Evaluation of Data Mining Techniques and Its Fusion with IoT Enabled Smart Technologies for Effective Prediction of Available Parking Space

Original Scientific Paper

Anchal Dahiya
Research Scholar, Department of Computer Science and Application, Maharshi Dayanand University
Rohtak, India
Assistant Professor MNS Government College, Bhiwani
aanchaldahiyaa@gmail.com

Pooja Mittal
Faculty of Computer Science, Department of Computer Science and Application
MDU, Rohtak, India
mpoojamdu@gmail.com

Abstract *– After experiencing the hard times of pandemic situations we learned that if we could have a smart system that can help us in automatic parking of the vehicles then it could be a great help to society. This idea motivated us to carry out this current work. Though, nowadays, in almost every application domain, IoT techniques are the buzzword. IoT techniques can also be used to achieve efficacy in predicting free available parking space in advance. But the biggest challenge with IoT techniques is that they generate numerous data, which makes its analysis intangible. It was realized that if IoT techniques can be fused with outperforming data mining techniques, more efficient predictions can be performed. Thus, for this purpose, the main objective of our paper is to firstly, select the most appropriate data mining technique, based on performance evaluation, and then to perform prediction of available parking space in advance by fusing it with IoT techniques. Due to the busy schedule, the drivers need to get information about free parking spaces in advance by using smart phones. With the help of this information, it will be easy for the drivers to park their vehicle in the exact location without wasting their precious time and will maintain social distancing in crowded areas too. Data mining techniques can play an important role in the prediction of available parking space, by extracting only relevant and important information when applied to the given dataset. For this purpose, a comparative analysis of five data mining techniques such as the Support Vector Machine, K- Nearest approach, Decision Tree, Random Forest, and Ensemble learning approaches are applied on PK lot data set by using Python language. For calculation of result anaconda (spyder) is used as a supportive tool. The main outcome of the paper is to find the technique that will give better results for the prediction of the available space and if we fused data mining techniques with IoT technologies results are improvised. Evaluation parameters that are used for finding the best technique are precision, recall, accuracy, and F1-Score. For numerical calculation of the results, the k-fold cross-validation method is used. As the empirical results are calculated using the Pk lot dataset, the decision tree outperformed the best among all the techniques that are selected for analysis.*

Keywords*: IoT-Enabled Smart Parking, Parking Sensors, Data Mining, Ensemble Learning, and Decision Tree*

1. INTRODUCTION

The number of vehicles on the streets of metropolitan cities and large urban areas has grown tremendously. It is also difficult to track down empty parking spaces due to such an enormous number of vehicles. Therefore, drivers also waste energy in searching for a parking space, which results in additional traffic. In metropolitan cities like Delhi, Bombay, or even in smaller cities like Rohtak most challenging task associated is to find the free parking space in public areas. One of the survey reports of IBM [15] states that about 45% of the road traffic in cities is due to drivers searching for the vacant parking space for parking their vehicles. Due to vehicles, many problems intensify such as pollution emission, consumption of fuel, congestion on roads, wastage of time, and also contributing to the accidents because of the focus of the drivers for finding free parking[1]. In near future, it is estimated, that there will be around 2 billion cars on the road. With the rapid growth of the vehicles in a small amount of time congestion increase with the passage of time and it is difficult for drivers to find the parking area .Lots of work have been done for the management of parking space such as the utilization of sensors that determines

vacant parking spots[2] and feedback by the users that will help the drivers to inform about other vacant space by a mean of smart applications that will identify the vacant parking space[3] .These systems are based upon the temporary data, so there is less probability for reserving and allocating the free parking space, so these techniques are practical in a short frame where the driver is in a nearby location of the parking space. These techniques don't assure if a parking space is available or not. At a particular point in time in the near future to predict the accessibility of vacant parking spaces these techniques are combined with artificial intelligence approaches that provide the smart solutions in the real environment. To obtain better results for the prediction of vacant space, a lot of data is generated by IoT sensors that are further coupled with other IoT devices, and data mining techniques are applied to the real-time sensor data to get hidden patterns from available data and it will provide the useful information to the drivers. IoT devices can produce a large volume of data and that data is locally processed and transferred to a centralized database, where it can be further processed and analyzed to produce knowledge. Data mining is defined as a family of techniques for analyzing such huge data that will collect the historical data which is generated by IoT devices and will preprocess the data to find the hidden patterns that will help in predictive analysis. Data mining is a strong contributor to deal with the huge amount of data that is generated from IoT devices. That is the reason why we fused data mining techniques with IoT devices that generate a tremendous amount of data. A large number of data mining techniques are available, one of the problems arise is to detect the appropriate data mining technique for the given problem and the size of the records i.e dataset because the performance analysis of each data mining model varies from problem to problem that is done for comparing data mining algorithms in numerous application domains. Data mining algorithms have been compared for different applications such as Djaneye-Boundjou et al.[4] applied K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Artificial Neural Network (ANN) methods to a malware sorting problem and find the solution in terms of accuracy and KNN outperformed SVM and ANN.

1.1 DESCRIPTION

The main objective of the paper is to find the optimized technique that will help the drivers for the prediction of free parking space based on the data that is generated by IoT sensors. When working with IoT sensors the biggest challenge is the size of the data generated by IoT devices. To deal with this challenge, we have fused data mining techniques that will handle large data efficiently. In this paper, we are analyzing different data mining techniques that can help us to regulate the best predictive technique among them for the accessibility of the free parking space. The PK lot dataset is used for the analysis of the results. For comparative analysis, we choose different data mining techniques such as the K-Nearest approach, Ensemble Learning approach, and decision Tree approach, Support Vector Machine, and Random Forest. Even though the number of data mining techniques are available in the literature, we choose these five data mining techniques because these are widely used by renowned researchers. At the end of the paper, the result is quoted with perspective to the main objective that to find the optimized technique for the prediction of vacant parking spaces.

