Al-Salam Journal for Engineering and Technology
Journal Homepage: http://journal.alsalam.edu.iq/index.php/ajest
ISSN: 2790-4822

Developed High Scale Bagging Algorithm for E-Tourism Advising System

Rula Amjed Hamid[1], Muayad Sadik Croock[2],*

[1]University of Information Technology and Communications, Institute of Informatics for Postgraduate Studies, Baghdad, Iraq

[2]University of Technology-Iraq, Computer Engineering Department, Baghdad, Iraq

*Corresponding Author: Muayad Sadik Croock

DOI: https://doi.org/10.55145/ajest.2022.01.01.005
Received December 2021; Accepted January 2022; Available online January 2022

ABSTRACT: Filtering huge amounts of data is a very critical issue with the explosion of data over the web and cloud storage. A need to classify and sort these data is linked to that issue to facilitate data management and database building for various applications. Machine learning techniques are the most suitable to deal with such big data.

One of the applications that can be implemented in machine learning is a tourist advising system that harvests data from tourism sites and aggregates different types of data about them (humidity, temperature, distance from user's country, etc…) and classifies them. These data should be updated constantly, since the system provides real-time decision based on real-time data, where they are used later on by a bagging system to provide the user with suggested tourism sites with percentage to how suitable these sites according to the preferences submitted in addition to some other criteria.

Keywords: Bagging, E-tourism, Data Mining, Internet of Things, Machine Learning, Advising system

1. INTRODUCTION

A need to classify and sort huge amounts of data is linked to having various and disperses sources of these data, mainly the web and smart devices, with the need to facilitate data management and database building for various applications.

The Internet of Things (denoted by IoT) technology allows the connection and communication of any smart device remotely, and they represent a method of collecting data about people as well as places with ease and comfort to their users.

Data mining is usually related to huge amounts of data; that comes with large volume, complex data combinations that grows and in different data types and storage size. The web and IoT connected devices are considered to be a main source of such data, and finding relationships that binds these data (sometimes not directly) and classifying them into categories is a very hard and time consuming process [1].

There are different approaches to data mining especially in such "big data" and data originated from IoT devices; one of the most famous approach is the "bagging" technique which relies on aggregating predictions of data from different data models in order to have more reliable and more accurate predictions of mined data [2]. This will in turn enhance the accuracy of classification and produce better prediction and more reliable advice. People who plan for vacation or a site-seeing trip usually use the internet before setting off on the trip. As a source of information that helps people in their decision making process, the internet and various web sites that provide different information (mostly not related) like the weather, destination from home, feedback from other visitors is obviously the source of data in an automated system. Hence; integrating data mining with such data for this specific domain shows as an obvious solution towards making such decisions easier and more reliable, and even provide suggestions that weren't even in mind of the users. The internet is used as a source of data to store into the database of the connected smart devices or web site

*Corresponding author: rula.amjed32@gmail.com

through which the users seek help in making the decision. Different information about each location are appended to the location's record that will be used later by the bagging algorithm to produce a sorted list of suggestions according to the preferences of the user. Bagging or bootstrap aggregation was adopted in forecasting systems, mainly in business and economic systems due to the algorithm's ability to deal with variables smoothly, like the one developed by [3] that forecasts tourists' hotel reservation cancellation, which would enhance the revenues of stakeholders of hotels based on certain attributes related to the tourist and the environment. Another application that adopted bagging in forecasting economic time series and inflation, like the research [4] among many others.In the proposed system described in this paper, tourism sites and information about them are collected from the internet, and the machine learning module classifies them according to predefined criteria. The user uses the mobile application or the web site to enter his/her preferences, which are then used by the bagging algorithm to map to pre-stored locations' criteria and give weighted suggestions for each of the suitable sites.

2. RELATED WORK

Big Data is defined as unstructured information which is large and complex to manage and store that is growing day by day. The process of extracting information from huge data sets using various algorithms is called Data Mining, in which many classification techniques are implemented to classify each data item in the datasets to produce predefined set of classes, The web has become a primary source of data for almost every aspect of life, like tourism systems to give an example. Travelers can get destination information and plan their trips with the help of information available on the web. New technologies need to be implemented to facilitate making better use of the growing amounts of data related to this kind of web service.

