders the management of independent sensor-based systems under a common infrastructure. An intermediate layer, independent of the application, enabling the sensor data semantically enriched to make it more useful is of crucial need.

Recently, researchers argue that semantic sensor web technologies can enrich the raw data obtained from sensors semantically and fill this intermediate layer (Wang et al., 2017; Arooj et al., 2017; Haller et al., 2019; Liu et al., 2019). Besides, a common framework is required for sensor-based information systems. Sensor data should be defined using URIs and delivered to sensor data consumers over HTTP (Patni et al., 2010). In addition, sensor data should be encoded in formats that can be read by machines such as RDF and OWL so that they can be easily read and processed by machines. However, at this point, the lack of a comprehensive and understandable standard for the enrichment of sensor data around the world appears to be a major problem in the common manageability and operability of sensor systems.

The World Wide Web Consortium established the Semantic Sensor Network Incubator Group (SSN-XG) in 2011 to fill this intermediate layer and identified a set of standards for sensor data (Haller et al., 2017). It has conducted many studies and defined certain standards for the semantic enrichment of raw sensor data obtained from SSN-XG sensor-based systems. The latest version of the Semantic Sensor Network (SSN), which is still used as a common framework in many studies today, was published in 2017 (Barnaghi et al., 2017). The core of SSN forms a lightweight but independent core ontology called SOSA (Sensor, Observation, Sample, and Actuator), which holds basic classes and properties. SOSA complies with the minimum interoperability limits, i.e. the sensor ontologies created with SOSA guarantees its sharing and interoperability with all other SSN and SOSA ontologies. Conceptual modules forming the infrastructure of sensor-based systems such as deployment, system, platform, procedure, and etc. are defined in the framework of SOSA and SSN. Some basic conceptual modules of SOSA/SSN are shown in Figure 2.

The semantic sensor network is an application-independent framework that needs to be expanded with a certain concept and provides the manageability of the sensor systems on different platforms under a common infrastructure (Calbimonte et al., 2012). Shortly, SOSA/SSN is a model that allows the scope of the sensor ontology framework to be extended with other ontologies and concepts. For instance, in a biosensor application planned to be created in the field of medicine, a medical ontology, specific to the related field, including the technical medical terminology, classes, object properties and data properties can be employed to expand the ontological framework of SOSA-SSN.

A domain ontology that includes chemistry-related sensor measurements might import chemistry ontology, which includes chemical terminology (atomic number, orbital number, noble gas, element, etc.), classes, and object properties can be depicted as an example of the expansion of the SSN core ontology. The basic components of the SSN ontology are shown in Figure 3.

The proposed ontology for laboratory environment parameters that affect the results of laboratory analysis and threaten the analyst's health during the analysis includes the general basic SOSA/SSN main classes. Only a few classes have been added to the basic SOSA/SSN framework. The added classes are described in detail in Section 3.2.

There is more than one purpose within the scope of the study. The main objectives of the study are listed below.

- Establishing a common infrastructure with a high capability to represent raw sensor data. Moreover, ensuring semantic integration of sensor data with each other by using ontological concepts such as Class, Object Property, Data Property. Hence, providing the capability to manage data obtained from different platforms, different systems, and different sensors under a common framework.
- To establish a system that provides real-time monitoring and control of laboratory environment parameters that negatively affect the laboratory analysis results and threaten the analyst's health.
- Determining the best algorithm for the designated laboratory environment parameters by using classical machine learning algorithms on ontological sensor data. And accordingly, detection of unforeseen environmental situations thanks to the ontological based proactive system created, and avoiding unwanted situations by executing appropriate action plans in time.

The remainder of the article is organized as follows. In Section 2, previous studies in the field of sensor ontology are examined, and the differences between those and the current ongoing study are clearly revealed. Setting up systems infrastructure, creating sensor nodes, and use case are presented in Section 3. The data collection, the experiments to prepare data for the machine learning algorithm, and choosing the appropriate machine learning algorithms for the proposed sensor ontology are presented in Section 4. Section 5 describes the comparison of machine learning algorithms, determination of the most suitable algorithm in every aspect, and integration into the proposed proactive system. Finally, the results and future studies are discussed in detail in Section 6.

2. Related Work

The concept of sensor data ontology was first introduced by Avancha et al. (2004). Since 2004, many studies have been carried out in this field, and sensor ontology has become an area of study that attracts more attention. Considering the components (machine learning, semantic web technologies, wireless sensor networks) that form the basis of the proposed study, there are many studies in the literature. Therefore, it is possible to classify the literature review under 3 headings by selecting articles that are similar to this study.

The works in the first group focus on the integration of machine learning algorithms built on data from wireless sensor networks. In this category, studies focusing

on machine learning algorithms processing sensor data and excluding semantic enrichment approaches are argued. In this context, many studies have been administered in different domains in the last 20 years. These studies cover the applications of machine learning approaches in the field of health in (Chen et al., 2018; Hu et al., 2018; Rathore et al., 2017). In (Sarangdhar & Pawar, 2017; Patil & Thorat, 2016) there are studies in which machine learning approaches are applied in the field of environment and agriculture. In addition to these, machine learning approaches have been used in areas such as smart cities (Din et al., 2019), security (Patel & Jhaveri, 2015; Canedo & Skjellum, 2016; Kotenko et al., 2015) where WSN's are frequently utilized. Studies in this area are not assessed in detail, as they are a bit far from the proposed study. The major difference between these studies in the first group and the proposed study is that the sensor data collected is not enriched by using the ontological concepts. The best advantage of the proposed system is that it enables the management of ontologically oriented application-specific sensor-based systems before the emergence of the SOSA/SSN common framework.

In the second group, studies focusing on structuring sensor data to be managed under a common framework are considered. Although the semantic sensor web is beneficial in ensuring analytical integration between different sets raw data, the complexity of semantic techniques is often unacceptable for some end-users and data consumers due to the long processing time. The suggested system in (Bermudez-Edo et al., 2017) proposes IoT-Lite to reduce complexity and shorten transaction times. The IoT-Lite contains a simple example of semantic sensor ontology. The greatest feature of this sensor ontology is an approach that provides interoperability of sensor data on heterogeneous IoT platforms and includes minimum concepts and relationships that can respond to most end-user questions in a reasonable time. In the work mentioned in (Jin & Kim, 2018), a semantic sensor network has been used to solve interoperability problems of different platforms and devices in an e-health system. Apart from these, Kuster et al. (2020), Wang et al. (2018), Ali et al. (2017) proposed different semantically based architectures to describe sensor information collected from different environments.

In these studies, the focus is on the management of sensor data feeding from different systems under a common infrastructure. The major difference between these studies and the proposed study is that machine learning algorithms are not operated on the sensor data of which ontology is created. In other words, these systems only perform real-time monitoring in real-world applications. In the suggested system, one of the main objectives is to find the most suitable machine learning approach for the proposed ontological sensor system.

In the third group, the studies cover the application of machine learning approaches to semantic sensor data. The studies closest to the proposed study are examined in this group. The system proposed in (Adeleke et al., 2017) mentions a sensor ontology which is presented using the W3C's SSN frame. Adeleke et al. developed a statistical machine learning-based prediction model using this proposed sensor ontology. In the respected study, in order to predict an unhealthy situation in the near future, their models are evaluated on PM2.5 and PM10 values. 5 different classification algorithms

are applied to ontological sensor data in their studies. By comparing these algorithms, they claim that the most effective algorithm on PM values is the Multilayer Perceptron.

In the work mentioned in Onal et al. (2017), another semantic sensor web-based proactive system is presented. This system has been applied and evaluated for clustering and sensor anomaly detection using a public data set. In this study, the LinkedSensorData and LinkedObservationData dataset containing different weather parameters such as air temperature, wind speed, relative humidity, pressure, and visibility are used. LinkedSensorData is an RDF dataset that describes approximately 8000 air sensor information. The K-means algorithm, which is widely used for proactive systems in the literature, has been chosen as the appropriate model in this system.

The studies that are the most similar to the proposed study in terms of technology and scope are evaluated in this group. Studies under this category have also created a semantic-based framework for the definition of sensor information, and classical machine learning approaches have been performed on ontological sensor information. The main purpose of SSN is to create a common identification frame for sensor information from different platforms, different domains, and different sensors. However, in these studies, the number of platforms, sensors, and domains are limited and the capacity of SSN to represent sensor information in different systems, platforms, and domains could not be fully utilized. In the proposed study, 3 different environments, 4 different platforms, 5 different sensors are used and 8 different parameter values are measured. In previous studies, machine learning algorithms applied to ontological sensor data are limited in number, so in this study, the number of algorithms running on sensor data is increased. Another difference is that many studies focused on either regression or binary classification. In this study, regression and binary classification approaches are evaluated together.

3. Material and Methods

3.1. Sensor Nodes

In order to measure the values of parameters determined in the selected use case, 4 different nodes to perform 4 different tasks have been established. These sensor nodes are named Type A, Type B, Type C, and Type D and the purpose of installation and fundamentals components are given below. Arduino Uno is used as a microprocessor in all sensor nodes due to its ease of use and low cost. Considering transmission distance, energy consumption, and compatibility with Arduino Uno, the nrf24l0 + antenna is chosen as the communication device. In order to reduce the load on the nodes and to provide the flexibility of deployment during the distribution of the sensors in the environment, two different sensor nodes are installed, and the sensors are placed on them.

Type A Sensor Node (Gateway Node): The gateway node is the most important node in the network, as it is the one to collect the data and transmit to the base station. In cases where the Type A sensor node fails to function due to physical obstacles or any problem arising from its electronics, or if communication with other nodes is interrupted, all data communication in the network stops. Thus, the Type A sensor node is vital for

the system. No sensor was placed on it as no measurement in the environment is expected from it.

