Activity Detection In Smart Buildings Using Sensor Data

Şevval Bulburu 19011038
Introduction to Data Mining/ *Computer Engineering* Yıldız Technical University
İstanbul / Türkiye
sevval.bulburu@std.yildiz.edu.tr

Mehmet Alperen Ölçer 20011023
Introduction to Data Mining/ *Computer Engineering* Yıldız Technical University
İstanbul / Türkiye
alperen.olcer@std.yildiz.edu.tr

***Abstract*—This article explores supervised learning in depth, focusing on classification methods typically employed in this discipline. It discusses three well-known techniques: K-Nearest Neighbors (K-NN), Logistic Regression, and Decision Tree. K-NN uses the idea of nearest neighbors to classify unknown data, whereas Logistic Regression analyzes the link between input factors and the target variable's log-odds for binary outcome prediction. Decision Tree uses a tree-like structure to recursively split the feature space depending on decision criteria. The need of data pretreatment, including cleaning, quantization, and feature selection, in preparing data for optimal analysis is also emphasized in the paper. This article seeks to provide readers with a full grasp of by integrating theoretical underpinnings with practical practices.**

Keywords—KNN, decision tree, logistic regression, data, precision, recall, accuracy.

I. INTRODUCTION

Data retrieval is the process of getting access to and retrieving certain data or information from a database or storage system. Information Retrieval (IR) is the science of searching for information within relational databases, documents, text, multimedia files, and the World Wide Web [1]. Information retrieval is the process of finding pertinent data or documents in response to a user's query from a sizable collection of data or documents.

The study employs a robust dataset composed of sequential observations from various sensors placed across different rooms within a hotel. This dataset also incorporates information regarding the occupancy of each room over time. After undertaking the required preprocessing tasks and appropriately structuring the data, the data is prepared for further analysis. Subsequently, various predictive models are engaged, including K-Nearest Neighbors (KNN), Logistic Regression, and Decision Trees, for an in-depth understanding of the dataset's features. To ensure transparency and comprehensibility of the processes, data visualization techniques are adopted at every stage of the analysis. This meticulously prepared introduction outlines the fundamental aspects of the research methodology, aimed at fostering a detailed, systematic approach to sensor data in the hospitality industry.

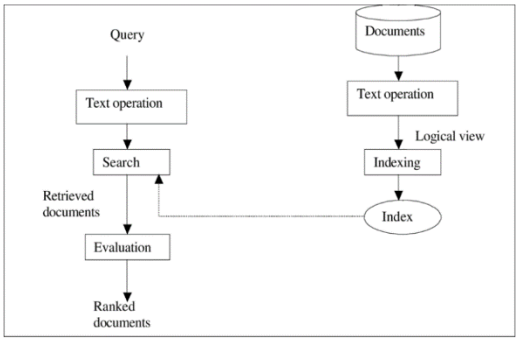

Fig. 1. A model of normal information retrieval systems

II. SUPERVISED LEARNING

In the field of machine learning, supervised learning is a crucial approach. With this method, a model is instructed to make predictions using a dataset in which the proper outputs (or labels) are already known and connected to each input. In order to develop a prediction model that can be applied to upcoming, unforeseen data, the model must be able to examine these known input-output pairs, spot patterns, and build a predictive model. This is the underlying idea behind supervised learning.

There are several mathematical changes involved in the supervised learning process. The model scrutinizes the input features, which are frequently numerically encoded aspects of the data, and maps them to the matching output labels during these transformations, which are the core of the learning process. The models' predicted categories or values are often represented by the labels.

Several components need to be established in order to implement the supervised learning approach. The "output space" reflects all potential predictions the model could be able to generate, whereas the "input space" represents all potential inputs the model might receive. The input-output pairings in the model's "training dataset" are well-known. All

possible mathematical representations of the algorithm's mapping from inputs to outputs are contained in the "hypothesis space." Last but not least, the process by which the model learns from the training data is known as the "learning algorithm".

The main objective of supervised learning is to identify the best hypothesis in the hypothesis space that minimizes the difference between the expected and actual results. This best-case scenario serves as a model that, under varied conditions, makes accurate forecasts. In other words, supervised learning seeks to produce a model that generalizes effectively from the training data to unknown scenarios, enabling accurate predictions and eventually satisfying the criteria of the particular job at hand.