- Identification of best performing algorithm among best known and widely used data mining algorithms.
- Endorsement of top-k free parking areas with admiration to the remoteness between the present location of the vehicle and vacant parking space.
- To determine how suitable prediction of available parking space is applied to the PK lot dataset

Impression of our Parking Space Prediction Model on Smart World. Now the day's widely used term is the smart world, it is an umbrella that will quarter numerous traits associated with urban research. Related to smart cities most important branch are transportation and flexibility. Smart flexibility and transportation have the latent that will make the important involvement in smart cities that will utilize IoT techniques. Traffic congestion problem occurs due to driver search for a parking place that will affect numerous procedures and fields of the smart cities some of these domains are parking space management, traffic management, and route planning management [15]. IoT-enabled smart parking system marks an effort that will diminish traffic jamming [7] that is presented by parking prediction data mining model that will make a significant impression on the smart world.

1.2 ORGANIZATION

The paper is organized as follows. In section 2 literature survey is presented. In Section 3 we will present a block diagram of IoT-Enabled smart parking system. In section 4 we will present an overview of the five data mining techniques that are used for analysis. In section 5 the performance of these data mining techniques is represented using evaluation parameters in which we will represent the result using precision, recall, accuracy, and F1-score, and finally conclusion is presented in section 6.

2. RELATED WORK

To deal with the parking spot reference problem, several systems have been proposed. A reference system based on real-time sensors is the most common solution to the problem that can detect the accessibility of the

parking space[2] . Wireless Sensor network is a real-time evaluation that is connected to a web server for gathering information that will decide the available parking space by J.Yang et al.[5] . The information is transmitted to the users via a cell phone. Another solution is proposed by Dong et al.[6] . Dong proposed a virtual reality-centered technique that will compact with factual time recognition of the free parking spaces. These systems gather the public information of free parking space. For example rented space, available space, price, etc all these types of information are collected by simulation-based methods after collection the public information is sorted using page rank algorithms. This algorithm is centered on examine actual data so they are not capable to handle the probability to forecast the accessibility of the parking space in the time frame (eg between 15 and 35 minutes from the present time) within the interest of the demand of the users. One solution is proposed by R.E Barone et al.[7] . An architecture is proposed by him and it is known as an intelligent parking assistant. The architecture does not provide the prediction of the availability of free parking space. This architecture allows the operators to reserve a parking space, to reserve for free space, the user needs to register with intelligent parking space. This architecture is used only by authorized users. Vlahogianni et al.[8] proposed a neural network-based model that will predict the occupancy rate of the parking spot. Y. Zheng et al.[9] accomplished a comparative analysis of different data mining techniques such as support vector regression, neural network, and regression tree that will predict the occupancy rate of the parking space. Y. Zheng et al. conclude that the regression tree method is best among the other two algorithms. A comparative analysis is also performed by C. Badii et al.[10] that uses different techniques for analysis, these techniques are Support vector regression, recurrent neural network, Bayesian Regularized Neural Network. It uses Auto-regressive integrated moving average method that will be used for the prediction of parking space accessibility within a particular parking area. There are two different areas of research for the availability of the parking space these are on-street parking area and off-street parking area [10]. Off–street parking approach is limited to inside the garage and On-street parking includes complex features such[11] as weather forecast in their data set. A. Camero et al.[11] proposed a Recurrent Neural Network approach that is based on the prediction of the number of free spaces available in the parking spot. The main aim of the paper is to improve the performance of the Recurrent Neural Network, for this, they introduced a Genetic Algorithm based technique that will find the best configuration for RNN using the Genetic approach.

3. IOT-ENABLED SMART TECHNOLOGIES

IoT was first introduced in 1999 by Kevin Asthon [17]. As this technology is evolving, it promises to connect all the things surrounded by a network and establish communication with less human involvement. Still, the IoT is in the beginning stage, and there is no common design present today. There are no boundaries or guidelines exists to define the definition of IoT. So, depending on this, the application of the IoT has various definitions. Shortly, it is defined as the physical world's things or an environment attached to embedded systems or sensors and connected to the network through wireless or wired networks. These connected devices are called smart objects or devices. IoT deals with linking all the world through the Internet. IoT helps to link trillions of nodes of different objects to the major supermarket web servers and clusters. IoT also helps to incorporate emerging software technology and networking technologies. IoT's main goal is to make the world around us smarter by supplying the data it needs through historical and real-time feeds with the help of data mining algorithms and automatically applying computational knowledge intelligence to make smart decisions. To understand and monitor dynamic environments around us, the data collected from IoT devices will be used to allow higher automation, better decision-making, greater efficiencies, accuracy, and productivity.

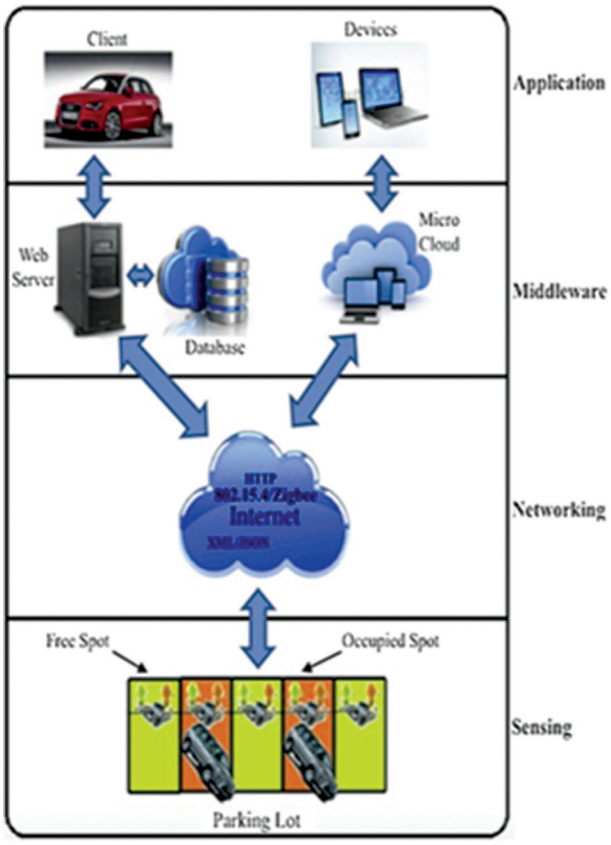

Fig. 1. IoT-Enabled Smart Parking Model

Due to an exponential increase in data volume and sophistication, data mining tasks help us to process such a large volume of data that is the reason we have combined data mining techniques with IoT. As the IoT- enabled system produces a large amount of data in today's time, our main task is to develop an effective model for analyzing, managing, and mining the data. Data mining techniques help us for determining interesting, novel,

