VEGUETA

Vol 9 Issue 01,June 2021
ISSN NO: 1133-598X

Using Deductive Decision Trees to Make Trip Suggestions

Dr. V Venkata Ramana[1], Dr S Nageswara Rao[2], B Gouri[3] S Riyaz Bhanu[4]

[1] Professor, [2] Assoc. Professor, [3,4] Asst. Professor
Department of CSE, K.S.R.M College of Engineering(A), Kadapa

Abstract—

Choosing a holiday destination from the abundance of available information is one of the most challenging aspects of traveling nowadays. While packing for a vacation and while traveling. Many previous Travel Recommendation Systems (TRSs) attempted to solve this problem. The technical aspects, such as the accuracy of the system, have been prioritized above the practical aspects, such as usability and enjoyment. To address this issue, we need to develop new models of how visitors search for information. In this research, we propose a novel TRS that puts people first and makes it easier for them to go through an unfamiliar city. We weigh technical and practical factors using a dataset we collected from the real world. The system is built using a two-step feature selection technique to reduce the number of inputs and provide recommendations using decision tree C4.5. The results of the tests show that the proposed TRS is capable of providing highly personalized suggestions for enjoyable holiday destinations.

I. INTRODUCTION

In 2013, tourism generated 9.5% of the world's GDP, making it a highly significant economic sector. Predictions for the travel industry are optimistic. Provide a GDP-boosting impact of almost 10.3 percent in 2023. When it comes to the economic impact of the travel and tourism industry, South East Asia is predicted to have the most expansion. Specific nations with the most appealing tourist attributes in 2013 were selected as Thailand, Indonesia, Singapore, and Myanmar [1].

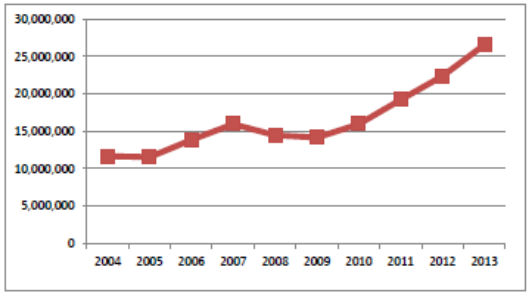


Figure 1: The Annualized Number of Overseas Visitors to Thailand, 2004–2013 [1]

The influx of tourists from other countries has doubled in recent years. In the last fifteen years (Fig. 1). In 2013, Thailand was the tenth most popular tourist destination worldwide [1]. With 26.5 million visitors from outside in 2016, the industry is seeing an increase of 18.76% compared to 2012 [2]. The government of Thailand has made it a priority to increase tourism arrivals and earnings from tourists from all over the world. In 2013, the tourism sector in Thailand generated revenue of 1.79 trillion BHT ($55.49 billion).Tourists now go to the Internet more than any other medium to research local establishments and their services [3]. Visitors may feel overwhelmed by the sheer volume of information available online while searching for locations (also known as trip planning). There are a lot of factors to think about while planning a trip, such as the quality of attractions, travel routes, hotels, numbers of tourists, number of travelers, leisure activities, weather, etc. Recent years have seen substantial improvements in the tourism business because to technological advancements, especially the Internet [5]. Recommendation Systems (RS) are a kind of decision-making technology that has simplified the process of finding a desired service or product, narrowing down available alternatives, comparing those options, and ultimately settling on a course of action for both consumers and service providers. To far, most TRSs have focused on providing tailored cost estimates for

booking travel-related services (such as restaurants, hotels, and transportation) at a specific destination, for a certain number of travelers, at a specific time. When it comes to technology, these TRSs only provide elementary methods for narrowing results to those that meet the user's exact criteria. On the other hand, they lack necessary theoretical and practical components (such as sparsity, scalability, transparency, system correctness, theories to improve personalisation, etc.). A significant challenge in developing a TRS that provides customized recommendations of tourist destinations is enhancing the traveler's capacity for decision making. In order to achieve this goal, it develops personalized models of visitors' information-seeking behavior and requires an in-depth understanding of the factors that influence their choices. During the research portion of a tourist's decision-making process, it is also crucial to allay any concerns they may have. Reducing the total number of system parameters might simplify the model. As a consequence, the system's ability to provide useful suggestions and the satisfaction of its users may both rise.