Type B Sensor Node (Sensor Node 1): In the proposed project, 5 different sensors are used to measure 8 parameters. These sensors are integrated into the two nodes, measuring an equal number of parameters. The DHT22 sensor, which measures the temperature and humidity parameters in the environment, and the CCS811 sensor that measures the carbon dioxide and total volatile organic compounds, are integrated on the Type B Sensor Node. The schematic design of the Type B sensor node, prepared with the Fritzing circuit modelling program, is shown in Figure 4.

Type C Sensor Node (Sensor Node 2): Another sensor node that makes measurements in the environment is the Type C sensor node. MQ-7 sensor measuring carbon monoxide, Nova SDS011 Sensor measuring PM2.5, and PM10 values, and light-dependent resistance (LDR) sensor measuring light intensity in the environment are integrated into this node. The schematic design of the Type C sensor node, prepared with the Fritzing circuit modelling program, is shown in Figure 5.

Type D Sensor Node (Repeater Node): After the nodes are installed in the measurement environment and WSN is established, a communication problem occurs due to the distance and obstacles between some nodes. In order to solve this communication problem and to ensure healthy data communication, repeater nodes are placed to strengthen the received signal and to enable the data received from the node to reach the gateway node. The sensors used, the nodes created, the technical infrastructure of this network, the characteristics, and detailed description of this system used are available in the previous study of the research team (Reference - Authors previous study).

3.2. Sensor Ontology

The SOSA/SSN provides an application-independent common framework that needs to be expanded with specific concepts and opportunity to manage sensor data for different domains. The concepts to be added can be classes, object properties, data property, or individuals depending on the application. In the proposed project, the core SSN ontology for the ontology of laboratory environment parameters is expanded by adding some classes, object properties, and individuals. This ontology is designed with the Protégé (Musen, 2015) ontology editor developed by Stanford University. Protege is a free open source framework that provides an interface for users to review ontologies. The Protege 5.5.0 editor has the capability to create classes and subclasses, define and visualize the relationship between classes to extend the SSN ontology.

Since this article focuses more on seeking the most appropriate machine learning approaches on ontological sensor data for proactive system design, the creation of sensor ontology is not explained in detail. Technical information on the proposed sensor ontology is available in the article previously written by the project team (Reference - Authors previous study). SSN core sensor ontology has been expanded to represent the environmental parameters that affect the analysis performed in the laboratory environment used and the indoor environment parameters that affects the health of the analyst. This extension includes appropriate classes, object properties, data properties, and instances. The following example is given in order to better understand the proposed

sensor ontology. In the core SSN ontology, the most significant concept is the “sosa:Observation” class, as it represents the sensor value and measurement date and time with the data properties attached to it. Figure 6 below shows an example of an extended sensor ontology from the point of view of the “sosa:Observation” class in the proposed sensor ontology.

"sosa:Observation" is the indicator representing the value of the property of a "sosa:FeatureOfinterest", or computing through a "sosa:Procedure". The alghoritm connects to "sosa:Sensor", subclass of "ssn: System" class with "sosa: madeBySensor" object property, to understand what shapes "sosa:Observation". In the above illustration, Nova SDS011 sensor used in the project is given as an example. An individual of the class "sosa:Observation" measured by this sensor is shown in Figure 6. Each measurement is given a unique value consisting of 32 characters and represented by it. So that, data consumers can access each sensor value they want to display with this unique id. The PM2.5 value measured by the Nova SDS011 value is "xsd:double" 7.73. As illustrated, the measurement date and time are "xsd:dateTime" 2019-08-30T06: 00: 00 + 03: 00. Since there is a good number of “unit”s for the same or different parameters in the literature, the "MeasurementUnit" class has been added to the basic SOSA/SSN framework to avoid unit complexity. The unit of the value measured by the Nova SDS011 sensor given in the example is assigned as "PartsPerMillion", which is frequently used in the literature.

Looking at the other concepts in the given example, to explain which parameter is measured by "sosa:Sensor" class , a link is given to "sosa:ObservableProperty", which is a sub-class of "sosa:Property" class, with "sosa:observes" object property. In this example, it is seen that the parameter measured by Nova SDS011 sensor, which is a member of the “sosa: Sensor” class, is PM2.5 “sosa:isFeatureOfInterestOf" object property is given as a link to "sosa:FeatureofInterest" classes to explain to which environment the value "sosa:Observation" is associated. "sosa:FeatureOfInterest" class is the area or environment where you want to measure. In this study, 3 laboratories that are frequently used in SITARC have been selected as the measurement area. One of them is the AoxMercury laboratory where various analyses are carried out. To summarize the example given above, in the AoxMercury measurement area, the value measured by the Nova SDS011 sensor on the AoxMercury13 platform at 06:00 a.m. on 30.08.2019 is 7.73 ppm.

In the proposed sensor ontology, each individual of the "sosa:Observation" class is expressed with 10 triples as shown in Figure 7. In this example, in the MaldiTof measurement area, the value of the CO2 parameter measured by the CCS811_22 sensor at 12:00 pm on 30.08.2019 is 1087,62 ppm.

3.3. Use Case

The proposed sensor ontology is created using the sensor data collected in the Scientific Industrial and Technological Application and Research Center (SITARC) within the Bolu Abant Izzet Baysal University (BAIBU). Data collection has been carried out in 3 laboratories frequently used in SITARC. These laboratories are MaldiTof, AoxMercury, and Chromatography laboratories. In these laboratories selected as Use Cases,

microorganism identification, proteomic analysis, bacteria count, fatty acid analysis, anion-cation determination, total halogen determination, solid-phase extraction, etc. analyses are done frequently.

During analyses, both the environmental parameters that will affect the analyst's health and the environmental parameters that will affect the analysis result must be monitored instantaneously in order to be kept under control. According to the report of the World Health Organization (WHO) one of the most important causes of disease and death in the world is an unhealthy living environment. Therefore, avoiding unhealthy conditions and monitoring the working environment effectively to keep the environmental parameters under control emerges as a serious issue.

In this study, a total of 8 parameters: temperature, humidity, carbon dioxide (CO2), total volatile compounds (TVOC), carbon monoxide (CO), particular matter 2.5 (PM2.5), particular matter 10 (PM10), and light intensity are measured by 5 sensors. For this, a total of 8 sensor nodes, including 1 Type A, 3 Type B, 3 Type C, and 1 Type D nodes, are established and deployed to measurement environments. 1 Type B and 1 Type C sensor nodes are placed in every 3 laboratories selected as measurement areas, one for each sensor. Type A sensor node (Gateway) is placed in AoxMercury Laboratory because it is close to the midpoint of all nodes. Once the sensor network is established, it is realized that there are communication problems between the Gateway and Type C sensor node, due to distance and physical obstacles such as walls, tables, and devices in the Chromatography laboratory from time to time. This problem is solved by placing a Type D sensor node between these nodes and strengthening the signal.

Technical information and properties about sensors and sensor nodes are detailed in the previous study (Reference - Authors previous study). How the sensor nodes are distributed in the measurement environment is shown in Figure 8.

4. Experiments

4.1. Collecting Data

After placing the sensor nodes in the measurement area and sending the data properly, the data collection process is started on 29.08.2019 at 16:05. Each sensor in the installed system is programmed to measure an average per minute and send it to the gateway. The hourly average of the collected data is added to Apache Jena Fuseki, which is frequently used as a triple database (Apache Software Foundation). Jena Fuseki is a SPARQL server. In addition, it has been preferred as a triple database in this project as it provides a clear user interface for server monitoring and management.

The data collection process has been terminated on 12.10.2019 due to the annual maintenance of the devices in the laboratory. A total of 45 days of uninterrupted data has been collected at the selected measurement sites. Between these dates, each sensor has made approximately 62,000 measurements, and a total of approximately 1,500,000 measurements have been made. Theoretical and practical training have been given twice in the first 10 days of September and October in the laboratories specified between the dates of measurement, and the 3 laboratories where the measurement is made have been used. This situation has been beneficial for the project results in terms of observing what

kind of changes may occur in the parameters during the analysis and training in the laboratory. In Figure 9 and Figure 10, daily average values of CO2 and temperature values between the measurement dates are shown.

The graph in Figure 9 is given as a box-whisker plot to clearly show the central position and spread of the mean of temperature data collected. Although the low values of some parameters such as temperature during analysis have a positive effect on analysis studies, it negatively affects the health of the personnel, especially in long-term analyses. Especially in MaldiTof and Chromatography laboratories, the ambient temperature must be below 18 °C for a proper analysis activity to be carried out. However, considering the health of the personnel, it is important to keep the temperature in these laboratories within a narrow range. Although there are air conditioners, keeping the ambient temperature at appropriate levels that do not expose a threat on human health and not negatively affect the analysis results in laboratories, is more complicated than in other environments.

The graphic in Figure 10 shows the daily average CO2 level in the laboratories selected for the measurement area within the specified date range. Especially during the dates of theoretical and practical training, it is seen that the amount of CO2 in the environment exceeded the value of 1000 ppm determined by the WHO health organization as a reference value for indoor environments. It has been observed that the value of many parameters measured within the scope of this study increased during the dates of formal education. The reason for this increase is thought to be directly related to the amount of gas released as a result of the analysis performed in the experiments and increasing the human activity in the environment.

4.2. Pre-Processing and Data Manipulation

4.2.1. Determination of Classes

The accepted reference values of important parameters that determine indoor air quality such as CO2, CO, TVOC, PM2.5, PM10 have been determined by the institutions that are accepted worldwide such as WHO, EPA, ASHARE in the literature. In this study, these reference values are used while classifying and labelling the data. However, while determining the limit values of parameters such as temperature and humidity, the past experiences of researchers who made analyses in other research and laboratories have been used. Although the level of light, which is the last parameter measured, is effective in many laboratory processes such as bacterial growth, no data indicating its impact on indoor air quality has been recorded. Generated classes and their limit values are shown in Table 1.