Web based e-tourism was found to be most appropriate and accommodating to both users and researchers, so [5] introduced the E-tourism web site and how it become an important system in current time because of the growing use to this kind of web service for travel planning. the site first recommends places depending on the user's selection from a list (Through the specific planning preferences (SPP) Module), then the list schedules a planning module producing a filtered recommended activity Controller (FRC).

Mobile applications are considered one of the most helpful equipment in modern life and can be utilized to help in the e-tourism industry. The authors in [6], developed the Mobile recommender systems for tourism with RSS software system which provide accurate situation-aware recommendations by capturing personal and social contextual parameters forecasting tourism demand was the subject of the research in [7] where the authors presented an implementation of Support Vector Regression (SVR) and novel neural network technique to tourism forecasting fields. the study aimed to build an advanced SVR model by implying a comparison between Back-Propagation Neural Networks (BPNN) and the Autoregressive Integrated Moving Average (ARIMA), and they presented the (GA)-SVR algorithm that use Gas value to search about SVR's optimal parameters notice that the users have to define and set the SVR's parameters carefully such as Kernel function so they can build the SVR models.

The decision making process for tourists was presented in [8] through a new implementation for e-tourism system using fuzzy logic and applied to Shiraz city as a case study, since its characteristics meet the tourists' requirements. The web application just asks tourist to fill in related data for their needs in the website since the system extracts data through forms, tests targeted traveling to shiraz in the simplest way.

The work in [9] presented an implementation of fuzzy inference system (Fuzzy decision-making) to help the tourist make a decision to select a hotel in the e-tourism industry, the experiment take a place in India for the international tourist hotel, the authors present a decision model to evaluation committees for location selection.

3. SYSTEM ARCHITECTURE

The system is mainly composed of two modules that integrate to provide the user with a weighted list of suggested tourism sites that meet their preferences. The first module is the machine learning module through which data about available sites for tourism from the internet is collect and classified into a lookup table with properties that are mapped to the users' preferences, nationality, proximity to users' location, in addition to the preference of this site to people from the same region as the user's.

The second module is concerned with providing the user with the decision on what are the suitable sites to visit based on the preferences and other criteria that are fed into the first module. These sites are displayed with weights according to the bagging algorithm that is used to make the decision based on the "degree" of closeness preferences are

to the classified data about each of the stored sites. Figure 1 illustrates the integration between these two modules, and the general system's architecture.

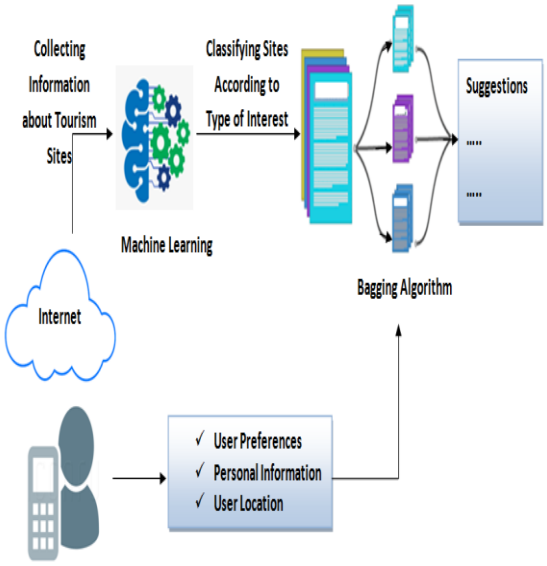


FIGURE 1. - System Architecture

3.1 DATA COLLECTION AND MINING

Large data about suitable places for tourism mainly originate from the internet, and more specifically from web sites that provide advises and display ratings from users around the globe. These provide a good source of data for the machine learning module to collect the data and learn what classes to add to each of the sites.

These data include the site's name and country, location (in terms of longitude and latitude, to facilitate finding the distance from user's current location) in addition to so many other attributes, like closeness to bodies of water, does this site hold religious values, does it have near hotels and accommodate services and so on.

Features extraction of collected data is necessary to find the best classification of each site's attributes, which will later help in producing the best "Advice" to users of the system. Static-predefined rules are used by the machine learning engine to define the correct feature for each site, and add it to the look-up table in its suitable category. Figure 2 shows the process of data mining with Machine learning as part of the process.