II. BACKGROUND

A. Method for Making Suggestions

Recommendation systems (RSs) are a kind of decision support system (DSS) that may provide advice on what to do next. Product depending on the user's choices as a whole [6]. It helps people out by giving them resources to utilize in making choices that are meaningful to them and address their issues [7]. Many well-known online retailers, like Amazon, Netflix, Pandora, etc., use RS extensively. The e-commerce RSs will recommend things according on the user's interests in news, publications, individuals, URLs, and so on [8].

B. Trip-Planning Software

Tourists have challenging decision-making processes because to the abundance of destinations, attractions, activities, and services from which to choose. This is why both academic and industrial researchers are interested in TRS. Many different kinds of platforms (including desktops, browsers, and mobile devices) have witnessed the creation or implementation of various TRS. Users may rate services and routes in order of preference, evaluate their level of interest in various destinations, choose their top-choice points of interest (POIs), and organize their whole trips with the aid of a TRS. Most TRSs are made with the individual traveler in mind [9], but there are a few that are made to help travel companies. While both have a foundation in similar frameworks, the technologies, philosophies, data inputs, interaction styles, and personalisation used by each are distinctive. The general structure of modern TRSs is seen in Fig. 2.

Data acquired from several sensors, GPS locations, questionnaires, evaluations, etc., are all housed in the repository. Each component of the recommendation engine—be it optimization, statistics, artificial intelligence, and so on—has its own function. The objective is to present the user with suggestions, rankings, or projections that take into account not just their wants and desires, but also any relevant hard and soft limitations (such as user demographics, trip duration, available cash, trip kind, etc.). The TRS often gathers the traveler's inputs (implicit, explicit, or both) before and throughout the journey in order to generate a user profile and send back the suggested outcome. Tourists may use the system's output in a number of different ways, such as displaying symbols denoting locations on a map interface, following a point-to-point route, or following an agenda or timetable. Typically, TRSs will use a spatial web service, such as Google Maps API, to show the conclusion. Newer TRSs take into account the user's location and the weather to personalize their recommendations. Some TRSs allow for tweaks to be made by the user and for the output to adapt based on ratings provided by the user [10, 11].

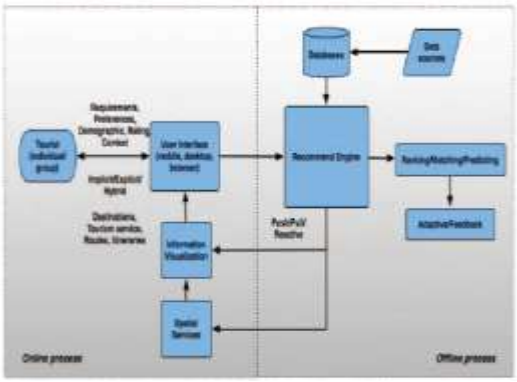
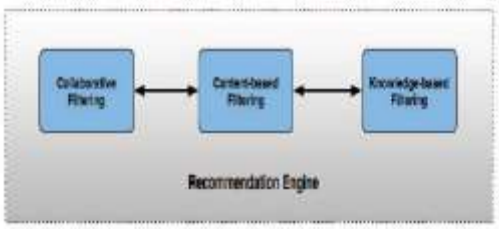


Figure 2.Overall structure of vacation advice apps

Procedures for making suggestions as stated in [12], RS may be categorized in terms of its severity. Of customization, such as how helpful and precise the suggestions are. No personalization, temporary personalization, and long-term personalization are all ways in which the level of customization may be described (long term). A no personalized RS is a basic system that makes suggestions without considering the user's tastes. For example, the RS only compiles a list of the most sought-after products (i.e., editors' picks or best-sellers) based on the total number of reviews and/or sales. Thus, the suggested outcomes would be helpful for other general users of the system. Non-personalized systems have not been a focus of RS research because of their lack of autonomy in making decisions [7]. An ephemeral and personalized RS is superior than a no personalized RS in terms of the inclusion of information connected to the system's users (i.e., user preferences, sociodemographic information, etc.). As a result, each user would be presented with a unique set of suggestions. As an example, Trip-advisor1 would suggest a place to visit based on the user's profile data, including their age, gender, and marital status. Previous studies have examined a wide variety of tailored RSs, and researchers have classified them based on the information-filtering mechanisms they use [7, [13]-[15]. Following this, we'll take a quick look at the recommendation engine (Fig. 3), which is built from a variety of recommendation methods based on research from [14]. We will talk about the merits and demerits of each, as well as the hybrid filtering strategy used (i.e., the interconnection of many RSs).