In many respected studies, generally, one parameter and two different classes are used, such as “Good” and “Poor” (Adeleke et al., 2017). Since the overall purpose of this study is to find a suitable prediction algorithm for ontological sensor data, the situation for the algorithms to be selected is shaped to present a more complicated state; 5 different classes are defined for 7 parameters and the limit values are determined. The class label of an instance is identified by the parameter with the worst value of class among the parameters that make up that row. Table 2 shows how the class value of the row is determined.

When the instances are classified according to the aforesaid rule, it has been seen that 65% of the total of 3168 rows of data are at the desired level for the laboratory interior environment. However, in the remaining 35%, timely preparation of necessary action plans is vital for laboratory analysis results, and employee health. The experiments reveal that laboratory air quality is monitored at ideal ranges when there is no biological analysis and no human activity in the environment.

Certain pre-processes are required to make logical inferences and obtain good conclusions on the data collected. Pre-processes such as removing noisy data, conveniently filling missing data, shift all parameter values to the same range (normalization) are absolutely necessary for determining a better prediction model. Pre-process operations performed before making estimates on the data and how they are applied are explained in detail below.

4.2.2. Missing Value Imputation

On the specified dates, approximately 25,920 data would be expected to have been saved to the Apache Jena Fuseki RDF database, though only 23,252 data have been recorded due to the malfunction of the devices operating in the system or human error. This number corresponds to approximately 90% of the data that should be recorded. It is important to fill the missing values with a reasonable approach, especially if the algorithms in effect that are sensitive to missing values such as "Decision Tree" and "Random Forest" are to be studied. In this manner, the missing 10% has been filled with the well-known and accepted methods, and data continuity was ensured.

In data mining, it is possible to deal with the missing value issue with different approaches, such as deleting the missing values, accepting the average of that feature as the standard, or accepting them zero. Deleting or statistically filling missing values causes bias and negative effects on the result. Therefore, unlike these approaches, inputting data can significantly improve the quality of the data set (Yang et al., 2017). Recently, many studies have shown that filling missing values with classification approaches has positive effects on the output (Deb & Liew, 2016; Darryl & Rahman, 2016; Abidin et al., 2018). In our work, missing values are filled by utilizing a hybrid approach of the K-NN algorithm and Decision Tree, and the quality of the data set is increased.

4.2.3. Outlier Detection

An Outlier can be defined as any observation different from other observations in the data set (Barnett & Lewis, 1994). Outliers in the data collected by WSN can generally be caused by sensor measurement errors or some problems arising from data communication. Occasionally the outliers can be caused by human error. For example, if someone blows or touches the sensor in an environment where the temperature parameter is measured, this is a human error that causes the sensor value to deviate upwards. Both system-based and human-based errors cause the estimation to be biased and wrong. Therefore, analysing the collected data and eliminating some inconsistent parts will increase the accuracy of the prediction.

There are some types of outlier detection approaches such as Probabilistic, Distance Based (cosine, Euclidean distance, etc.), algorithm-based (neighbour based, neural networks based, etc.). We evaluated outlier detection in two stages. First, the outlier data in each attribute has been found in itself and eliminated. During this process, the cosine distance approach, which is one of the distance-based outlier detections measures, is used, and a total of 10 observation data inconsistent with the other data have been deleted from each column. In the second step, after the class label of each instance (row) is assigned, outliers have been determined over this class label and eliminated. While determining outliers, the K-NN neighborhood approach has been used (k =10) and a total of 20 observation data eliminated.

4.2.4. Normalization

The measurement ranges, limit values, of each sensor used in this study are different. The measuring range is the total range that the instrument can measure under normal conditions. Table 3 shows the maximum and minimum values that can be measured by the sensors used in this study.

Absolute distance measuring methods such as Euclidean and Minkowski consider features in the same value ranges in the similarity calculation with equal importance. When using such distance measuring methods, calculating the similarity between instances without any pre-processing on the data set causes the feature with a large variance to have a high effect on the result (Jain et al., 1999). In other words, the feature with large variance dominates the effect of other features on the result. It is called "feature domination". Moreover, the feature with high variance may not have a positive correlation with data in the same class, so it may not have the capability to parse data properly. In this case, the classification process might be misleading. To avoid feature domination; (i) all features are shifted to a certain interval. Normalization has significantly increased the performance of the classifiers used in this study. (ii) Cosine like similarity measures can be used that are not affected by the feature domination problem.

As demonstrated in Table 3, the values of some parameters can be between 0 and 100, while some parameter values may go up to 10,000. Therefore, it is certain that the prediction algorithms will decide according to the parameter with high values. In order to prevent this situation and to ensure that the parameters affect the estimation algorithm equally, all parameters have been shifted to the range of [0-1].

4.3. Integration of a Predictive Model into Proposed Sensor Ontology

Machine learning is a rising trend in the field of computer science, due to its capability to extract hidden features and patterns even in highly complex data sets. The performance of machine learning approaches may vary according to each situation, case study, and data set. However, the duration and difficulty of the training phase may also be different. Considering the factors that determine performance such as scalability, flexibility,

accuracy, precision, training time, and total working time, it is difficult to decide on which method is more efficient for a situation or data set.

To be able to say that a machine learning approach is the best for any data set, it must strike a delicate balance between performance, flexibility, scalability, and training time. For example, it is not possible to say that an algorithm with the highest accuracy is always the most appropriate approach. Although the accuracy rate of an algorithm is high, if the total running time in a system that has to respond in real-time is too long, it may not be suitable for the system to be installed.

Classical machine learning methods, which are widely used in the literature, have been applied to predict the near future on the proposed sensor ontology. The predictive methods are processed with the "Rapid Miner" tool on the sensor data collected and the results obtained are compared. As a result of the comparisons, performance criteria such as training time, total running time, accuracy, and gain were examined and the best machine learning approach has been chosen and integrated into the system. Algorithms performed on the proposed sensor ontology data are described in detail below.

Naive Bayes: The Naive Bayes technique takes its name from Thomas Bayes and his conditional probability theorem. It is one of the oldest supervised learning algorithms among machine learning methods. The most important features are the simple operation and speed. The algorithm accepts all variables as independent, but this assumption is rarely valid in the real world. Bayes theorem uses the probability function in Eq.1.

$$P(A/B) = \frac{\mathrm{P(B/A).P(A)}}{\mathrm{P(B)}} \quad (1)$$

where P(A) is the probability of the occurrence of event A and P(A/B) is the probability of the occurrence of event A, when event B occurs.

Generalized Linear Model (GLM): GLM is a method developed by John Nedler and Robert Wedderburn by combining various statistical models (Nelder, & Verrall, 1997). It is a flexible and generalized form of ordered linear regression that can classify regardless of the normal distribution of the dependent variable (Garrido & Zhou, 2009). Linear regression uses Eq.2, while GLM uses a similar formula as a link function.

$$\mathrm{Y} = \beta 0 + \beta 1 \mathrm{X} + \varepsilon \quad (2)$$

Linear regression consists of two parts; (i) the first two terms in the formula are structural components and (ii) the third term is ε (error term), which is also the random component.

Logistic Regression (logit): In logistic regression, as in other regression models, the aim is to establish a model with a certain number of variables and with an acceptable error rate. It is preferred for multivariate data, especially if the dependent variable is not continuous. The main difference between linear regression and logistic regression is that the value of the dependent variable in linear regression is estimated, whereas in logistic regression, the probability of realization of the values that the dependent variable can take is calculated. Therefore, logistic regression takes values between 0 and 1. In addition,

linear regression uses the Ordinary Least Squares (OLS) method for estimating, logit uses the Maximum Likelihood (MLE) method.

Fast Large Margin: Algorithms such as Support Vector Machines (Vapnik, 1995) that position the decision border in order to maximize the distance between two classes are called Large Margin algorithms. In other words, data estimated to belong to separate classes are mapped to have as clear a distance as possible. This type of linear classifiers complies with multidimensional data sets. The larger the natural margin between classes is, the higher the success of the classifier is.

Deep Learning (DL): Deep Learning is an emerging machine learning technique that has become popular recently. DL is closely related to the artificial neural network (ANN). ANN is the general name of the algorithms that learn to generalize the whole data set from a small data set by modelling the working principles of the human neural system and the brain. It contains mechanisms, like humans, that ensure making decisions on situations that they have not seen but based on their past experiences. In other words, ANN is used to model the relationship between input data and output data. DL is a more sophisticated and structurally more complex form of ANN. Because DL has more intermediate layers, learning is relatively slow compared to ANN and requires more processing power.

Decision Tree (DT): DT is one of the most common algorithms of machine learning and data mining. It is used in both classification and regression analysis. The most important feature of DT is that it simplifies and clarifies the decision-making mechanism in any process. It handles the decision-making process like a tree structure, naming the algorithm. It recursively divides the search space into subsets according to an attribute in each decision node. The division process ends when the data remaining in the subset cannot be separated according to any attributes. The lowest node of the tree specifies the classes. Establishing an optimal DT is usually an NP-Complete problem. Therefore, it requires applying heuristic ways to establish a good near-optimal DT (Hartmann & Varshney, 1982).

Random Forest (RF): RF (Breiman, 2001) is an ensemble learning algorithm that generates multiple decision trees and combines the results obtained from these decision trees with the bagging method. It is one of the popular algorithms used recently because RF can be applied to both regression and classification problems and also achieves good results in these areas. In addition, since the RF algorithm runs on different data sets for each feature, overfitting is reduced. The RF algorithm is preferred because it can employ the function of distinguishing the classes for each feature.