and potentially useful patterns from big data sets and applying various techniques to extract hidden information. The IoT enabled smart parking system model is divided into four layers, i.e., application, middleware, networking, and sensing, as shown in Fig. 1[12] . IoT system is classified as the as processing, connectivity, and sensing. The sensing layer involves sensing the speed of cars and humans or any object like accelerometer, temperature sensing, pressure sensing, etc. These can be processed by using various processors such as the hybrid processor, network processor, etc., and these devices connected with the help of technologies like Wi-Fi, GPS, RFID, etc. Processing unit act as an intermediate between the cloud and the sensors. The sensors are connected wirelessly to the processing unit and the processing unit process the data with the help of data mining techniques that will exact the patterns from the large volume of data to predict the location is vacant or not. Using a smart parking model which is based on the sensor and contains a pi-camera to detect the vacant spaces and sends the data to the server, this stored data is accessed by the user [5] .This increases the user's ability to check the status or availability of spaces before setting up their ride. The task here is to optimally use the available resources to reduce the search time and traffic congestion in the region. The application layer serves as an interface to communicate the system with the end-user.

4. DESCRIPTION OF DATA MINING TECHNIQUES

Data mining techniques such as KNN, SVM, and Decision tree, Random Forest, and Ensemble Learning are evaluated and analyze a data set to predict the availability of the parking space is compared here.

4.1. K- NEAREST NEIGHBORS (KNN)

One of the simplest data mining techniques is KNN that is based on the supervised learning technique. It works on the principle of similarity between the new data and the existing data. KNN puts the new data into the category that is most similar to the existing categories. It is a non-parametric algorithm. Samples are classified based on the distance between them. Observations are classified based on the form such as X and Y in the training data set. Here Xi is a vector that will contain the feature values, and Yi is the class label against Xi. Let's take an observation of Xj and we want to predict its class label that is Yj using KNN. The equation used for this observation is given in equation (1), KNN finds J number of observations in X that is similar to the observation Xj:

$$DIS\ Xj,Xi = D(Xj, Xi)1\_i\_n. \quad (1)$$

By using equation (1), the distance between observation Xj and all other observations in X can be calculated using the above formula. When we perform the training on the dataset using the KNN algorithm, then it just stores the new data and at the time of classification, it performs an action on that data and that is the reason it is also known as lazy learner algorithm [20]. Steps used for KNN are as follow:

- Select the n number of neighbors.
- Calculate the Euclidean Distance for the selected number of neighbors.
- Now take n nearest neighbors that are calculated using the Distance formula.
- Among these n neighbors, count the number of data points in each category.
- Assign the new value of the category for which the number of neighbors is maximum.

Let's understand based on an example. Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. By using K-NN, we can easily identify the category of a particular dataset.

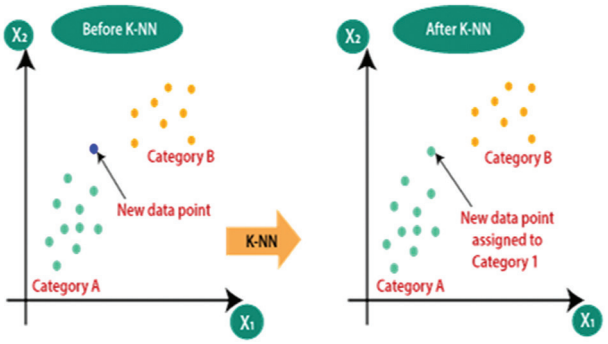

Fig. 2. Example of KNN for assigning the new value to the category [20]

4.2. DECISION TREE

It is a non-parametric supervised learning method that is used for classification and regression. In a decision tree algorithm a tree is constructed by setting different conditions on different branches. Workflow of the decision tree is shown in fig. 3. A decision tree consists of different nodes that are the root node (starting point of the tree), internal nodes (where nodes are divided into different branches), and the external nodes (these are the terminal nodes that will contain homogenous class)[13] .

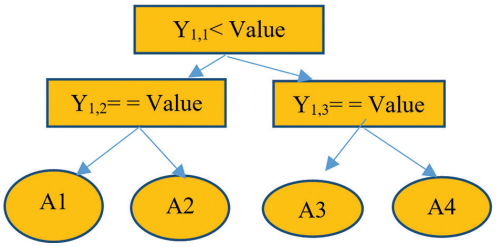

Fig. 3. Decision tree for checking the condition of value

The main aim of using this technique is to create a model that predicts the value of the target variable by learning simple decision rules that are inferred from the data features [20].

4.3 SUPPORT VECTOR MACHINE (SVM)

It is a supervised learning technique that is associated with the learning algorithms that will analyze the data for classification and regression analysis. It is one of the predictive models that is based on statistical learning. It can take two groups of data that is training data and testing data. The SVM learning algorithm maps the training data sets samples to the points in the space to maximize the width of a gap between these two groups. Now the new variables are mapped into the same space and they are predicted that they belong to the group based on which side of the gaps they fall. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper plane [20]. Pictorial representation of workflow of SVM is shown in fig. 4. In the figure the SVM takes the data from the database that will contain heterogeneous data, it maps the training data into different groups according to their category.