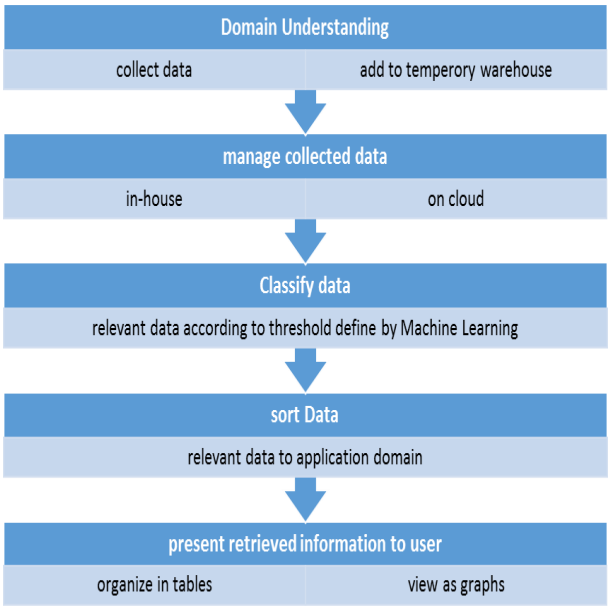


FIGURE. 2 - Data Mining process

The training set used by the machine learning module is organized as Decision Trees, where predicted classes are mapped based on simple rules that resemble the human's way of thinking about problems. For the tourist advising system, to see if a certain location is near a body of water for example, the distance between the center of it and the closest body of water is calculated, if this distance was >150 km, then this place is classified as a dry place that is suitable for people who have fears from water. Another example is when a place contains ruins, and then this place is suitable to tourists who prefer historical or heritage places. A sample decision tree is shown in figure 3.

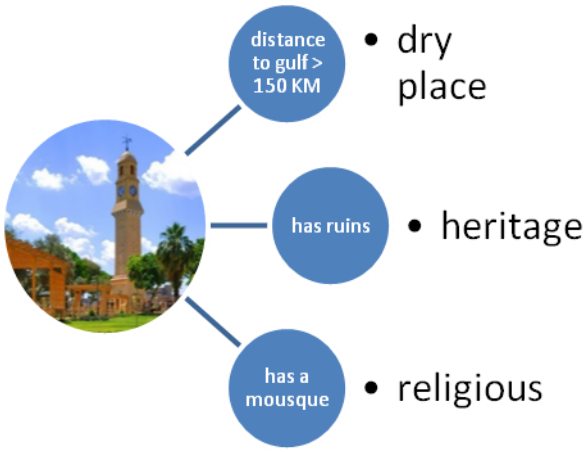


FIGURE. 3 - decision tree illustration for the site "Al Qashla Building" in Baghdad (building image was adapted from Wikipedia.com)

A look-up table is constructed to be used to train the machine learning algorithm and find the best classification for each new site. Some data are justified and classified by a human user are kept apart to test the machine learning's classification. This look-up table will be then fed into the bagging algorithm to find matches with users' preferences (the ones they enter in addition to the ones collected from other smart devices and the internet, like users' current location). A segment of this look-up table is shown in figure 2.

				Environm									
City	Place names	Latitude	Longitude	Adventure	Culture	ental	Health	Nature	Religious	sport	shopping	business	leis
بغداد	الروضة الكنفية	33.379507	44.339878		1				1				
بغداد	جامع ومرقد الإمام أبو حنيفة النعمان	33.371865	44.358385		1				1				
بغداد	جامع برثا	33.351238	44.361236		1				1				
بغداد	جامع ومرقد الشيخ عبد القادر الكيلاني	33.33788	44.410124		1				1				
بغداد	جامع الحفاء	33.338887	44.397739		1				1				
بغداد	منتزه الزوراء	33.314855	44.377068	1	1			1		1			
بغداد	المدرسة المستنصرية	33.338494	44.389989		1								
بغداد	القصر العباسي	33.342913	44.38353		1								
بغداد	مبنى المكتبة	33.341267	44.386225		1								
بغداد	المتحف البغدادي	33.34022	44.389768		1								
بغداد	متحف التاريخ الطبيعي	33.354717	44.392675		1			1					
بغداد	جزيرة بغداد	33.443686	44.343758	1		1		1		1			
بغداد	مدينة العاب السندباد	33.278072	44.311123	1		1		1		1			
بغداد	سور بغداد	33.340998	44.412757		1								
بغداد	عقرقوف	33.353535	44.202124		1								
بغداد	ملعب الشعب الدولي	33.324922	44.43547							1			
بغداد	المسرح الوطني العراقي	33.303843	44.432322										
بغداد	منتزه ابو نؤاس	33.312993	44.417574					1		1			
بغداد	كورنيش الاعظمية	33.364252	44.369969					1		1			
بغداد	تل حرمل	33.323218	44.482462		1								
بغداد	المدائن	33.098498	44.581211		1				1				
بغداد	ساحة بغداد	33.311676	44.364501								1	1	

FIGURE. 4 - Classified data about sites

Theses classifications are used by the bagging algorithm to help make a decision on which are the most suitable places for the tourist to go, given the preference they submit to the system and other information that are collected from IoT devices related to the user.