Gradient Boosted Trees: Boosting is a technique used to strengthen weak classifiers. Gradient Boosted Trees, like other techniques described earlier, is used both in regression analysis and classification but is also an ensemble technique that uses decision trees. As the name suggests, the ensemble technique uses the boosting approach, making the classification sequential rather than independent. Therefore, this technique tries to make better predictions using the mistakes of previous estimators. If the terminate criterion is not selected properly, overfitting may occur, unlike the ensemble bagging technique.

5. Experimental Results

The results of classification algorithms on the aforementioned data set are presented in this section with different aspects. In order to reveal the achievements of algorithms, they have been run on the collected data set and it has been evaluated that the testbed established as a real-life case is sufficient for a fair evaluation of the classifiers. The algorithm performance tests have been performed on a computer with windows 10 operating system and equipped with Intel I7 7700HQ 2.8 GHz processor, 16 Gb DDR4 Memory, Nvidia Geforce Gtx 1050 video card.

When algorithm performance tests have been enforced on ontological sensor data, 70% percent of the data has been divided into the training set, and 30% percent test set. 6 out of 9 machine learning algorithms evaluated are used with default parameter values. However, depths and the maximum number of tree parameters of RF, GBT, and DT algorithms negatively affect the time performances at their default values. Therefore, these parameters have been optimized for these algorithms, without much compromise on accuracy. The Maximal Depth and Number of Trees parameters are set to 10 in order to compete with other algorithms in terms of time.

All of the algorithms obtained acceptable accuracy values except NB and LR. But the most successful algorithms in terms of accuracy among them are RF, DL, and DT with the value 90%, 89%, 88% respectively. Therefore, it has been observed that these three algorithms are equally suited for this case. Generally, complexity and accuracy performance specify a trade-off in many cases, for this scenario the performance/complexity ratio of DT is better than others. The comparison of the accuracy percentages of the algorithms used in the case study is shown in Figure 13.

The results obtained in terms of time comparison of the algorithms can be seen in Figure 14. According to the results, we see that the most effective algorithms at the total time aspect are DT and GLM methods, respectively. The biggest reason underlying the high speed of DT is the fast decision-making mechanism thanks to its tree structure. Also, DT doesn't need a large training set to get good results. GLM is a regression-based method and it is obvious and known that regression-based methods are effective especially in terms of running time. So it is not surprising that DT and GLM achieve the fastest scores. However, DT, NB, and RF algorithms have shown a tendency to learn faster. For this reason, the training time of these algorithms is the lowest. In addition, the duration of time spent in a training set with 1000 records are observed in the time graph in Figure 14. According to this statistic, DT again gets the lowest score while DL gets the second place. This graphic demonstrates that the DL method has good scalability.

Figure 15 shows the average correlations calculated by all models between labels and attributes. According to these correlations, the most important parameters affecting the result is PM10, PM2.5, and Temperature. While it is predictable that PM10 and PM2.5 are active attributes, it is a surprise that the temperature is effective. However, the lectures in labs have increased the human presence and activity and the linear relationship of the temperature attribute with CO2 has been caused by this situation.

The results in Figure 15 revealed that the parameter of light does not have much effect on the results obtained however, it is an expected result. While setting the label

value of each row in advance, it has been thought that the parameter of light would not affect the result and it is stated that it is not used in defining the line label.

In addition to the run time and accuracy comparisons of the selected algorithms, the amount of gain and loss is also an important parameter in the selection of the algorithm, especially in multi-class labelling. In a multi-class dataset, more acceptable it is for a predicted value to be in a class close to the real value than if in a class far from the true value is. The benefits and costs of the wrong and correct estimates are given in Table 4. Losses are represented as negative numbers while benefits or gains are represented as positive numbers.

For example, if the label value of an instance with an actual label value of Excellent is estimated as Excellent with any classifier, the prediction is correct and takes 1 as the gain point. On the other hand, if the classifier labelled the same Excellent instance as a Good, Moderate, Poor, or Terrible the classifier takes -1, -2, -3, -4 loss point respectively and this prediction becomes wrong. These loss points give the value of the wrong prediction. In some cases, it may be more beneficial to choose the best performing algorithm by looking at gain rather than accuracy. An example of how Cost Matrix is used is shown in Table 5.

When the performances of the algorithms are compared in terms of gain, it is seen that the sum of the costs of Naive Bayes and Logistic Regression algorithms is negative, while the remaining algorithms are positive. When the performances of algorithms are assessed via gain metric, it is seen that the algorithms that give the best results in parallel with their accuracy rates are RF, DL, and DT. A comparison of algorithm performances in terms of gain is given in Figure 16.

6. Conclusions and Future Work

In recent years, sensor-based systems have rapidly spread to all areas of daily life as a result of the physical minimization of sensors in size, enabling the use in every field, the developments in the academic community, and the decrease in their prices. The intensive use of sensor and sensor based systems in every field has caused an exponential increase in sensor data in the internet environment. However, the heterogeneous nature of the sensor data makes it difficult to manage them under a single infrastructure. In addition, the absence of a common framework for the representation of sensor data makes it difficult for the machines to be understood and interpreted. Although a syntactic relationship has been established between sensor data in studies conducted so far, this is insufficient to make meaningful inferences from sensor data.

Semantic Sensor Web technology has been suggested and used by many researchers to address all these problems. Creating semantic relationships instead of establishing syntactic relationships between sensor data will provide more meaningful inferences. In addition, sensor data must be encoded in languages that machines can understand and interpret, such as RDF and OWL. Each sensor data should be represented by URIs and it should be easier for data consumers to reach it. In the first step of this study, a different model has been created by using the SSN framework to manage the data collected from different platforms, different environments, and different sensors under

the same infrastructure. In the second step of the proposed study, in order to establish a proactive system design, some traditional and state-of-art prediction algorithms on ontological sensor data are tested and compared by using data from this model. When the values obtained by running the algorithms on the collected sensor data are compared, it is seen that the most effective algorithms are RF and DT in terms of run time, accuracy, and gain.

The proposed model can be combined with different domains, different platforms, and different systems to expand its scope in future studies. With this extended model, sensor data can be used to make a common inference. Although the proposed sensor ontology associates the data semantically, the complexity of the semantic techniques often causes an increase in processing times. A new model that includes minimum concepts to ensure that the proposed semantic systems respond in a reasonable acceptable time to data consumers can be created. Object properties and data properties can be used within the scope of the minimum concept. Thus, the triple number in the RDF database is reduced and the system can be more efficient.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Scientific, Industrial and Technological Application and Research Center of Bolu Abant İzzet Baysal University for utilization of MaldiTof laboratory, AoxMercury laboratory, and Chromatography laboratory, as real-world use-case in proposed sensor ontology.

Data availability

All data of 8 measurements collected over 45 days using 5 different sensors from 3 different laboratories are in the link: 10.6084/m9.figshare.14742858

References

Bermudez, L., Delory, E., O'Reilly, T., & del Rio Fernandez, J. (2009). Ocean observing systems demystified. In *OCEANS 2009* (pp. 1-7). IEEE.,

Abd Hakim, S., Tarigan, K., Situmorang, M., & Sembiring, T. (2018, November). Synthesis of Urea Sensors using Potentiometric Methods with Modification of Electrode Membranes Indicators of ISE from PVA-Enzymes Coating PVC-KTpClPB. In *Journal of Physics: Conference Series* (Vol. 1120, No. 1, p. 012024). IOP Publishing.

A. Sheth, "Interoperating Geographic Information Systems," *Interoperating Geogr. Inf. Syst.*, pp. 5–30, 1999.

Wang, F., Hu, L., Zhou, J., Hu, J., & Zhao, K. (2017). A semantics-based approach to multi-source heterogeneous information fusion in the internet of things. *Soft Computing*, *21*(8), 2005-2013.

Arooj, M., Asif, M., & Shah, S. (2017). Modeling Smart Agriculture using SensorML. *IJACSA) International Journal of Advanced Computer Science and Applications*, *8*(5).

Haller, A., Janowicz, K., Cox, S. J., Lefrançois, M., Taylor, K., Le Phuoc, D., ... & Stadler, C. (2019). The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation. *Semantic Web*, *10*(1), 9-32.

Liu, J., Li, Y., Tian, X., Sangaiah, A. K., & Wang, J. (2019). Towards semantic sensor data: An ontology approach. *Sensors*, *19*(5), 1193.

Patni, H., Henson, C., & Sheth, A. (2010). Linked sensor data. In *2010 International Symposium on Collaborative Technologies and Systems* (pp. 362-370). IEEE.

Haller, A., Janowicz, K., Cox, S., Le Phuoc, D., Taylor, K., & Lefrançois, M. (2017). Semantic sensor network ontology.

Barnaghi, P., Compton, M., Corcho, O., Castro, R. G., Graybeal, J., Herzog, A., ... & Page, K. (2011). Semantic sensor network XG final report. *Recommendation REC-rdf-syntax-grammar-20040210, Online, http://www. w3. org/TR/rdf-syntax-grammar/, World Wide Web Consortium, Tech. Rep. XGR-ssn-20110628*.

Calbimonte, J. P., Jeung, H., Corcho, O., & Aberer, K. (2012). Enabling query technologies for the semantic sensor web. *International Journal On Semantic Web and Information Systems (IJSWIS)*, *8*(1), 43-63.

Avancha, S., Patel, C., & Joshi, A. (2004). Ontology-driven adaptive sensor networks. *UMBC Student Collection*.

Chen, M., Zhou, J., Tao, G., Yang, J., & Hu, L. (2018). Wearable affective robot. *IEEE Access*, *6*, 64766-64776.