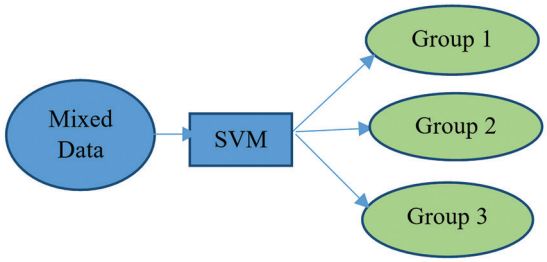

Fig. 4. Block Diagram of SVM

4.4 RANDOM FOREST

Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset [20].Random forest is similar to the decision tree algorithm. Multiple independent decision trees are the subparts of the random forest. Each tree in the random forest divide out the class prediction and the class with the most vote become the subtree. Each tree sets a conditional feature differently. Whenever a sample arrives at a root node, then it can be forwarded to all among the subtree. Class label is predicted by each subtree for that specific tester. At the ending stage, the class in the majority is assigned to that specific tester. The working of the random forest algorithm is represented in figure 5.

4.5 ENSEMBLE LEARNING APPROACH

This approach combines multiple data mining techniques. In this paper, we have combined SVM, KNN, Random forest, and Decision tree that will solve the predictability problem of available parking space. In the ensemble learning approach the training data properly trains each model. When the training process is performed the ensemble learning approach feeds the testing data to different models and then each model predicts a class label for each sample in the testing data available in the training set. In the next step, the voting process is performed for the prediction of each sample. Normally two types of voting are available hard and soft voting. Hard voting assigns a class label that is voted by the majority to the sample. In soft voting, average the probability of all the expected outputs. Block diagram of the ensemble learning approach is shown in the given fig. 6.

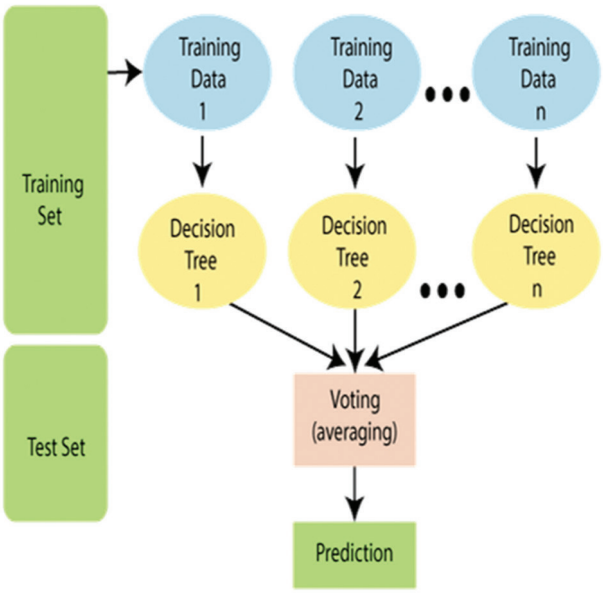

Fig. 5. Working of Random Forest algorithm [20]

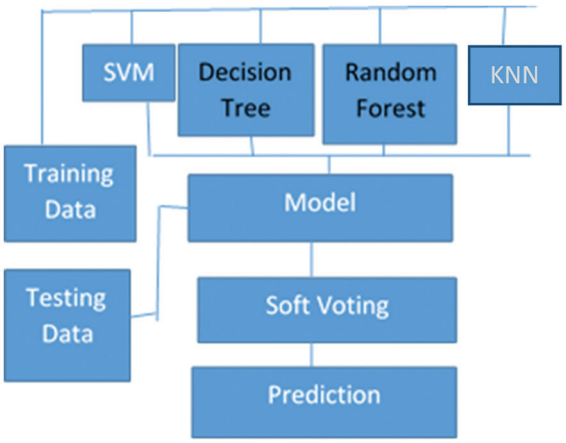

Fig. 6. Block Diagram of Ensemble Learning

5. EVALUATION AND RESULTS

In this paper, five data mining techniques are used that are described in the previous section. These techniques analyze the free parking space and help the drivers to get the most relevant information about the nearest parking space. Sensors are used for the collection of the data that are deployed in a real-time environment i.e. smart city Santander (a city in Spain). In the section below, we are analyzing the performance of five data mining techniques that can be used for the

prediction of the free parking space and also provide a relative analysis of the initial results.

5.1. IOT-ENABLED PARKING DATA SET

The data set used for prediction is PK lot data that contains IoT data which collects the data from the sensors that are deployed in different locations of the smart cities. Parking sensors [16] collect over 3-month of data, this data set was constructed as a part of the Pk lot data. The objective is to predict the available parking space within a time interval of 15 to 25 minutes and also analyze the accuracy of the prediction. The organization of the collected data set is as follow:

1. Parking Identity: A unique identity connected with each parking area.
2. Timestamp: It is a parking space data collection.
3. Period: It refers to the total period for which specific parking space is available or engaged.
4. Start and Finish Time: It refers to the interval of the time during which a parking space status continued to be the same i.e. accessible or engaged.
5. Status: It represents the status of the parking space i.e. accessible or engaged.

The above features are organized in the form of the table 1 given below.

Table 1. Mined Geographies

Key Features	Range
Parking Space Identity	Unique identity of a sensor
Day	7 days of a weak
Starting Hour	0-23 hours of the day
Starting Minute	0-59 minutes of the hour
Ending Hour	0-23 hours of the day
Ending Minute	0-59 Minutes of the hour
Status	0-1(Accessible or Engaged)

Starting an hour and ending hour in Table 1 represent 15 or 25 minutes interval status for any specific space. Data set is collected every minute to provide the exact status of the parking space.

5. 2. SOFTWARE TOOL

Spyder is used for implementation. It is an open- source cross-platform integrated development environment for scientific programming in the python language.