a) Collaborative filtering: This strategy is generally utilized by the most deployed recommendation engines (see Figure 3). Systems. Users with similar characteristics are taken into account when making recommendations, and well-liked products are also suggested. However, this method still has a cold-start issue since it requires an initial rating of the new item or user before making a suggestion. The second kind of recommendation strategy is called "content-based filtering," and it makes suggestions to the user based on the user's prior searches and queries. The user has to start from scratch and submit a lot of information before the algorithm can provide a suggestion, which is the biggest negative. Unless a sufficient amount of previous data has been stored, the system will not be able to provide reliable findings [13]. Overspecialization is another prevalent issue [7] due to the system's tendency to propose the item that the user loved the best. Expertise-based filtering (c): Recommends products to the user based on prior subject knowledge. That is to say, the system understands how the item pertains to the person in question. Case-based reasoning and ontological approaches are particularly useful for this purpose. Both [9] and [16] use systems that use the prior experiences of travel firms and groups of experts to provide recommendations.
All the above-mentioned recommendation methods have their advantages and disadvantages, which is why d) hybrid filtering was developed. Reasons for Advising a Hybrid Approach the goal of combining techniques is to maximize

performance while eliminating any drawbacks to one approach. In addition, there are a plethora of hybridization approaches, such as the mixing of several recommendation systems (weight, switching, mixed, feature combinations, cascades, feature augmentations, and met levels) [13]. Researchers now have the tools they need to create a TRS that is intelligent, interactive, and adaptive; that can be automated; that can support a higher level of user satisfaction than ever before thanks to advancements in ICT like Artificial Intelligence (AI), the Semantic Web, communication networks, and so on. To that end, we're working on a system design to meet those goals.

III. METHODOLOGY

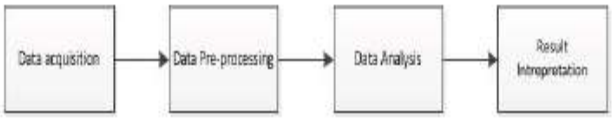


Figure 4. Data Mining Framework

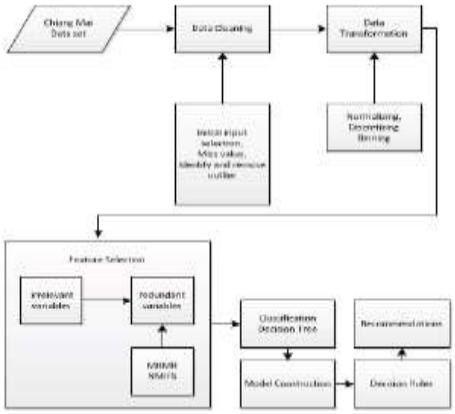


Figure 5. Theoretical framework for the targeted radio spectrum
Figure 4 depicts the proposed DM architecture, which includes four stages: data collecting, data preparation,

analysis of data and speculation about its implications. (1) In order to gather information, a four-part questionnaire was created, distributed, and collected in Chiang Mai, Thailand. (2) The collected data undergoes a number of pre-processing operations before it is utilized, such as data cleaning, data transformation, and feature selection techniques. Finally, a C4.5 decision tree is utilized as a classifier in the third and final step. down this last phase, we focus on honing down on the most helpful features and locating the most effective models. Finally, we'll examine the rules of thumb and optimal decision trees that we've developed. The whole procedure is shown in Fig. 5.

Information Collection In order to learn more about tourists' search behavior while analyzing travel information and their decision-making process when picking a place, we choose to use a questionnaire as our data collecting method. Researchers ran preliminary studies to determine the many factors that influence vacationers' top choices of destinations to visit. This allowed them to better adapt the questionnaire to probable responders.