Hu, L., Yang, J., Chen, M., Qian, Y., & Rodrigues, J. J. (2018). SCAI-SVSC: Smart clothing for effective interaction with a sustainable vital sign collection. *Future Generation Computer Systems*, *86*, 329-338.

Rathore, H., Al-Ali, A., Mohamed, A., Du, X., & Guizani, M. (2017). DLRT: Deep learning approach for reliable diabetic treatment. In *GLOBECOM 2017-2017 IEEE Global Communications Conference* (pp. 1-6). IEEE.

Sarangdhar, A. A., & Pawar, V. R. (2017). Machine learning regression technique for cotton leaf disease detection and controlling using IoT. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)* (Vol. 2, pp. 449-454). IEEE.

Patil, S. S., & Thorat, S. A. (2016). Early detection of grapes diseases using machine learning and IoT. In *2016 second international conference on Cognitive Computing and Information Processing (CCIP)* (pp. 1-5). IEEE.

Din, I. U., Guizani, M., Rodrigues, J. J., Hassan, S., & Korotaev, V. V. (2019). Machine learning in the Internet of Things: Designed techniques for smart cities. *Future Generation Computer Systems*, *100*, 826-843.

Patel, N. J., & Jhaveri, R. H. (2015). Detecting Packet Dropping Misbehaving Nodes using Support Vector Machine (SVM) in MANET. *International Journal of Computer Applications*, *122*(4).

Canedo, J., & Skjellum, A. (2016). Using machine learning to secure IoT systems. In *2016 14th annual conference on privacy, security and trust (PST)* (pp. 219-222). IEEE.

Kotenko, I., Saenko, I., Skorik, F., & Bushuev, S. (2015, May). Neural network approach to forecast the state of the internet of things elements. In *2015 XVIII international conference on soft computing and measurements (SCM)* (pp. 133-135). IEEE.

Bermudez-Edo, M., Elsaleh, T., Barnaghi, P., & Taylor, K. (2017). IoT-Lite: a lightweight semantic model for the internet of things and its use with dynamic semantics. *Personal and Ubiquitous Computing*, *21*(3), 475-487.

Jin, W., & Kim, D. H. (2018). Design and implementation of e-health system based on semantic sensor network using IETF YANG. *Sensors*, *18*(2), 629.

Kuster, C., Hippolyte, J. L., & Rezgui, Y. (2020). The UDSA ontology: An ontology to support real time urban sustainability assessment. *Advances in Engineering Software*, *140*, 102731.

Wang, C., Chen, N., Wang, W., & Chen, Z. (2018). A hydrological sensor web ontology based on the SSN ontology: A case study for a flood. *ISPRS International Journal of Geo-Information*, *7*(1), 2.

Ali, S., Khusro, S., Ullah, I., Khan, A., & Khan, I. (2017). Smartontosensor: ontology for semantic interpretation of smartphone sensors data for context-aware applications. *Journal of Sensors*, *2017*.

Adeleke, J. A., Moodley, D., Rens, G., & Adewumi, A. O. (2017). Integrating statistical

machine learning in a semantic sensor web for proactive monitoring and control. *Sensors*, *17*(4), 807.

Onal, A. C., Sezer, O. B., Ozbayoglu, M., & Dogdu, E. (2017). Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 2037-2046). IEEE.

Musen, M. A. (2015). The protégé project: a look back and a look forward. *AI matters*, *1*(4), 4-12.

World Health Organization. (2014). *Global status report on noncommunicable diseases 2014* (No. WHO/NMH/NVI/15.1). World Health Organization.

"Apache Software Foundation. 'Apache Jena.' A free and open source Java framework for building Semantic Web and Linked Data applications," 2011. [Online]. Available: https://jena.apache.org/documentation/fuseki2/index.html. [Accessed: 12-Oct-2019].

Yang, J. H., Cheng, C. H., & Chan, C. P. (2017). A time-series water level forecasting model based on imputation and variable selection method. *Computational intelligence and neuroscience*, *2017*.

Deb, R., & Liew, A. W. C. (2016). Missing value imputation for the analysis of incomplete traffic accident data. *Information sciences*, *339*, 274-289.

Darryl, N. D., & Rahman, M. M. (2016). Missing value imputation using stratified supervised learning for cardiovascular data. *J Inform Data Min*, *1*, 13.

Abidin, N. Z., Ismail, A. R., & Emran, N. A. (2018). Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *9*(6), 442-447.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. John Wiley & Sons, Chichester.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

Nelder, J. A., & Verrall, R. J. (1997). Credibility theory and generalized linear models. *ASTIN Bulletin: The Journal of the IAA*, *27*(1), 71-82.

Garrido, J., & Zhou, J. (2009). Full credibility with generalized linear and mixed models. *ASTIN Bulletin: The Journal of the IAA*, *39*(1), 61-80.

V. Vapnik, (1995) *The Nature of Statistical Learning Theory*. New York, New York, USA: Springer,Verlag.

Hartmann, C., Varshney, P., Mehrotra, K., & Gerberich, C. (1982). Application of information theory to the construction of efficient decision trees. *IEEE Transactions on information theory*, *28*(4), 565-577.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Table 1: Classes and Limit values to be used in this study

	Excellent(5)	Good(4)	Moderate(3)	Poor(2)	Terrible(1)
Temperature	19-21	18-19	17-18	16-17	<16
		21-22	22-23	23-24	>24
Humidity	40-60	30-40	20-30	10-20	<10
		60-70	70-80	80-90	>90
CO_2	<700	700-900	900-1100	1100-1300	>1300
TVOC	<40	40-70	70-100	100-150	>150
$PM_{2.5}$	<10	10-20	20-30	30-40	>40
PM_{10}	<20	20-40	40-60	60-80	>80
CO	<25	25-50	50-75	75-100	>100
Light	Nan	Nan	Nan	Nan	Nan

Table 2: Determining the class values of parameters and rows (color print).

Temperature	Humidity	CO_2	TVOC	$PM_{2.5}$	PM_{10}	CO	Light	Nominal
22.93	54.16	534.55	20.86	10.66	12.85	27	74.63	Moderate
23.01	53.78	541.1	21.68	10.09	11.83	27	67.1	Poor
21.03	42.12	422	2.48	0.88	1.12	21.6	26	Good
20.99	42.2	417.45	1.71	1.32	1.38	21	4	Excellent
20.27	50.94	879.46	71.31	5.08	5.78	32.59	78.07	Moderate
20.31	50.94	554.24	23.08	4.67	5.73	32.8	76.56	Good
20.25	52.34	1348.59	142.37	7.58	8.96	37.28	28	Terrible
20.31	52.3	1223.55	128.47	7.79	9.22	34.65	28	Poor
19.66	52.25	1306.33	138.5	6.53	7.71	255.35	79.43	Terrible
19.59	55.33	407.04	0.28	3.42	3.73	22.57	26	Excellent

Table 3: Value ranges of measured parameters.

No	Sensor	Parameter	Unit	Measurement Range
1	DHT22	Temperature	°C	-40 °C-125 °C (± 0.5)
2	DHT22	Humidity	% rh	0%-100% {± 2.5-5}
3	CCS-811	Carbon Dioxide	ppm	400-29206 ppm
4	CCS-811	Total Volatile Organic Compounds	ppb	0-32768 ppb
5	Nova PM	Particular Mattter 2.5	ppm	0.0-999.9 ppm
6	Nova PM	Particular Mattter 10	ppm	0.0-999.9 ppm
7	MQ-7	Carbon Monoxide	ppm	10-10.000 ppm
8	LDR	Light Level	%	0%-100%

Table 4: Cost matrix referenced when comparing the gain performance of algorithms used.

Cost Matrix	True Terrible	True Poor	True Moderate	True Good	Ture Excellent
PredictedTerrible	1	-1	-2	-3	-4
Predicted Poor	-1	1	-1	-2	-3
Predicted Moderate	-2	-1	1	-1	-2
Predicted Good	-3	-2	-1	1	-1
Predicted Excellent	-4	-3	-2	-1	1

Table 5: Example of cost matrix use.

Status	Status_Prediction	Cost	Accuracy
Excellent	Teribble	-4	FALSE
Excellent	Poor	-3	FALSE
Excellent	Moderate	-2	FALSE
Excellent	Good	-1	FALSE
Excellent	Excellent	1	TRUE