5.3. HYPER-SPECIFICATIONS OF THE DATA MINING TECHNIQUES

In Table 2 we will represent the five specifications for the data mining model that will be used for the analysis of different data mining techniques. Grid search [17] is used to the get best result using the specification for each data mining model.The parameter that is tuned for SVM is "C" and it is the regularization parameter. The strength of this parameter is inversely proportional to C which must be strictly positive. Kernel specifies the type of algorithm to be used and we are using a widely used algorithm that is "rbf".The degree of the kernel is 3 by default. And coefficient used for gamma is "scale "and scale uses 1 / (n_features * X.var()) as value of gamma. Here tolerance is used for stopping the criterion and the size of the cache is 200 by default. Parameters that are tuned for KNN are n_neighbors, distance metric that is Euclidean is used in our case, and n_jobs parallel jobs are used for searching the nearest neighbor in our case. "n_Neighbor" experimented with different numbers of neighbors (2, 6, 8,10,24,49 and 99). "n_neighbors=10" provide the best result. Weights are set as a uniform for all the neighbors' points that are weighted equally. For Decision tree 3 parameters are tuned. Here max depth defines the maximum depth of the tree, over fitting problem occurs if the depth of the decision tree is increased. In our case, max depth is set to 100. Min sample_ leaf defines the number a leaf node can contain, in our case, we set this value equal to 5. For random forest, we tuned 4 parameters that are max depth that is similar to the decision tree parameter. The value of both the parameters are the same in our case. "n_estimator" is used to define the number of trees in the forest and in our case we set this value equal to 100. "Criterion=entropy" is works on information gain. Here information gain is related to the decrease in entropy after every split. For ensemble learning 3 parameters are tuned "estimator" defines the techniques used by ensemble learning, in our case we use SVM, KNN, Random forest, Decision tree. Another parameter used for ensemble learning is the weight that is used for defining the priority of each estimator used. The voting classifier is used for assigning the weights to the estimators and we have assigned equal weights to all the classifiers except the decision tree. High priority is assigned to the decision tree because the performance of the decision tree is better among all data mining techniques when it is used to predict the available parking space. These hyper-specifications are shown in the tabular form in table 2.

5.4. EVALUATION PARAMETERS

The parameters that are used for the valuation and comparing data mining techniques are given below.

1. **Precision:** It can be defined as the division of all the samples that are labeled as positive samples and that samples are positives[14]. The mathematical formula is given below.

$$\textbf{Precision} = \frac{\textbf{True Positive}}{\textbf{True Positive+False Positive}}$$

2. **Recall:** It is defined as a fraction of all the positive samples; all the samples are labeled as positive [16].The equation of recall is given below.

$$\textbf{Recall} = \frac{\textbf{True Positive}}{\textbf{True Positive+False Positive}}$$

3. **Accuracy:** It is the ratio of the appropriately forecasted values in the total number of the available samples that is communicated in the following equation.

$$\textbf{Accuracy}=\frac{\#Accurate\ Predictions}{\#Total\ Samples}$$

4. **F1-Score:** The harmonic mean of recall and precision is defined as F1 score [14], the mathematical definition is given below.

$$\textbf{F1-Score}=\frac{2*(Recall*Precision)}{Recall+Precision}$$

Table 2. Hyper-Specification of Data Mining Techniques Used for Analysis

SVM		KNN		Decision Tree		Random Forest		Ensemble Learning	
Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
Kernal	rbf	n-neighbors	10	Max-depth	90	Max_depth	90	estimator	SVM, Decision Tree, KNN, Random Forest,
Degree	3	metric	euclidean	criterion	entropy	criterion	entropy	voting	soft
Gamma	Scale	n-jobs	nill	Min_sample	6	Min_samples	1	weights	1,2,1,1
CoefO	0.0	weights	uniform			N_estimators	100		

K-Fold Cross Validation: This method is used for testing the over fitting and to evaluate the consistency of the specific data model. In the k-fold authentication method, a dataset is divided into k equal sets in our case the value of k is 5. In the given k set one data set is used as a testing data set and all the remaining data sets are used as training data sets. In our case, we will configure it to generate 1000 samples each with 20 input features, 15 of which contribute to the target variable.

5.5. PERFORMANCE EVALUATION

In this section, we will provide the evaluation performance of KNN, decision tree, random forest, SVM, and Ensemble learning algorithms. A comparative analysis of 15-minute and the 25-minute predictions was done after considering 50% and 70% thresholds for both predictions. "Data mining algorithm has a capability of predicting a probability of class membership and this must be interpreted before it can be mapped to a crisp class label. The threshold is used to achieve the goal where all the values equal or greater than the threshold are mapped to one class and all other values are mapped to another class". To improve the performance of the results we have considered 50% and 70% threshold values. Why we need to set threshold values because when we are working with real-time data then we have received a variety of data values and if we compare all the values on a single threshold then results never be that much impressive. We have chosen the 50% and 70% threshold because we want results to be more reliable. When we work on real-time data the values are not of the same nature it may vary and depending on different values we are showing the result using different thresholds.

5.5.1. 15-Minute Prediction Rationality (50% Threshold)

Table 3 given below represents the cross-validation score of SVM, KNN, Decision tree, random forest, and Ensemble learning technique that uses 15-min prediction with 50% threshold. During computation, it is shown that SVM has the lowest performance with an average of 64.64% precision, 51.08% recall, 71.72% accuracy, and 56.57% F1-score. One of the simplest data mining models is KNN that is outperformed with SVM and the results are 72.03% precision, 66.35% recall, 77.72% accuracy, and 69.13% F1-Score. The performance of the random forest is even better and the results are 85.91% precision, 79.12%recall, 85.49%accuracy, and 82.36%F1-Score. As shown in the table that the performance of the decision tree and the ensemble technique both techniques are quite close to each other. The average precision of decision tree is 92.13% while an average precision value of the ensemble learning is 93.71%.The average recall score of the decision tree is 90.39% and the average recall score of the ensemble learning is 89.23%. The average accuracy of the decision trees 93.16% while the average accuracy of ensemble learning is 93.23% an improvement of only 0.07%. The average F1-Score of the decision tree is 91.70%, while ensemble learning shows 91.99%. Results are shown in the form of a graph in figure 7.