The four parts of the survey all focus on various sets of characteristics that are important to the destinations that people choose to visit on vacation. Following:

1) Trip characteristics: these are the most important [17]. Relevant aspects include the length, purpose, and substance of the trip. Travelers' mental state, worldview, and bank account balance all have a role in their ultimate choice [17]. Third, vacationers' motivations are a key factor in determining where they go on vacation, as shown by reviews of relevant research. The reason why tourists want to visit a certain place [18]. Information on tourists'

socioeconomic status and other personal characteristics that can influence their propensity to seek out new knowledge [19]. It was decided that 4,000 surveys would be distributed and collected from visitors to Chiang Mai's top five attractions. User feedback from the travel website Trip Advisor was used to generate this list of the most well-liked destinations. The survey was sent to both international (60%), and domestic (40%). Notable tourist destinations were Art in Paradise (27.7%), Mae SA Waterfall (22.06%), Hay Tung Tao Lake (19.18%), the Museum of World Insects and Natural Wonders (16.97%), and Boa Thong Waterfall (14.09%). The majority of responders (65%) finished the survey in within 30 minutes. For 3,695 valid surveys, 145 variables were input during the data pre-processing stage; 35 samples were thrown out owing to inadequate data.

This proposed framework classifies the tourist's ideal destination using characteristics gathered from questionnaires on the tourist's interests, budget, and personality. This page provides a demographic breakdown of visitors with information on their typical activities, spending habits, and motivations for taking trips.

Preparation and Cleaning of Data

In the actual world, it is not unusual for data to be incomplete, noisy, and inconsistent. Mistakes in data entry or the submission of incorrect information by respondents who seek to preserve their identity are both possibilities in surveys like the ones we conduct. For precise labeling, high-quality data is required. To complete the task at hand, we used feature selection approaches for data integration, data cleaning, data transformation, and variable selection. Feature selection, also known as variable selection, is the process of narrowing down available variables to those that best describe the target classes. This process is essential for better accuracy as well as greater efficiency and usefulness. In this investigation, we attempt to keep as much information as possible while using as few variables as feasible. The objective, therefore, is to boost classification model performance with little more input from the user. In this study, we provide a two-stage filtering strategy for prioritizing features and removing duplicates based on Mutual Information (MI). Measures of similarity (MI) are used to characterize the significance and redundancy of variables in feature selection. The MI value would be zero if the variables were unrelated. In most cases, the dependent variable's significance increases as the MI value increases. It can be shown that p(x) and p(y) are the marginal probability distribution functions of X and Y, respectively, and that p(x, y) is the joint probability distribution function of X and Y. The MI is written like this:

$$MI(X;\ Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log\ \left(\frac{p(x,y)}{p(x)p(y)}\right)$$

1.) The Basic Filtering Technique

The first stage of filtering is meant to order the variables and get rid of the ones that aren't independent. Separated from the dependent variable. To filter out superfluous information, we used the Max- Relevance feature selection method [20], using MI as the metric of choice. For every set of explanatory and criterion variables, we calculated the MI score. To exclude characteristics that contributed less or were unrelated to the predictive capacity, we sorted them in decreasing order and applied a threshold value (the threshold value is determined manually).

2) An alternate filtering strategy

As a second round of filtering, we employed the mutual information-based feature-selection algorithms Minimum Redundancy Maximum Relevance (MRMR) [20] and Normalized Mutual Information Feature Selection (NMIFS) [21] to get rid of superfluous information. Taking the highest MI G value into account, we determined that this was the best feature space. When G 0, further feature selection will be halted.

a) Algorithm for MRMR

The MRMR method [20] is based on the concept of utilizing the MI value to order features according to minimum redundancy and maximum relevance. Redundancy between features and their relevance to a class are both calculated by MRMR. It may be stated as (1).

$$MRMR = max_{1\epsilon\Omega_s}[I(i,h) - \frac{1}{|S|}\sum_{j\epsilon s} MI(i,j) \quad (1)$$

$$MI2(i;j) = \frac{MI(i;j)}{\min\{H(i),H(j)\}} \quad (2)$$

$$NMIFS = max_{1\epsilon\Omega_s}[I(i,h) - \frac{1}{|S|}\sum_{j\epsilon s} MI2(i,j) \quad (3)$$