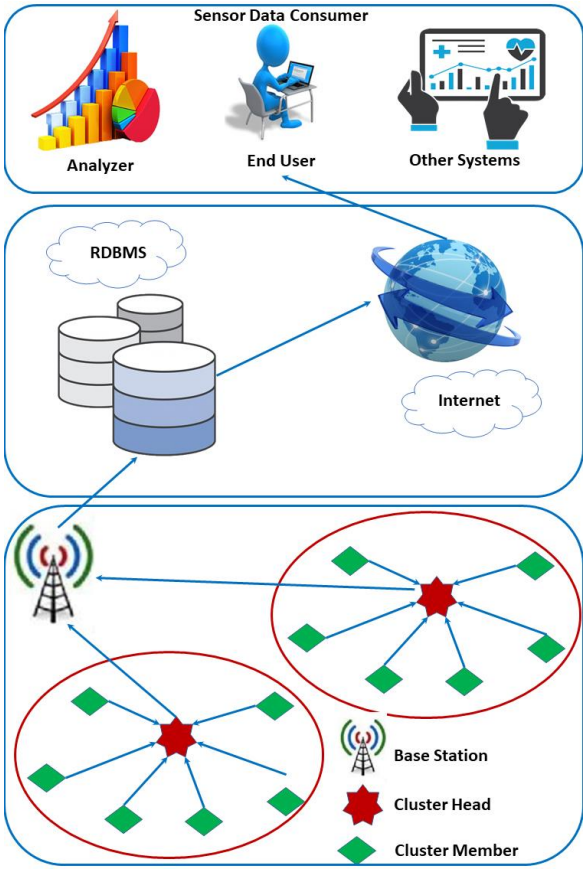

Figure 1: The simple structure of a sensor-based system

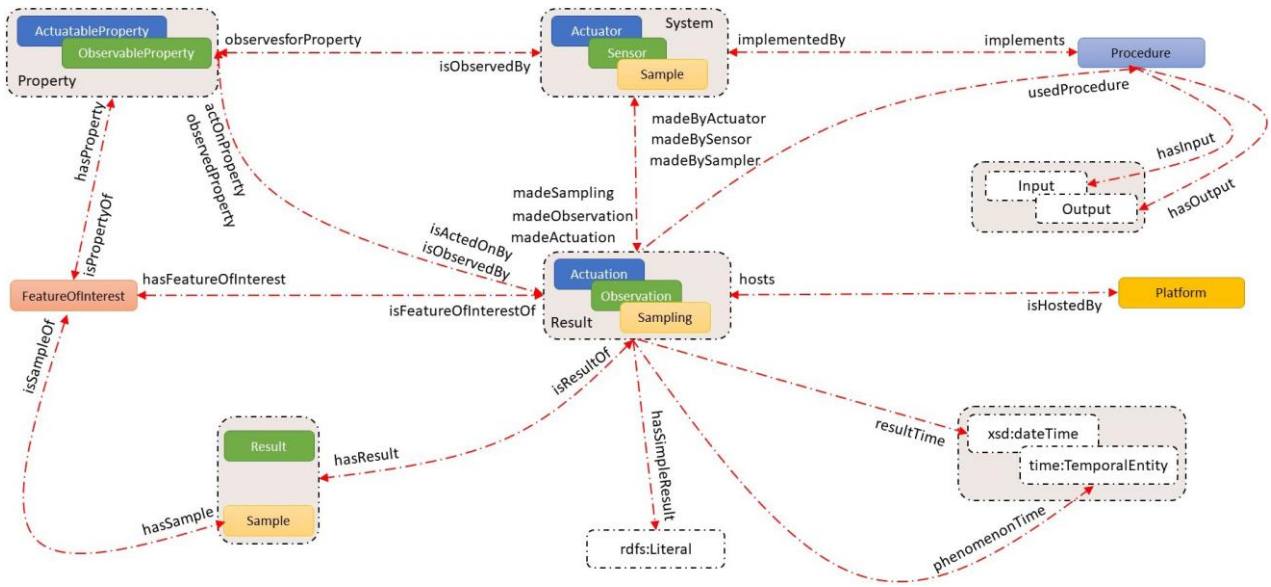

Figure 2: Overview of the core structure of the SOSA classes, object properties, and data properties.

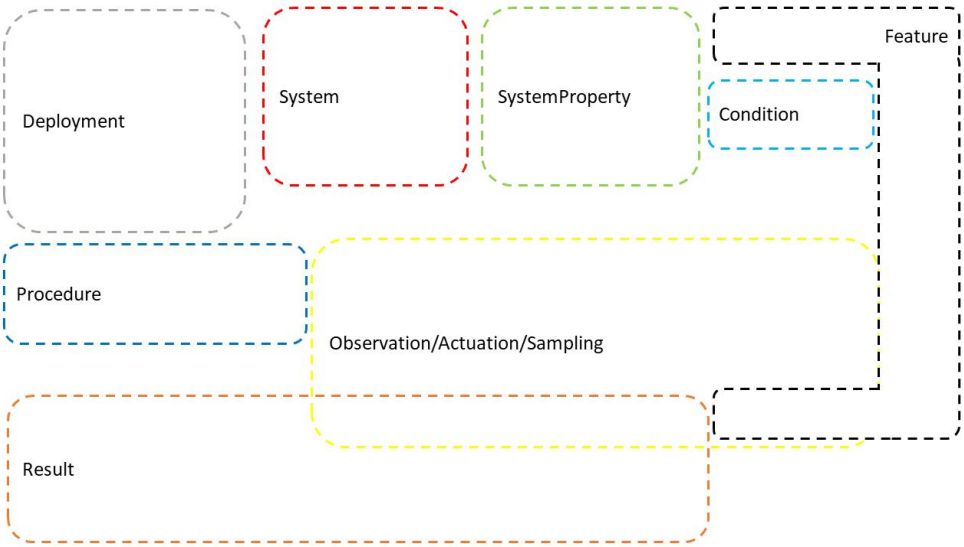

Figure 3: Basic conceptual ontology modules of SOSA/SSN frameworks.

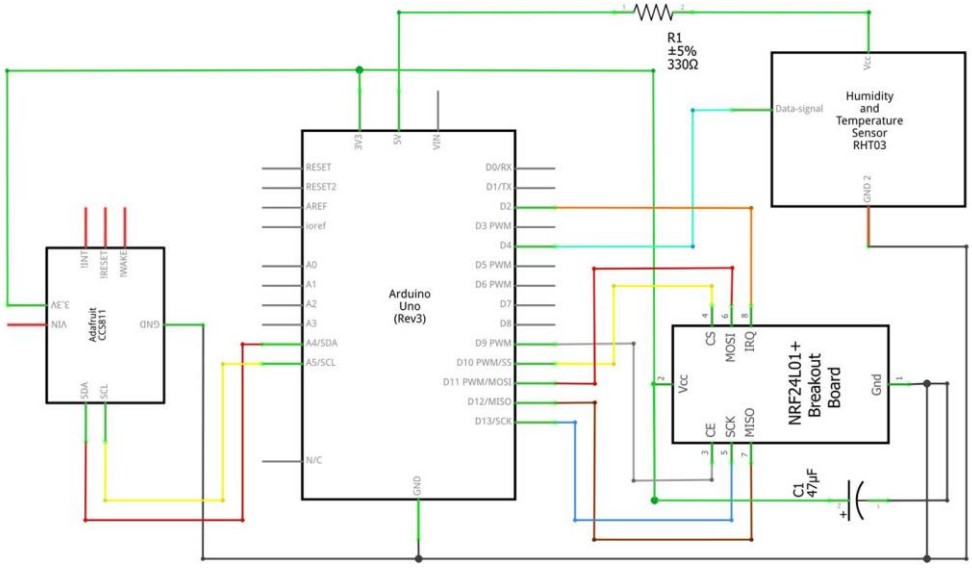

Figure 4: Fritzing-drawn circuit modelling of a Type B sensor node

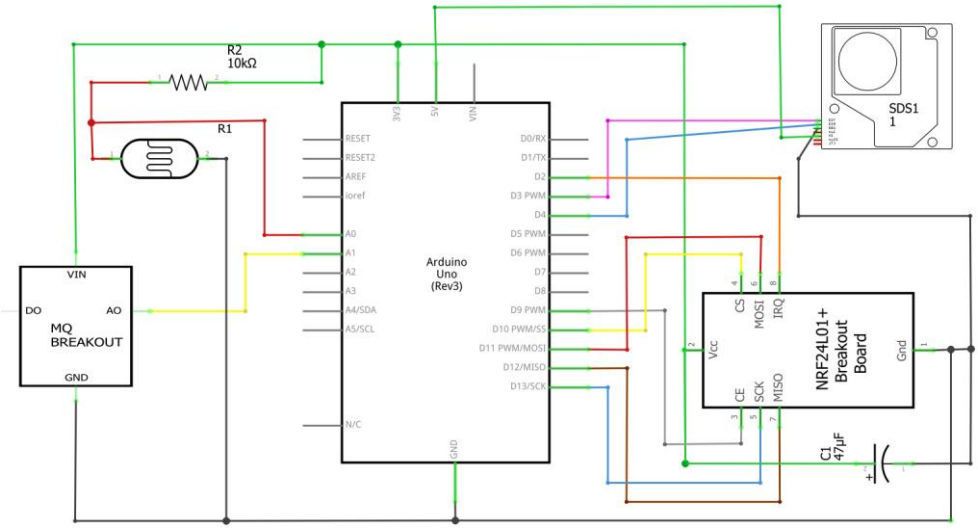

Figure 5: Fritzing-drawn circuit modeling of a Type C sensor node.

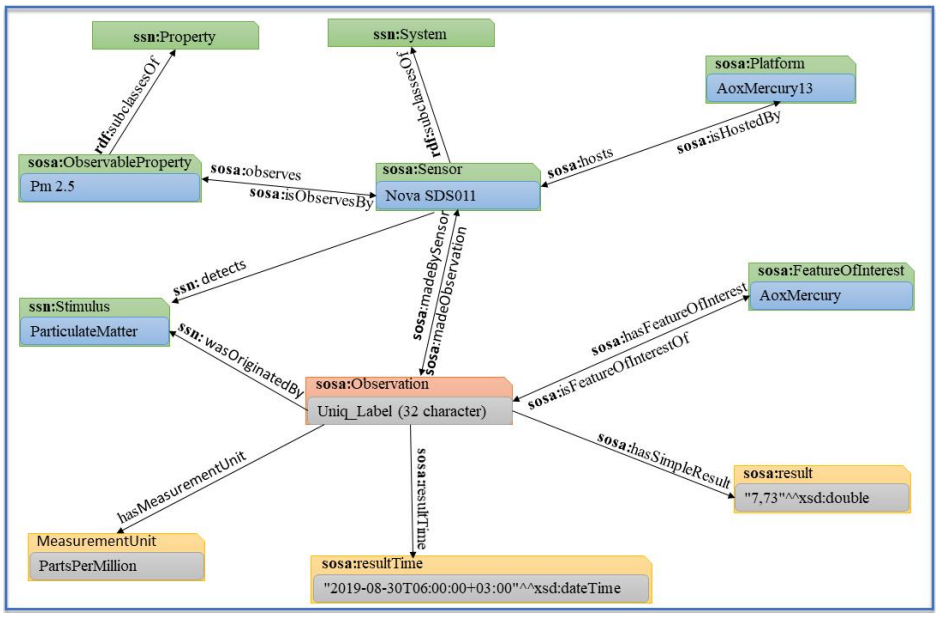

Figure 6: Overview of some SOSA/SSN classes and properties from the sosa:Observation class perspective.

```
ssn:6e00befd_968e_4d4d_8d5e_c3618ad43c44
        a                           sosa:Observation , owl:NamedIndividual ;
        rdfs:label                  "6e00befd_968e_4d4d_8d5e_c3618ad43c44"@en ;
        sosa:hasFeatureOfInterest   ssn:MaldiTof ;
        sosa:hasSimpleResult        1087.62e0 ;
        sosa:madeBySensor           ssn:CCS811_22 ;
        sosa:observedProperty       ssn:CO2 ;
        sosa:resultTime             "30.8.2019 12:00"^^xsd:dateTime ;
        ssn:measurement_unit        ssn:partsPerMillion ;
        ssn:wasOrginatedBy          ssn:CO2_Level .
```

Figure 7: Representation of an individual belonging to the "sosa:Observation" class with triples in the RDF file.