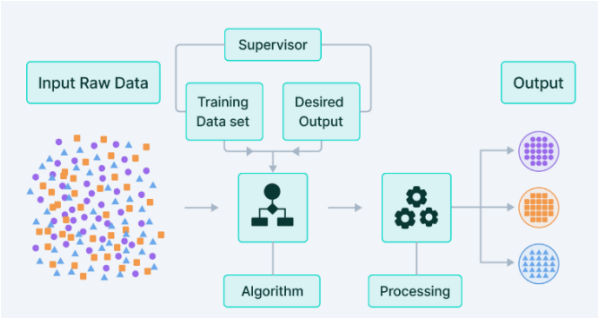

Fig. 2. Explanatory abstract of supervised learning

III. CLASSIFICATION METHODS USED

A. *K-Nearest Neighbors (K-NN)*

A foundational approach that is frequently used in the field of supervised machine learning is K-Nearest Neighbors (K-NN). The core of this algorithm's functioning is on the idea of data similarity, which states that new data instances are put into categories based on existing data points that most closely match them. The algorithm works by determining how similar two instances are to one another and storing all available information, building a solid knowledge base. As a result, the K-NN algorithm effectively classifies new data when it is presented by making comparisons to well-known examples, assuring a high level of accuracy in data categorization.

The K-NN algorithm's adaptability is one of its main advantages. Although it is frequently used for classification tasks, it is also skilled at solving regression issues, demonstrating its versatility in a variety of machine learning contexts. Since K-NN is a non-parametric technique, it may be used to a wide range of data sets because it makes no assumptions about the distribution of the underlying data.

The K-NN algorithm's built-in "lazy learning" feature is an intriguing feature. Lazy learning methods postpone entire model development until a classification is needed, in contrast to "eager learners" who try to generalize a decision model based on the training data before receiving fresh data examples. Consequently, K-NN doesn't actually learn anything during the training phase; it only saves the dataset. When fresh information enters the context, true learning and the related categorization action take place.

In summary, the K-Nearest Neighbors algorithm is a strong and effective method for supervised machine learning problems. It can handle both classification and regression jobs by utilizing the idea of data similarity and enabling correct categorization of fresh data instances. Because K-NN is non-parametric and has a special "lazy learning" feature, it may effectively be used across many domains and adapt to different data distributions.

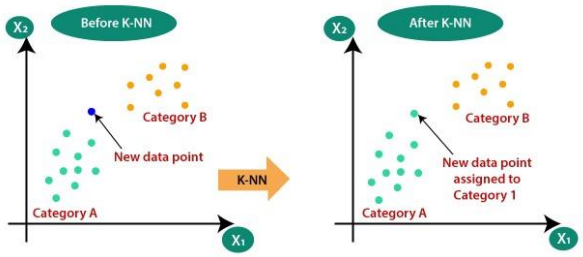

Fig. 3. The mechanism of the KNN algorithm

The K-Nearest Neighbors (K-NN) algorithm uses a structured classification procedure. The number of neighbors, represented by K, is chosen first. The Euclidean distance between the K neighbors and the new data point is then calculated. The K nearest neighbors are identified in the third phase using the previously estimated Euclidean distances. The algorithm then counts how many data points are in each category among the K neighbors. The new data point is then assigned to the category with the greatest number of neighbors. Finally, the model is judged complete. The K-NN method guarantees an effective and formal approach to classification tasks by carefully following these stages.

Formula of Eucledian Distance according to showed in figure 4.

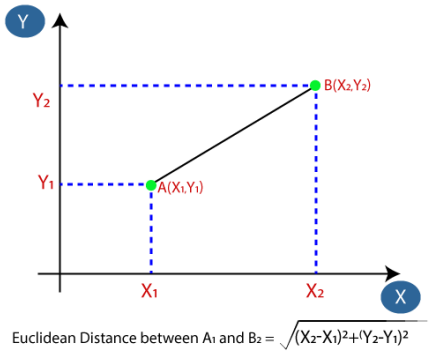

Fig. 4. Eucledian Distance Formula

B. *Logistic Regression*

Logistic regression is a well-known supervised learning technique that is commonly used in machine learning. It's major goal is to forecast categorically dependent variables based on a set of uncorrelated factors. Logistic regression, as opposed to linear regression, works with categorical outcomes, producing values such as 0 - 1, or true - false. Rather than giving precise values of 0 and 1, logistic regression produces probabilistic values ranging from 0 to 1.