3.2 DECISION MAKING WITH BAGGING

Bagging algorithm (also known as Bootstrap Aggregation), as one application of ensemble algorithms, allows the system to make a decision by aggregating different, sometimes, heterogeneous models of the data samples into a uniform decision with percentage on level of confidence about the decision. Ensemble algorithms are used to get a better understanding of the various biases, variances and features that exist in the original data samples to draw decisions and produce output to the user.

According to [10] bagging provides the highest predicted values accuracy in big data, due to the accumulation of accuracy measures in the training sets calculated separately.

Using Bagging in decision making reduces variance of data models elements that the system was trained on, in addition to new data elements that are input to the system to map the learnt data sets to other data sets that are "new" to the system, to produce a decision more quickly and efficiently with minimal storage and processing requirements. Figure 4 shows how bagging is done.

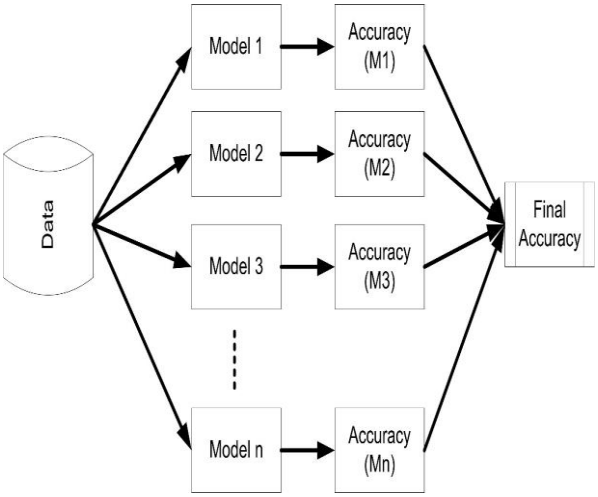


FIGURE. 4 - flowchart of bagging training sets to get the final estimation

In machine learning, input data are separated into a set of training data and a segment for testing as in any other machine learning algorithm, these training sets are divided into subsets by randomly sampling data and assigning them to different training sets; this generate a classifier to be included in the data mining model.

The accuracy of estimation is calculated for each sample set (data model) and then all these estimations are averaged to reach the final estimation of accurate data. So the trained data and the data that are entered by the user (both the users' preferences and the data collected by the IoT devices) are added to the samples of the data and resample to get the accuracy of estimation for each set of data model.

4. SYSTEM IMPLEMENTATION

The tourist advising system uses machine learning and bagging algorithm in the back end as mentioned earlier, while the user can input his/her preference through a user-friendly web page or mobile application, while connected IoT devices (like the GPS sensor mounted on the mobile phone, or the temperature thermometer in the user's area) submit simultaneously with the user. The snapshots in figures 6 and 7 show the main page of the system as a web page, and on android mobile emulator.