3. Analyzing the Information

The suggested TRS uses a decision tree as its classifier/model because of the many advantages it offers. A decision maker, like ease of use and clarity. The decision-making process is modeled as a flowchart, making it simple to grasp. In terms of technological considerations, it solves the sparsely and scalability problems plaguing the TRS. There are nodes and leaves in the decision tree. Test set instances begin their journey to a leaf node at the root node. Internal nodes entail checking a specific property, which results in a binary or multi-way split. A class label (the result of the classification) or the instance's ultimate verdict from the test data is represented by the leaf nodes. [22]. we must follow the decision tree from its trunk all the way out to its leaves before we can confidently advise tourists on where to go. There are several decision tree algorithms available, such as Hunt's algorithm, Top-down Induction of Decision Tree (TDIDT), ID3, CHAID, CART, and C4.5. The criteria for splitting, the extent of pruning, the types of characteristics, etc., are all different.

IV. EXPERIMENT DESIGN

1. Data set representation

Details about the data set utilized for this analysis are provided in Table 1. There are five travelers' records in the dataset. Locations of choice. In spite of include all five locations in the decision tree model; the classification accuracy was just 36.1%. The decision tree model was also overly complicated, with a high tree size and a significant number of leafs, both of which made the model opaque to the decision-maker. This multi-classes classification problem is broken down into manageable chunks by first learning which types of tourists visit which cities; then using that data along with insights from Chiang Mai's tourism experts and Trip Advisor to determine which cities are most popular with each of those tourists. This led to the development of the two groups, which are shown in Table 2. Decision tree models were built using these taxonomies as inputs. The data from the Museum offers a binary classification challenge, whereas the data from Nature presents a multi-classification problem. Because both types of museum in the Museum data set serve distinct purposes, we divide them into distinct categories. There are a total of three categories in the Nature dataset, with two of them representing the waterfall and one representing the lake.

TABLE 1. CHARACTERISTICS OF THE DATA SET USED IN THIS STUDY

Data set	# Features	# Classes	# Sample
Tourist destination choice	145	5	1,632

B. Preparing the Data

The data cleansing procedure starts with the first selection. At this stage, we use what we've learned from the tourist industry to filter out characteristics that aren't directly linked to the final product. Next, we subject both sets of data to missing value analysis. The binning technique was used to discredited continuous variables. With a bin size of 10, we can divide the data well. In order to standardize some of the discrete variables, we tapped the expertise of professionals in the tourist industry. The suggested two-stage filtering procedure was implemented once the data set had been cleansed and modified. This was completed in an effort to cleanse the data collection of superfluous or duplicative elements. In the first filtering stage, we employed between 17 and 18 criteria to determine which attributes were most important, depending on the data set. The features in the subset were then run through the MRMR and NMIFS feature selection algorithms to get rid of the ones that weren't necessary.

C. Grouping and Model Building

We used a decision tree to build a classifier after removing superfluous characteristics and honing in on the intended ones. The effectiveness of the two feature selection methods in C4.5 is analyzed.This study used a technique known as K repeat holdout. Sixty percent of each data set was chosen at random for training, whereas 20 percent was stratified (i.e., the percentage of each class in the training, validation, and testing sets is the same). Training and validation sets' prediction accuracy over iterations was averaged. Finding the best models for each batch of data requires a variety of confidence level settings for decision tree pruning. With a step size of two, the confidence levels may be anywhere from 0.1 to 0.5. In terms of validation sets, the best-case scenario is when their mean accuracy is highest. Second, the validation set's mean accuracy must be lower than the training set's mean accuracy.

V. RESULTS AND SYSTEM EVALUATION

C4.5's categorization rate results are shown in Table 2. The Museum dataset was successfully classified at an 80% rate using the single best learner. An overall classification success percentage of 49.72% was found in the Nature dataset. When comparing the two feature selection techniques, NMIFS is regarded as more effective than MRMR for both datasets.