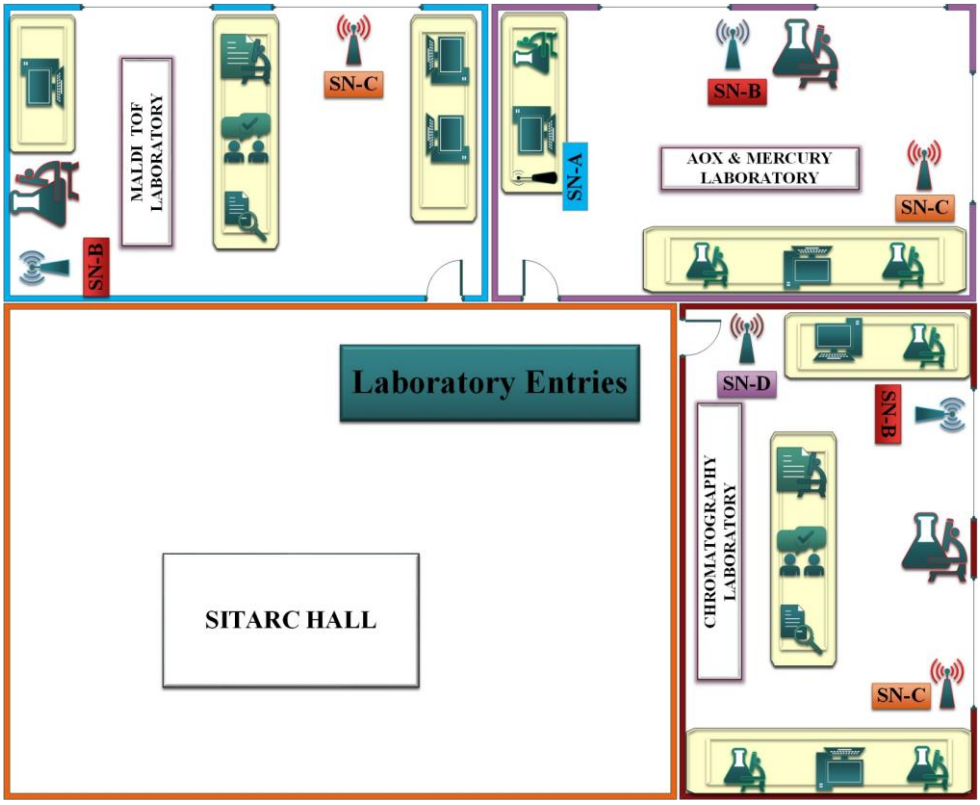

Figure 8: Deployment of sensor nodes to measurement areas.

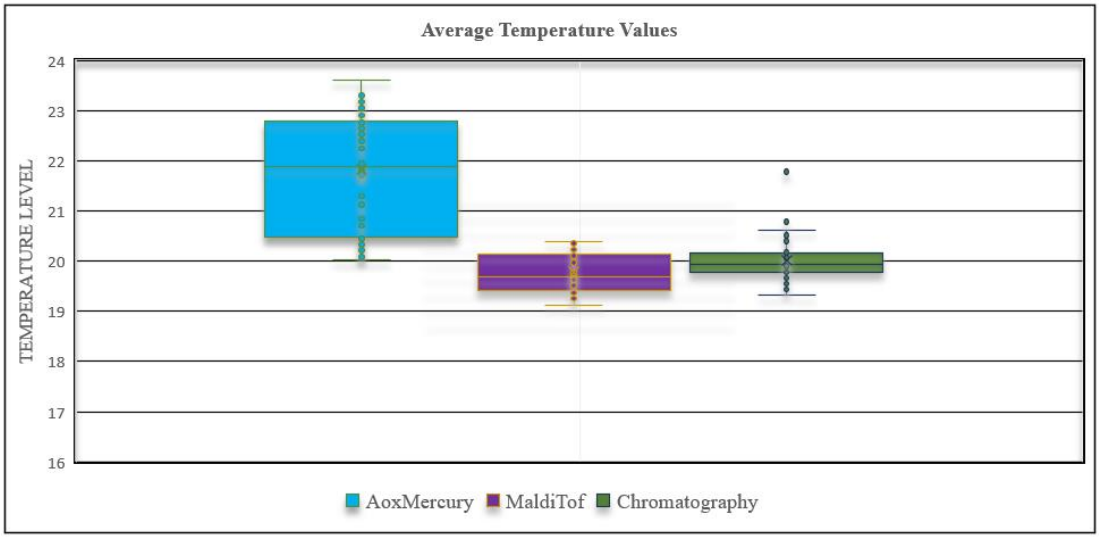

Figure 9: Box and Dispersion (spread) graph of average temperature values between 29.08.2019 and 12.10.2019 in laboratories.

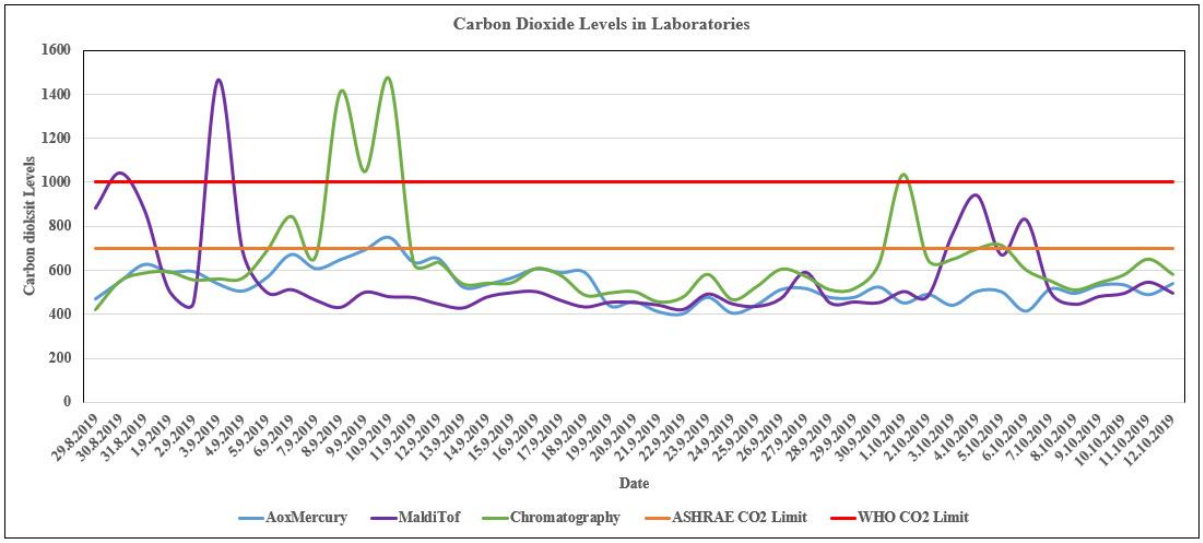

Figure 10: The daily average CO2 value between 29.08.2019 and 12.10.2019 in all laboratories.

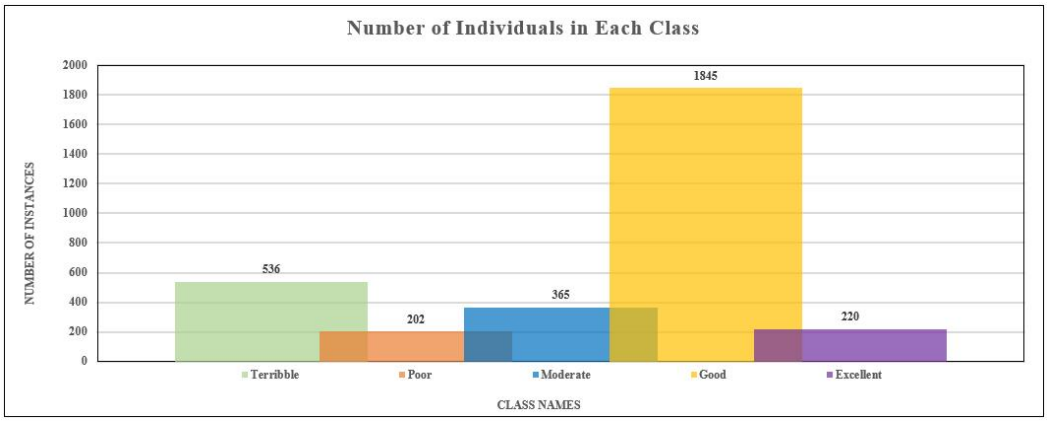

Figure 11: Scatter graph of classes by row.