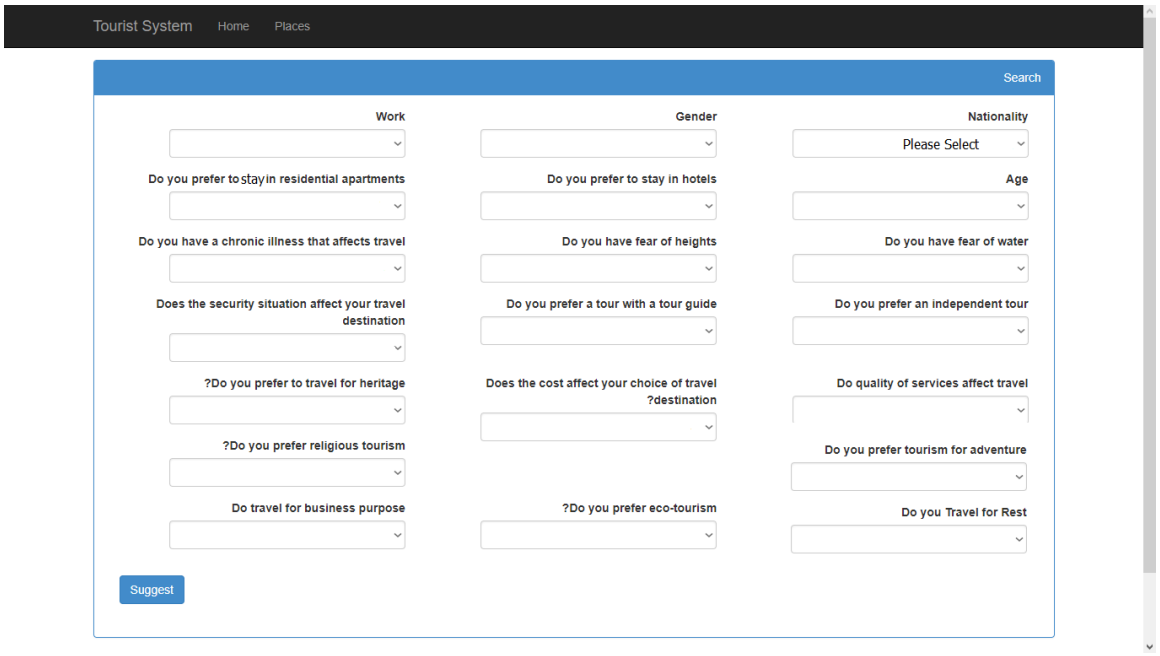


FIGURE. 5 - System's Web Page Screenshot

The entries in the web page and the mobile application are almost the same that the system used as features to classify the sites. As the users select their preferences from the application, the bagging algorithm uses these preferences and makes a decision on which are the best sites that suites this user. The results are produced as a sorted list from the most suitable to the least suitable, while also considering national preferences of the tourism site (how many people with the same user's nationality have visited this site within the past year).

The results show only the sites that match the users' preferences and hide all other sites stored in the database. Figure 8 shows the user's selections against the results that were displayed.

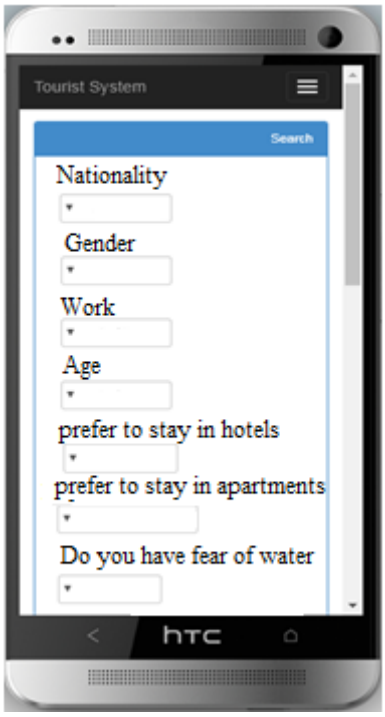
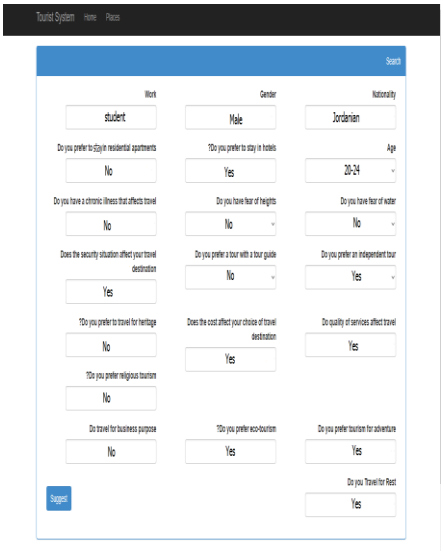