TABLE 2. ACCURACY RATE FOR EACH DATA SET

Data set	# of classes	#Sample	Confidence level	Single best learner accuracy rate
Museum	2	729	0.39	80%
Nature	3	903	0.24	49.72%

As may be seen in Fig. 6, the results of the data pre-processing on the Museum dataset are shown. The MI value from the first filter technique is shown in Fig 6(a), where the threshold was 0.022, 128. The data set was cleaned up by removing variables. The MI G values for both feature selection methods are shown in Fig. 6(b). Whenever a negative number was reached, feature selection ceased.

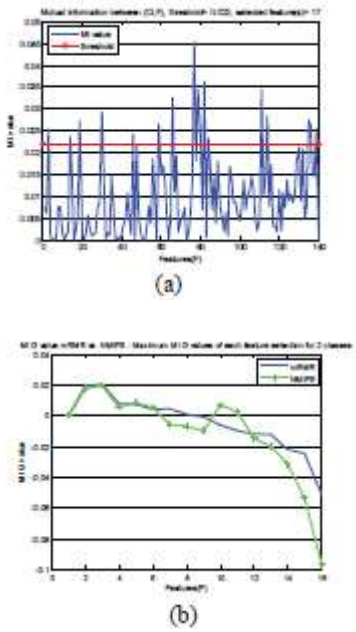


Figure 6: The MI value (a) and the MI G value (b) from the Museum data set's two-step feature selection approach Feature selections from each are shown in Table 3.

Methods for selecting features from the Museum's data collection. Features that are part of the "optimal subset" are denoted by bold variables. After using the second filtering strategy, the MRMR algorithm chose eight ideal features from the Museum data set, whereas the NMIFS chose six. Feature a stands out as the most crucial. Assuming that one of the museums specializes in insects provides an explanation for this observation. Three features (c, d, and b) were used in tandem to aid in the data set's classification. Using a combination of four characteristics taken from the NMFIS, the best decision tree for the Museum dataset is determined, and decision rules are constructed (See Fig 7 and 8). Due to its small size (17 nodes) and plenty of leaves (10 in total); the resulting decision tree is considered very easy to comprehend. An analysis of the Nature dataset revealed that b2 (travel goal) is the most crucial variable to consider.

TABLE 3. FEATURE RANKING BASED ON THE MRMR AND NMIFS ALGORITHMS (MUSEUM DATA SET)

Algorithm	Selected feature
MRMR	*a c d b e f g h i j n k l m o p*
NMIFS	*a c d b g h f j i o k e n l m p*

a:deepest impression is wildlife b:to visit place I have never been before c: The people who are companying are friend d: books and guides influences your decision to visit Chiang Mai

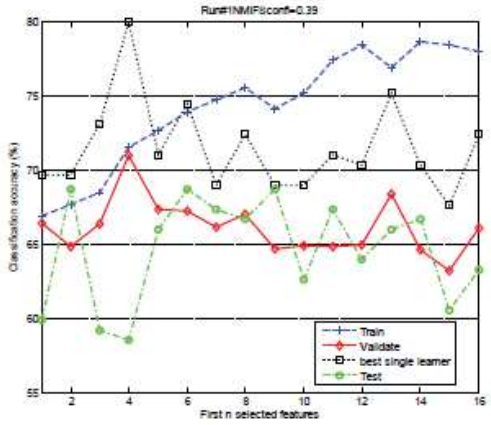


Figure 7. Accuracy rate for the Museum data set

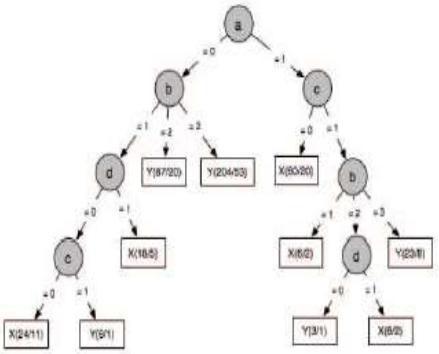


Figure 8: The validation data-driven best-case scenario decision tree for the Museum dataset. (Art in Paradise, Chiang Mai 3D Art Museum and Museum of World Insects and Natural Wonders, X and Y, respectively) The confusion matrix, which includes both the true and anticipated classifications made by the best decision tree, is used with the accuracy rate to assess the model's efficacy. There were a greater number of false positives (samples from the Museum of World Insects that were wrongly labeled as samples from the 3D Arts Museum) in the Museum of World Insects, as seen in Table 4.