Logistic regression, like linear regression, involves applying a curve on the data. Instead of a straight line, it uses a "S"-shaped logistic function with two maximum values, which often indicate the classes being predicted. This curve depicts the probability of a given event, such as assessing whether cells are malignant based on certain traits or predicting whether or not a mouse is fat depending on its weight. The importance of logistic regression stems from its ability to provide probabilities and accurately categorize new data using both continuous and discrete datasets. It can handle a wide range of data and quickly discover the most influential variables for classification. As a result, logistic regression is an important tool in machine learning for classification tasks.

Logistic regression is relatively fast compared to other supervised classification techniques such as kernel SVM or ensemble methods (see later in the book) but suffers to some degree in its accuracy. It also has the same problems as linear regression as both techniques are far too simplistic for complex relationships between variables. Finally, logistic regression tends to underperform when the decision boundary is nonlinear [2].

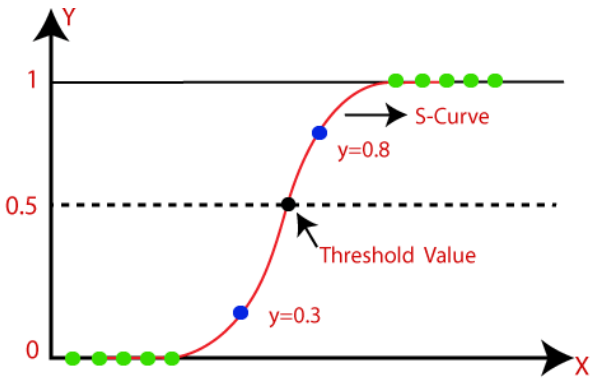

Fig. 5. Logistic Regression S-Curve

Logistic Regression uses the equation displayed below.

$$L(\beta;\mathbf{y},\mathbf{X}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$
$$= \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{X}_i\beta)}{1+\exp(\mathbf{X}_i\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(\mathbf{X}_i\beta)}\right)^{1-y_i}.$$

Fig. 6. For a sample of size n, the likelihood for a binary logistic regression

C. *Decision Tree*

One of the widely used techniques in data mining is systems that create classifiers. In data mining, classification algorithms are capable of handling a vast volume of information. It can be used to make assumptions regarding categorical class names, to classify knowledge on the basis of training sets and class labels, and to classify newly obtainable data [3]. Although it is capable of being used for both kind of problems, it is more commonly used for classification tasks. The attributes of the dataset are represented as internal nodes, decision rules as branches, and outcomes as leaf nodes in this tree-structured classifier. There are two sorts of nodes in a decision tree: decision nodes and leaf nodes. Decision nodes aid decision-making by splitting out into many paths, whereas leaf nodes indicate the ultimate result with no further branches. The decision-making process in a decision tree is based on running tests depending on the properties of the dataset. This graphical depiction allows you to investigate all possible solutions to a given problem or decision under specific parameters. The decision tree, as the name implies, is figuratively like a tree, beginning with a root node and expanding through successive branches to form a tree-like structure. The Classification and Regression Tree (CART) technique is used to build a decision tree, which allows for the production of a complete tree. A decision tree, in essence, asks questions and divides the tree into subtrees based on the answers.

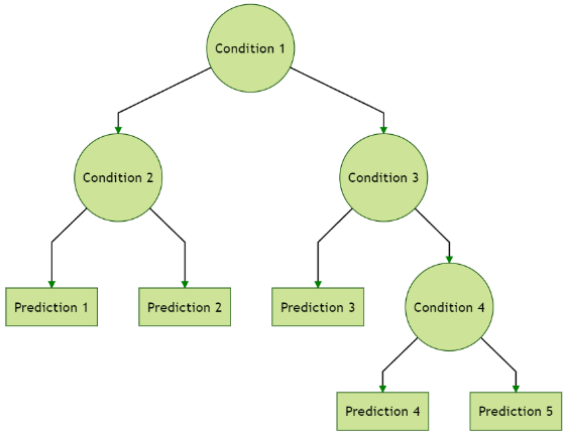

Fig. 7. An illustration for a decision tree with 5 leaves

IV. DATA PREPROCESSING

The dataset was built from 255 sensor time series that were used in 49 hotel rooms. Each room's dimensions fall into one of five categories:

• CO2 concentration: This describes the level of carbon dioxide in a space.