Table 3. Average cross Validation of each Data Mining technique (15-minute prediction validity with 50% threshold)

Metrics	SVM	KNN	DT	RF	EL
Precision	64.64	72.03	92.13	85.91	93.71
Recall	51.08	66.35	90.39	79.12	89.23
Accuracy	71.72	77.72	93.16	85.49	93.23
F1-Score	56.57	69.13	91.70	82.36	91.99

"SVM= Support Vector Machine, KNN=K-Nearest Neighbors, DT= Decision Tree, RF= Random Forest, EL= Ensemble Learning"

5.5.2. 15-Minute Prediction Rationality (70% Threshold)

Table 4 given below represents the cross-validation score of KNN, Decision tree, random forest, SVM, and Ensemble learning techniques that use 15-min prediction with a 70% threshold. After considering the 50% threshold it is clear that the performance of the SVM

is worst among all the given data mining techniques. SVM shows 72.24% accuracy with 62.91% average precision, 50.54% average recall, and 56.12% average F1-Score. The KNN shows 78.28% accuracy, 72.20% precision, 66.24% recall, and 69.10% average F1-score. The Random forest average accuracy was 85.60%, while its average precision was 86.02%, its average recall value was 79.80%, and F1-Score was 82.27%. From the given table 4 it is clear that the performance of the decision tree and the ensemble learning techniques shows the quite same performance both at the top end. The average accuracy value for decision tree and ensemble learning is 93.40% and 93.30% respectively, precision value is 92.12% and 94.14% respectively, average recall is 90.31% and 88.25% respectively, and F1-Score is 91.70% and 91.51% respectively. Results are shown in the form of a graph in figure 8.

Table 4. Average cross Validation of each Data Mining technique (15-minute prediction validity with 70% threshold)

Metrics	SVM	KNN	DT	RF	EL
Precision	62.91	72.20	92.12	86.02	94.14
Recall	50.54	66.24	90.31	79.80	88.25
Accuracy	72.24	78.28	93.40	85.60	93.30
F1-Score	56.12	69.10	91.70	82.27	91.51

"SVM= Support Vector Machine, KNN= K-Nearest Neighbours, DT= Decision Tree, RF= Random Forest, EL= Ensemble Learning"

5.5.3. 25- Minute Prediction Rationality (50% Threshold)

In the given section, we will represent the comparative analysis using 25-minute prediction validity with a 50% threshold. Table 5 represents the average cross-validation score for each data mining technique. The SVM shows the lowest score among the given techniques. The average precision value of the SVM is 64.90 % and the average recall value is 51.28%. F1-score depends on the precision and the average recall value, SVM remains low at 56.67% and the average accuracy value of the SVM is 71.93%. Now consider the average value of KNN as we compare the performance of KNN is better than SVM. The performance of KNN is 73.16% average precision value, 67.77% is average recall, 78.72% is average accuracy, and average F1- Score is 70.36%. If we compare the performance of the Random forest with these two techniques then the result is much better than these first two, with 81.43% precision, 72.77% average recall, 81.48% average accuracy, and an F1-Score is 76.86%. And if we take Decision tree and ensemble learning they are following the same trend. The performance of these two techniques is similar to the previous one. The average accuracy of the decision tree and the ensemble learning is 86.65% and 87.72% respectively, the average precision value is 84.63% and 87.64% respectively, and the average recall is 83.36% and 82.55% respectively, and F1-score is 84.01.and 85.02 respectively. Results are shown in the form of a graph in figure 9.

Table 5. Average cross Validation of each Data Mining technique (25-minute prediction validity with 50% threshold)

Metrics	SVM	KNN	DT	RF	EL
Precision	64.90	73.16	84.63	81.43	87.64
Recall	51.28	67.77	83.36	72.77	82.55
Accuracy	71.93	78.72	86.65	81.48	87.72
F1-Score	56.67	70.36	84.01	76.86	85.02

"SVM= Support Vector Machine, KNN= K-Nearest Neighbors, DT= Decision Tree, RF= Random Forest, EL= Ensemble Learning"

5.5.4. 25-Minute Prediction Rationality (70% Threshold)

In this, we will represent the performance result of all the data mining techniques with 25-min rationality with 70% threshold value. From the given table it is clear that the threshold value did not affect the standing of the data mining techniques for the configuration. The performance of the decision tree and the ensemble learning is always remains in the top two among all the techniques in terms of all the evaluation of the metrics. Decision tree shows 84.41% precision 83.12% recall, 86.81% accuracy and 83.76% F1-Score. The performance of the ensemble learning is 88.03% precision, 81.51% recall, 87.71% accuracy, and 84.63% F1-score. The next best technique is the Random forest technique with an average precision value is 81.85%, the average recall is 72.57%, the average accuracy is 82.16%, and F1-Score is 76.94%. If we compare the performance of the KNN and SVM then the KNN outperformed SVM with 74.35% precision, 67.35% recall, 79.37% accuracy, and 70.23% F1-score. The performance of the SVM is as follows. The average precision value is 64.34%, the average recall is 50.84%, the average accuracy is 73.08% and F1-Score is 56.79%. Results are shown in the form of a graph in figure 10. Table 6 represents the cross-validation of each technique using the 70% threshold.

Table 6. Average cross Validation of each Data Mining technique (25-minute prediction validity with 70% threshold)

Metrics	SVM	KNN	DT	RF	EL
Precision	64.34	74.35	84.41	81.85	88.03
Recall	50.84	67.35	83.12	72.57	81.51
Accuracy	73.08	79.37	86.81	82.16	87.71
F1-Score	56.79	70.23	83.76	76.94	84.63

"SVM= Support Vector Machine, KNN= K-Nearest Neighbours, DT= Decision Tree, RF= Random Forest, EL= Ensemble Learning"

For a better understanding of the tables, results are shown in the form of figures. In figures 7-10, comparative analysis results of five data mining techniques that are fused with IoT datasets are shown. Comparative analysis is performed using well-known techniques that are SVM, KNN, Decision tree, Random Forest, and

Ensemble Learning. For evaluation of the results, we used 4 parameters that are precision, recall, accuracy, and F1-Score. The K-fold method is used to evaluate the consistency of the specific data model. In the k-fold authentication method, a dataset is divided into 5 equal sets. In the given k set one data set is used as a testing data set and all the remaining data sets are used as training data sets. In our case, we will configure it to generate 1000 samples each with 20 input features, 15 of which contribute to the target variable. An experiment is conducted using 50% and 70%threshold using 15-minute and 25-minute predictive rationality. Using a 50% threshold, accuracy is around 93% and the threshold is increased to 70% then accuracy is around 82%. After analyzing all the results of the data mining techniques it is clear from evaluation parameters that the performance of the decision tree and the ensemble learning is better among all other data mining techniques. And if we compare the performance of the decision tree and ensemble learning technique then the decision tree outperformed the ensemble technique. So we can say that the optimized technique among all the techniques is the decision tree.