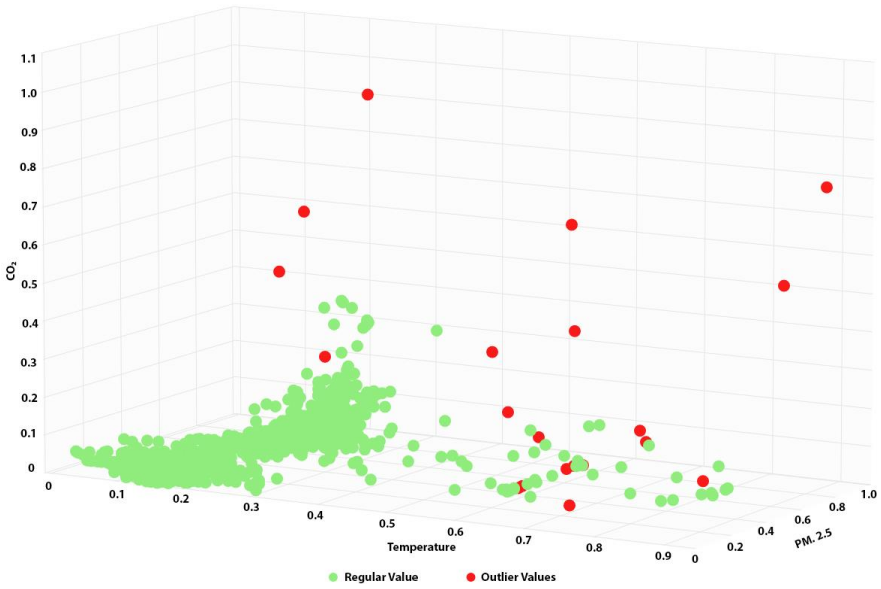

Figure 12: Detection of outlier data based on the K-NN approach. (colour print).

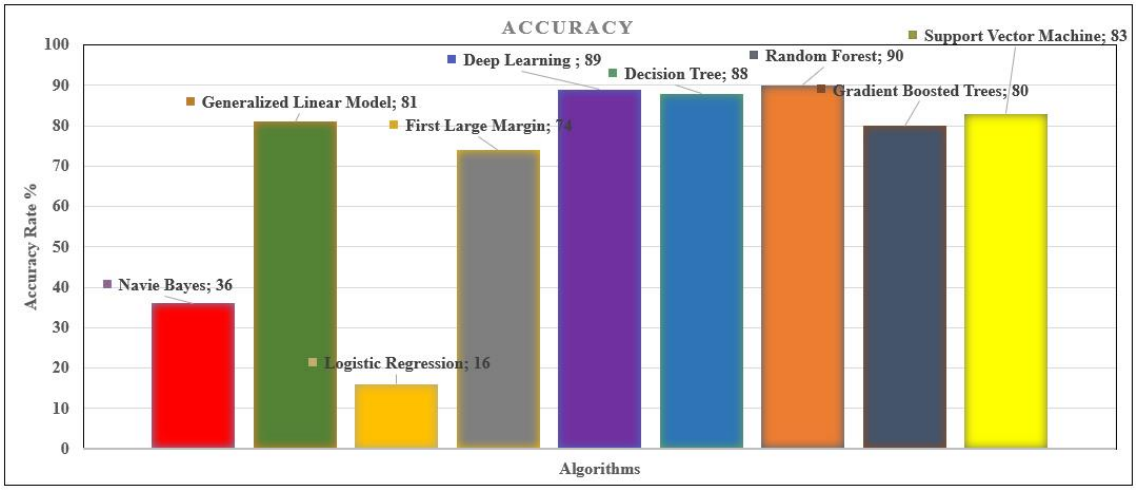

Figure 13: Comparison of accuracy percentages of algorithms used in the case study (colour print).

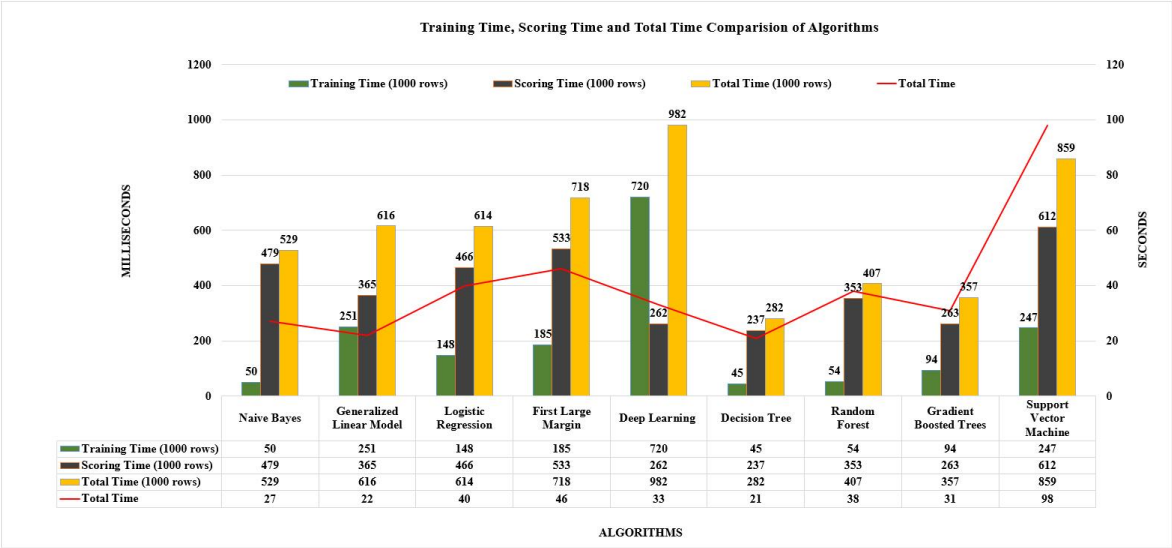

Figure 14: Comparison of the training time, the scoring time, and the total time of algorithms used in the case study (color print).

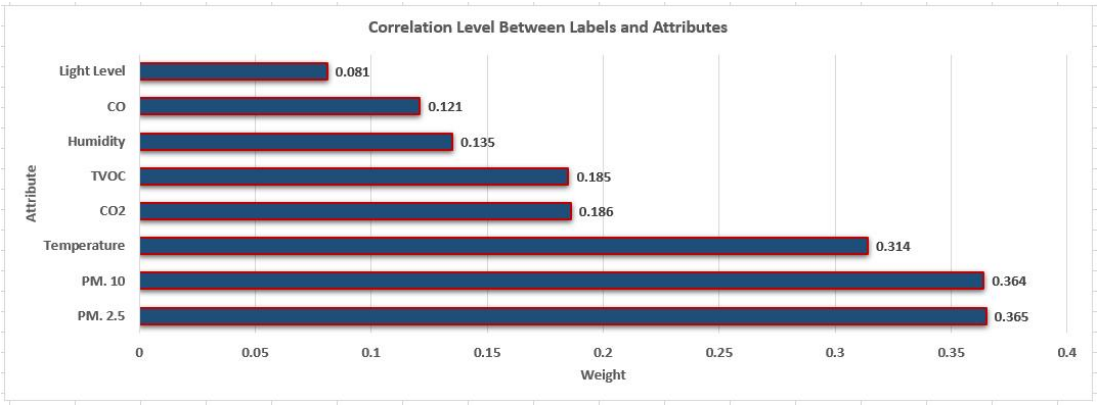

Figure 15: The average correlations calculated by all models used between labels and attributes are seen.

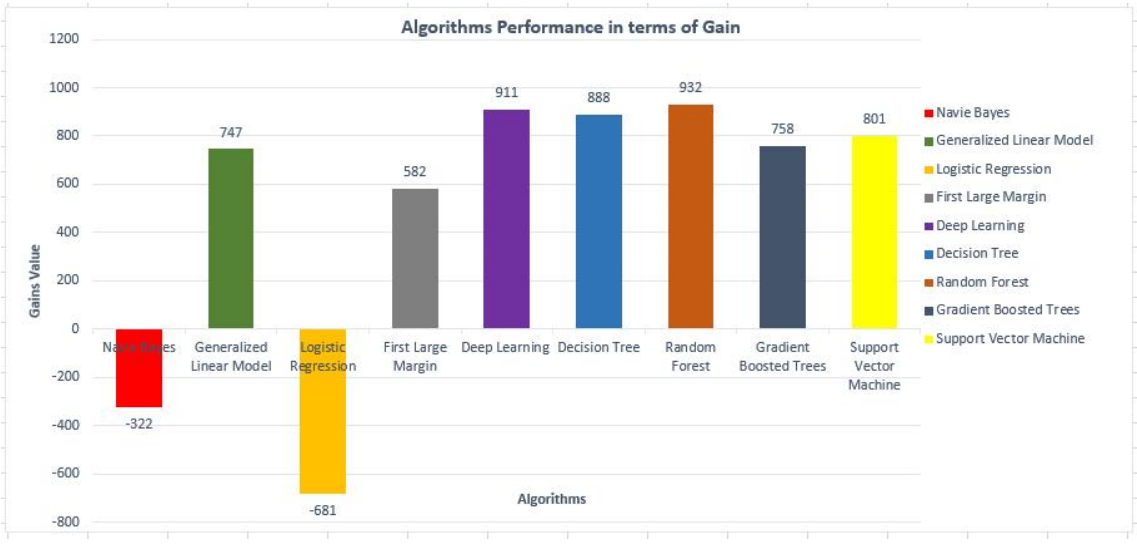

Figure 16: Comparison of algorithms used in the proposed system from the perspective of gain (color print).