• Humidity: This describes the amount of moisture in the air of the space.

• Temperature: The temperature of the space is shown by this data point.

• Light: This parameter measures the room's brightness.

• PIR: Information from passive infrared (PIR) motion sensors.

From Friday, August 23, 2013, to Saturday, August 31, 2013, data was gathered during a one-week period. Every 10 seconds, the PIR motion sensor was sampled, but every 5 seconds, the other sensors were. Each file includes the sensor readings itself as well as timestamps.

The passive infrared sensor, or PIR sensor, is an electrical sensor that gauges the amount of occupancy in a space by measuring the infrared (IR) light emitted by objects in its range of view. 94% of the PIR data is zero, suggesting an unoccupied room, whereas just 6% of it is non-zero, indicating an occupied room.

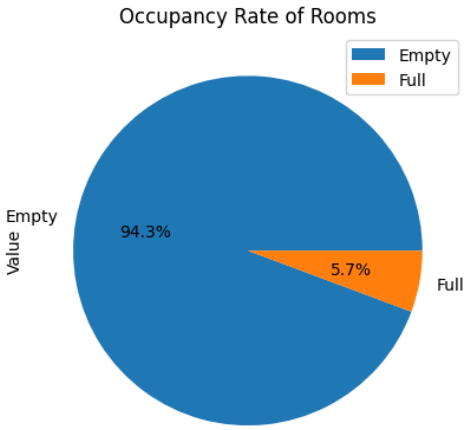

Fig. 8. The occupancy rate of rooms

Preprocessing began by combining the data from each room into a single consolidated table. This required reviewing and integrating the data from each room to make sure the combined dataset was devoid of any null values. It was essential to exclude these values to preserve the precision and dependability of following studies.

Synchronization became a crucial preprocessing step due to the asynchronous nature of the sensor readings, which were not gathered at regular intervals of time. Measurements from all sensors were aligned throughout this procedure, giving the data a uniform timeframe.

Quantization was carried out to ensure uniformity and comparability across all data points. This process was crucial in standardizing the data and preparing it for accurate analysis.

The dataset was split into two unique subsets for training and testing after the first preparation steps. This separation made it possible to evaluate any future forecasting models objectively.

The data was then scaled using a conventional scaler. Through the process of normalizing the dataset's range, intrinsic patterns were preserved but possible bias brought on by different measurement scales was reduced. The success of the machine learning algorithms that were utilized afterward depended heavily on this standardization.

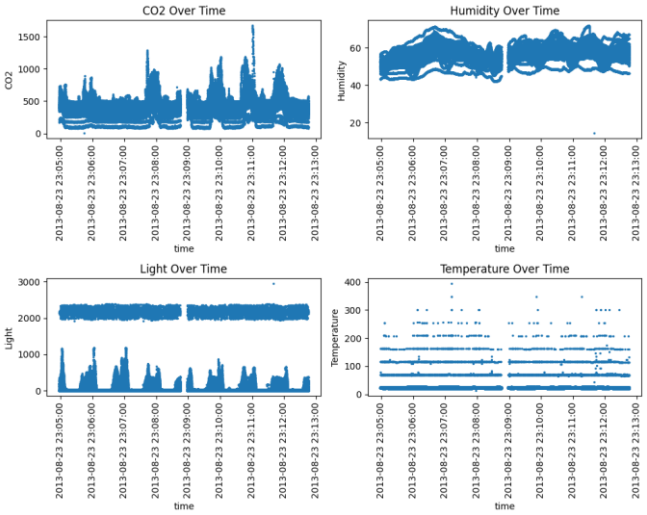

Fig. 9. Data distribution after preprocessing

V. RESULTS

KNN, Logistic Regression and Decision Tree machine learning models are trained with the dataset. Model's achievement measured with precision, recall and f1-score. Precision is a metric that assesses the accuracy of positive predictions. It is calculated as the ratio of true positives (positive instances that were correctly predicted) to the sum of true positives and false positives (positive instances that were incorrectly predicted). Recall, also known as sensitivity or true positive percentage, is a statistic that evaluates the proportion of true positive cases that are accurately detected. It is determined as the percentage of true positives to the sum of true positives and false negatives(positive instances mislabeled as negative).The F1-score is the harmonic mean of precision and recall, providing an unbiased assessment of a classifier's performance. It is calculated by multiplying the precision and recall products by the total of these two values. The support value represents the number of instances in each class, which is utilized as the denominator in precision, recall, and F1-score computations.