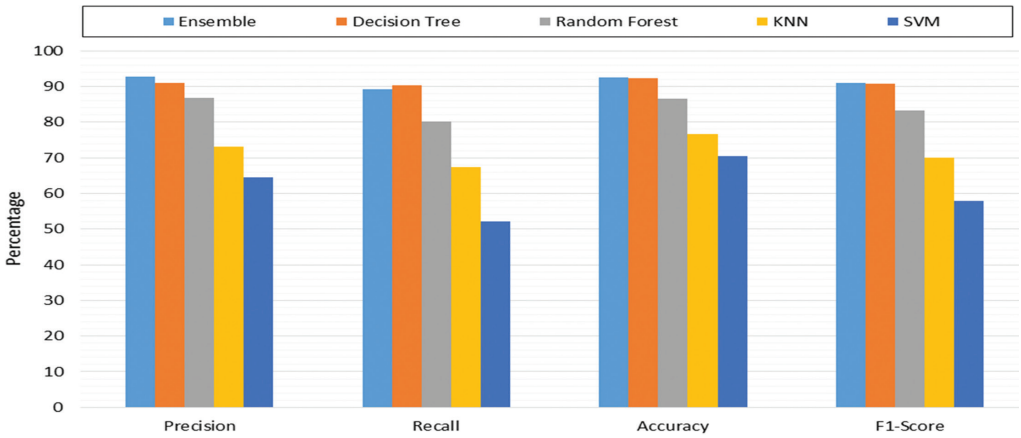

Fig. 7. Comparative analysis using different Data Mining Techniques where predictive rationality=15 minute, threshold=50%

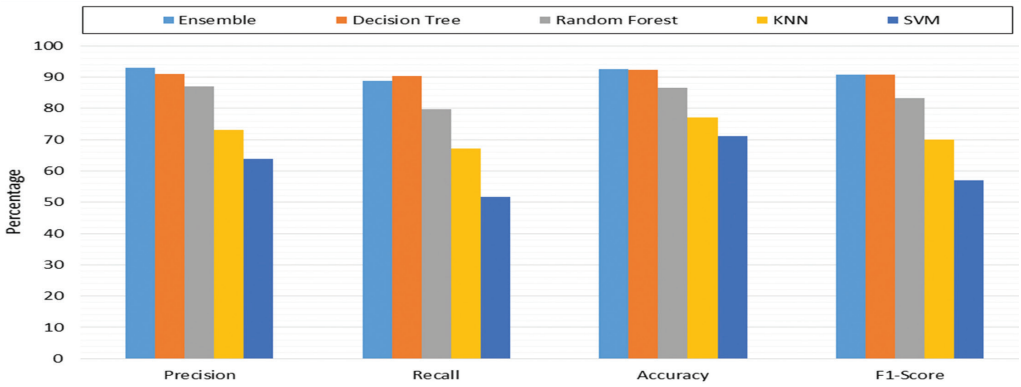

Fig. 8. Comparative analysis using different Data Mining Techniques where predictive rationality=15 minute, threshold=70%

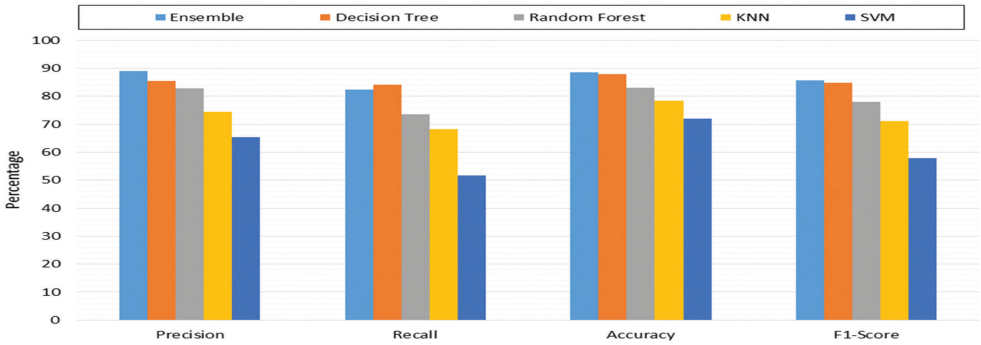

Fig. 9. Comparative analysis using different Data Mining Techniques where predictive rationality=25 minute, threshold=50%

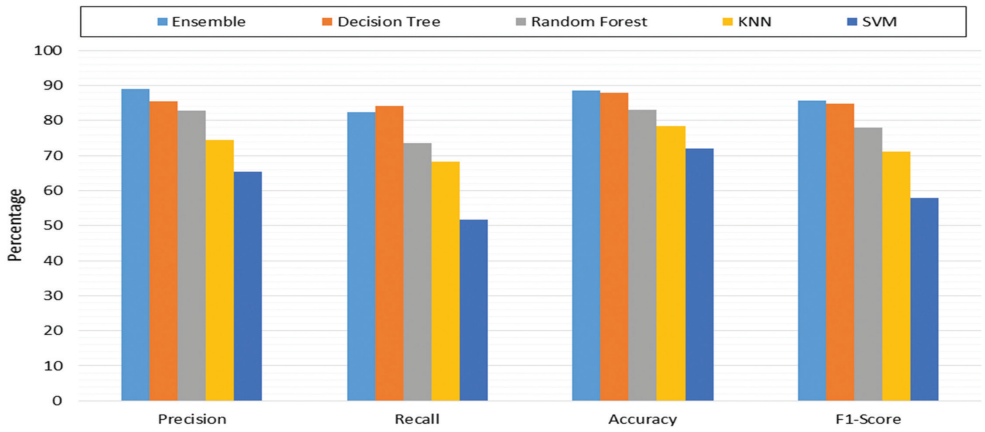

Fig. 10. Comparative analysis using different Data Mining Techniques where predictive rationality=25 min, threshold=70%