TABLE 4. CONFUSION MATRIX OF THE MUSEUM DATA SET

		Predict	
		Museum of world insect	3D arts Museum
Actual	Museum of world insect	26	21
	3D arts Museum	8	90

To aid in decision making, the derived optimum decision tree is used to produce decision rules of the Museum data set, which are then presented as Table 5 shows. For the Museum dataset, eight rules are developed.

TABLE 5. THE DECISION RULES OF THE MUSEUM DATA SET

```
if a == 0, then
        if b==1 then
                if d==0
                        if c==0 then , class = X;
                        elseif c== 1 then, class = Y;
                        end
                elseif d==1 then, class = X
                end
        elseif b==2, then class = Y;
        elseif b==3, then class = Y;
        end
elseif a == 1
        if c==0 then, class = X;
        elseif c==1
                if b==1, then class = X;
                elseif b==2
                        if d == 0 then, class = Y;
                        elseif d==1, then class = X;
                        end

                elseif b==3, then class = Y;
                end
        end
end
```

VI. CONCLUSION

In this research, we introduce a decision tree-based tourist recommendation system to address the problem with existing TRS for specific destinations. The employing expertise in the tourist industry, the data set was partitioned into two sub-sets. This was executed to lessen the decision tree's complexity and improve its categorization accuracy rate. As a result of analyzing the data from NMIFS, the most accurate and straightforward (i.e., smallest in terms of number of leave and overall size) decision trees possible have been crafted for final destination selection. Rules for making selections were mined from decision trees. It is clear that NMIFS is the superior strategy since it employs a smaller feature set than MRMR does while dealing with both data sets. In conclusion, testing findings support the practicality of the suggested TRS. The needs of visitors coming to or already in Chiang Mai are met by the projected TRS. To further improve the data sets' categorization accuracy in future study, several classifiers might be evaluated. As an added bonus, a front-end web application with an interactive and adaptable UI will be developed.

REFERENCES

, "Economic Impact of Travel & Tourism 2014 Annual Update: Summary." International Tourism Organization. According to [2] "Thailand Annual Report 2013."
J. Hosp. Tour. Technol., vol. 4, no. 3, pp. 211-227, 2013. [3] E. Pentane and L. D. Petro, "From e-tourism to f-tourism: emerging issues from negative tourists' online reviews."

(4) B. Pan and D. R. Fesenmaier, "Semantics of Online Tourism and Travel Information Search on the Internet: A Preliminary Study," Inf. Commun. Technol. Tour. 2002 Proc. Int. Conf.Innsbr. Austria 2002, pp. 320-328, Jan. 2002.

According to E. Pitoska (see reference 5), "E-Tourism: The Use of Internet and Information and Communication Technologies in Tourism: The Case of Hotel Units in Peripheral Areas," Tour. South East Eur., volume 2, pages 335-344, December 2013.

"Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids," by G. Häubl and V. Thrifts, published in Mark. Sci., volume 19, issue 1, page 4, Winter 2000.

[7] F. Ricci, L. Roach, and B. Shapiro, "Introduction to Recommender Systems Handbook," Recommender SystemsHandbook, F. Ricci, L. Roach, B. Shapiro, and P. B. Kantor, Eds. Springer US, 2011, pp. 1-35.

According to [8] "Recommender Systems," by P. Redneck and H. R. Varian, published in Common ACM, volume 40, issue 3, pages 56-58, March 1997.

(9) Alptekin, G. I., and Buyukozkan, G. "An integrated case-based reasoning and MCDM system for Web-based tourism destination planning," EXPERT Syst. Appl., vol. 38, pp. 2125–2132, 2011.

According to [10] "Mobile application to provide personalized sightseeing tours," R. Analects, L. Figueiredo, A. Almeida, and P. Novas published in J. Newt.Compute.Appl., vol. 41, pp. 56-64, May 2014.

According to [11] "e- TOURISM: A TOURIST RECOMMENDATION AND PLANNING APPLICATION," published in "Int. J. Art if. Intel. Tools," volume 18, issue 5, pages 717-738, October 2009.