The result metrics are shown in figure10, figure11 and figure12.

-	precision	recall	f1-score	support
0.0	0.98	0.99	0.99	144717
1.0	0.84	0.72	0.78	8838
accuracy	-	-	0.98	153555
macro_avg	0.91	0.86	0.88	153555
weighted_avg	0.97	0.98	0.98	153555

Fig. 10. Confusion metric for K-NN

-	precision	recall	f1-score	support
0.0	0.97	0.98	0.97	144717
1.0	0.61	0.44	0.51	8838
Accuracy	-	-	0.95	153555
Macro_Avg	0.79	0.71	0.74	153555
Weighted_Avg	0.95	0.95	0.95	153555

Fig. 11. Confusion metric for Logistic Regression

-	precision	recall	f1-score	support
0.0	0.97	0.98	0.97	144717
1.0	0.61	0.44	0.51	8838
Accuracy	-	-	0.95	153555
Macro_Avg	0.79	0.71	0.74	153555
Weighted_Avg	0.95	0.95	0.95	153555

Fig. 12. Confusion metric for Decision Tree

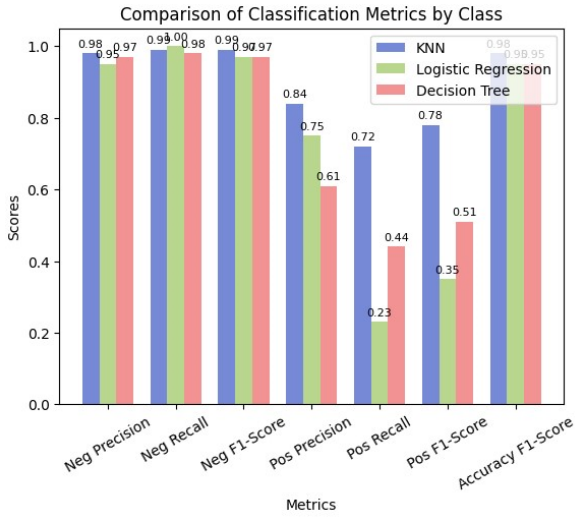

Fig. 13. Comparesion of the Results

Based on the outcomes of the algorithms in figure 13, the performance of the Decision Tree and Logistic Regression classifiers for the provided dataset is similar. However, when compared to the other two algorithms, the approach using KNN provides more successful results. Given the KNN algorithm's dominance in performance, it can be concluded that it is the best choice for training this dataset.

VI. Conclusion

In conclusion, the dataset completed a thorough set of preparation procedures, producing a cleaned, standardized dataset that was ready for usage. The data's compatibility and integrity for next tasks were assured by these processes. A number of visualizations were developed and presented in order to better comprehend and describe the nature of the dataset. The subsequent steps of data analysis were aided by these graphical representations, which offered insightful information about the dataset's structure.

The dataset was partitioned into two independent subsets, training and testing data, after preprocessing and visualization. This separation made it possible to evaluate the prediction models objectively, guaranteeing their efficacy and generalizability for eventual use.

Then, using several estimators, including K-Nearest Neighbors (K-NN), Logistic Regression, and Decision Trees, a number of prediction models were built. These models, which were created using several algorithmic tenets, were assessed for their capacity to predict outcomes while providing various viewpoints on the properties of the data. To comprehend how each model handled the dataset, comparisons were made and graphically shown.

Evaluation revealed that the K-Nearest Neighbors (K-NN) algorithm model produced the best results for this particular dataset in a training environment. This result emphasizes how crucial it is to choose a machine learning algorithm that is matched to the particular characteristics of the data and the study's aims. It further highlights the K-NN algorithm's adaptability and strength in handling a variety of datasets, reiterating its fit for this specific purpose.

REFERENCES

[1] Bijalwan, Vishwanath, et al. "KNN based machine learning approach for text and document mining." International Journal of Database Theory and Application 7.1 (2014): 61-70.

[2] Nusinovici, Simon, et al. "Logistic regression was as good as machine learning for predicting major chronic diseases." Journal of clinical epidemiology 122 (2020): 56-69.

[3] Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." Journal of Applied Science and Technology Trends 2.01 (2021): 20-28.