FIGURE. 6 - System's Android's Application Screenshot

(a) user entries



(b) displayed results in sorted format

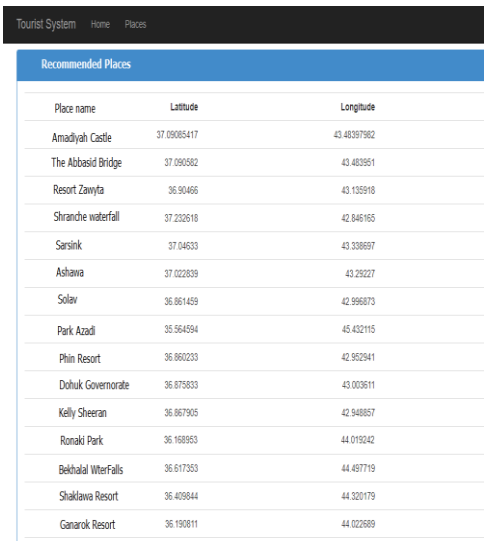


FIGURE. 8 - user's preferences and how they were mapped to resutls

5. SYSTEM TESTING

The system was tried by several users from different background and with various preferences from different locations. All users gave positive feedback about the system starting from the user-friendly and easy to use application, to the satisfying results that they thought were suitable and would like to go on a tour to.

A small set of data that was used as training set by the machine learning algorithm has been classified by a human user and measured for accuracy, recall and positive predictive values (PPV), and the values are arranged in a confusion matrix. The results show that the system has very rates of accuracy and recall, as well as a high PPV. The data in table 1 lists the values of the confusion matrix that was used to test the efficiency of the machine learning procedure.

Table 1. - Confusion Matrix for testing the Machine Learning module

		Actual Class: P	Actual Class: N	
Predicted Class	P	True Positive(TP) 15	False Positive(FP) 3	
	N	False Negative(FN) 0	True Negative(TN) 7	25

As stated earlier, a subset of the data that was used to train the machine learning algorithm to build the decision trees was used and evaluated by a human user to classify according to the criteria established for the preferences of users. True Positive values indicate the number of sites (in the test data set) that both the system and the user agreed it is the most suitable site based on preferences (produced highest percentage in the system), while the True Negative values indicate the number of site where both the system and the user gave lowest values as being a suitable tourism site.

The accuracy of the system is calculated based on the values of the confusion matrix as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

When applying equation (1) the system registered an accuracy of 22/25 = 88% Other efficiency measurements are computed, and the results show that the recall (sensitivity) of the system registered 69% (TP / (TP+FN)) and PPV of 83% (TP/(TP+FP)). These figures are good indications that the system's performance is very good and satisfactory to users, where the suggested sites meet their preferences in addition to other factors they don't submit directly.

6. CONCLUSIONS

Tourist guidance and advising system was developed in this research to assist users who seek advice on tourism sites based on their preference. The system uses machine learning decision trees to build a look-up table with a list of sites, and classifies these sites with certain properties that most people look for when going on tourism trips.

For the decision making process, bagging algorithm was applied to the users' submitted preferences in addition to other criteria collected through IoT devices in the users' are like location and current temperature, and mapped against the look-up table built earlier to display a list of suggested sites sorted from the most suitable to the least suitable places.

The user friendly interface of the system in addition to its mobility through a mobile application, with the high percentage of accuracy and recall values, makes it a suitable system to adopt by tourist, the quick response and fast execution of the system adds to the benefits of adopting such system.

REFERENCES

[1] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

[2] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

[3] Morales, D. R., & Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. European Journal of Operational Research, 202(2), 554-562.

[4] Inoue, A., & Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of US consumer price inflation. Journal of the American Statistical Association, 103(482), 511-522.

[5] Sebastia, L., Garcia, I., Onaindia, E., & Guzman, C. (2009). e-Tourism: a tourist recommendation and planning application. International Journal on Artificial Intelligence Tools, 18(05), 717-738.

[6] Gavalas, D., Konstantopoulos, C., Mastakas, K., & Pantziou, G. (2014). Mobile recommender systems in tourism. Journal of network and computer applications, 39, 319-333.

[7] Chen, K. Y., & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. Tourism Management, 28(1), 215-226.

[8] Hamedi, Z., & Jafari, S. (2011). Using fuzzy decision-making in e-tourism industry: a case study of Shiraz city e-tourism. International Journal of Computer Science Issues, 8(3), 1.

[9] Gupta, G., Srivastava, M. K., & Kumar, S. (2016). Decision on the Selection of Tourist Hotels for an Hotelier Using the Concept of Fuzzy Inference System.

[10] Budhani, S. K., Jha, C. K., & Ahmad, A. (2018). Comparative Study of Meta Classification Algorithm: Bagging, AdaboostM1 and Stacking with Concept Drift based Synthetic Dataset Hyperplane1 and Hyperplane2. International Journal of Engineering Science, 15927.