6. CONCLUSION

In this paper, we have fused two technologies that are IoT and data mining techniques to predict free parking space. IoT devices generates a large volume of raw data that cannot be recognized for meaningful knowledge unless the data are processed. For that reason, we chose the data mining techniques that will be preprocessed such a large volume of historical data to model the behavior so that it will help to predict the availability of the free parking space. We have analyzed well-known data mining techniques and the originality of the paper is the comparative analysis of the techniques using PK lot data that will imitate the actual environment. The main aim of this paper is to find the optimized data mining technique that will help us to predict the free parking space availability in the parking lot. We performed comparative analysis using well-known 5 data mining techniques: Support Vector Machine (SVM), K-Nearest Neighbor, Decision tree, Random Forest, and Ensemble Learning. The K-fold cross-validation method is used for numerical calculation of the results. For evaluation metrics, we have used precision, recall, accuracy, and F1-Score. An experiment is conducted using 50% and 70%threshold using 15-minute and 25-minute predictive rationality. One of the main aims of the paper is to find the technique that will give better results for predictive analysis. Among all the techniques one of the simple techniques is the KNN mining technique. Based on the result, we can conclude that the decision tree is an optimized solution for the prediction of the availability of the parking space. Ensemble learning is the next closest technique to get better results. So we can say that the optimized technique among all the techniques is the decision tree.

7. REFERENCES:

[1] A. Koster, A. Oliveira, O. Volpato, V. Delvequio, F. Koch, "Recognition and recommendation of parking places", Proceedings of the 14th Ibero-American Conference on AI, Santiago de Chile, Chile, November 24-27 2014, pp. 675–685.

[2] W. J. Park, B. S. Kim, D. E. Seo, D. S. Kim, K. H. Lee, "Parking space detection using ultrasonic sensor in parking assistance system", Proceedings of the IEEE Intelligent Vehicles Symposium, Eindhoven, Netherlands, 4-6 June 2008, pp. 1039–1044.

[3] M. Rinne, S. Törmä, "Mobile crowdsensing of parking space using geofencing and activity recognition", Proceedings of the10th ITS Euro Conference, June 2014, pp. 1–11.

[4] B. N. Narayanan, O. Djaneye-Boundjou, T. M. Kebede, "Performance analysis of machine learning and pattern recognition algorithms for Malware classification", Proceedings of the National Aerospace and Electronics Conference and Ohio Innovation Summit, Dayton, OH, USA, 25-29 July 2016 pp. 338–342.

[5] J. Yang, J. Portilla, T. Riesgo, "Smart parking service based on Wireless Sensor Networks", Proceedings of the 38th Annual Conference on IEEE Industrial Electronics Society, 25-28 October 2012 pp. 6029–6034.

[6] S. Dong, M. Chen, L. Peng, H. Li, "Parking rank: A novel method of parking lots sorting and recommendation based on public information", Proceedings of the International Conference on Industrial Technology, Lyon, France, 20-22 February 2018 pp. 1381–1386.

[7] R. E. Barone, T. Giuffrè, S. M. Siniscalchi, M. A. Morgano, G. Tesoriere, "Architecture for parking management in smart cities", IET Intelligent Transportation Systems, Vol. 8, No. 5, 2014, pp. 445–452.

[8] E. I. Vlahogianni, K. Kepaptsoglou, V. Tsetsos, M. G. Karlaftis, "A Real-Time Parking Prediction System for Smart Cities", Journal of Intelligent Transportation Systems, Vol. 20, No. 2, pp. 192–204, 2016

[9] Y. Zheng, S. Rajasegarar, C. Leckie, "Parking availability prediction for sensor-enabled car parks in smart cities", Proceedings of the IEEE 10th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Singapore, 7-9 April 2015, pp. 7–9.

[10] C. Badii, P. Nesi, I. Paoli, "Predicting Available Parking Slots on Critical and Regular Services by Exploiting a Range of Open Data", IEEE Access, Vol. 6, 2018, pp. 44059–44071.

[11] A. Camero, J. Toutouh, D. H. Stolfi, E. Alba, "Evolutionary deep learning for car park occupancy prediction in smart cities", Proceedings of the 12th International Conference on Learning and Intelligent Optimization, Kalamata, Greece, 10–15 June 2018 pp. 386–401.

[12] F. Al-Turjman, A. Malekloo, "Smart parking in IoT-enabled cities: A survey", Sustainable Cities and Society, Vol. 49, 2019.

[13] R. Sharma, A. Ghosh, P. K. Joshi, "Decision tree approach for classification of remotely sensed satellite data using open source support", Journal of Earth System Science, Vol. 122, No. 5, 2013, pp. 1237–1247.

[14] Z. C. Lipton, C. Elkan, B. Naryanaswamy, "Optimal thresholding of classifiers to maximize F1 measure", Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, Nancy, France, 15-19 September 2014 pp. 225–239.

[15] IBM Global Parking Survey, https://www- 03.ibm.com/press/us/en/pressrelease/35515.wss (accessed: 2021)

[16] "Worldwide Interoperability for Semantics IoT (PK-Lot -IoT)", PK-LoT Dataset https://public.roboflow.com/ object -detection/ pklot/1 (accessed: 2021).

[17] "Efficient tools for predictive data analysis, Scikit Learn", https:// scikit-learn.org/ stable/ modules/ generated/sklearn.model_selection.GridSearchCV.html (accessed: 2021).

[18] T. N. Pham, M. F Tsai, D. B. Nguyen, C. R. Dow, D. J. Deng, "A cloud-based smart-parking system based on Internet-of-Things technologies", IEEE Access, Vol. 3, 2015, pp. 1581-1591.

[19] F. A. Turjman, A. Malekloo, "Smart parking in IoT-enabled cities. A survey" Vol. 49, 2019, pp 101-608.

[20] "Data mining algorithm basic concepts", www.javatpoint.com/k-nearest-neighbor-algorithm-for - machine-learning (accessed: 2021).