Open Access **Baghdad Science Journal** **P-ISSN: 2078-8665**
2021, Vol. 18 No.1 (Suppl. March) **E-ISSN: 2411-7986**

DOI: http://dx.doi.org/10.21123/bsj.2021.18.1(Suppl.).0737

Application of Data Mining Techniques on Tourist Expenses in Malaysia

***Cai Miao*[1*]** ***Tan Shi An*[2]**

[1]University Sains Malaysia, China.
[2] University Sains Malaysia, Malaysia.
[*]Corresponding author: caimiao000@gmail.com[*],aarontanshian@gmail.com
[*]ORCID ID: https://orcid.org/0000-0001-5492-3028[*] , https://orcid.org/0000-0002-5769-5187

Received 11/12/2020, Accepted 15/3/2021, Published 30/3/2021

Abstract:

Tourism plays an important role in Malaysia's economic development as it can boost business opportunity in its surrounding economic. By apply data mining on tourism data for predicting the area of business opportunity is a good choice. Data mining is the process that takes data as input and produces outputs knowledge. Due to the population of travelling in Asia country has increased in these few years. Many entrepreneurs start their owns business but there are some problems such as wrongly invest in the business fields and bad services quality which affected their business income. The objective of this paper is to use data mining technology to meet the business needs and customer needs of tourism enterprises and find the most effective data mining technology. Besides that, this paper implementation of 4 data mining classification techniques was experimented for extracting important insights from the tourism data set. The aims were to find out the best performing algorithm among the compared on the results to improve the business opportunities in the fields related to tourism. The results of the 4 classifiers correctly classifier the attributes were JRIP (84.09%), Random Tree (83.66%), J48 (85.50%), and REP Tree (82.47%). All the results will be analyzed and discussed in this paper.

Key words: Classification, Data mining, J48, REP Tree, Tourism

Introduction:

Tourism is an important economic source for Malaysia, which was once ranked 9th in the world for tourist arrivals (1). Tourism has become Malaysia's third largest source of foreign exchange income (2). This means that there are many entrepreneurial opportunities and problems to be solved in the tourism industry in Malaysia.

This paper mainly based on business needs and customer needs solves the problem of investors investing by obtaining the income level of different tourist destinations and the problem of managers' judgment on the area of tourists to different destinations. Judgment of the area where the tourists belong is conducive to the manager of the destination to make relevant adjustments to attract more tourists and obtain the maximum benefit. For example, if the managers judge that the majority of tourists in the hotel belongs to the European people, then the menus, prompts and other places with text in hotel can add European languages, while adding some European customs and elements.

This paper needs to use data mining technology, which can not only reduce costs, but also use this technology to increase business opportunities (3). Data mining is the process of using data as input and generating output knowledge. For example, customer and tourist destination as the data input and provide output on recommending tourist destination. Business managers can use data mining technology to obtain the maximum benefit while reducing the cost of customer research, thereby prompting more people to start a business. This research conducted data mining on the simulated income data of various tourist destinations in Malaysia and the simulated tourist location data to determine the tourist area, so as to help real merchants in Malaysia use data mining to make correct judgments.

In the following part, this paper will study the key data mining task: use WEKA to implement 4 data mining classification techniques experiments, extract important information from the travel data set. The goal is to find the best performing algorithm in the comparison results to improve business opportunities in the travel-related fields. The correct classification results of the four

classifiers are: JRIP (84.09%), Random tree (76.62%), J48 (85.39%) and REP Tree (83.44%). This article will analyze and discuss all the results. Before end, this paper provides a list of data mining resources and tools for people who want to get more information on this topic.

Problem Statement

After investigation, this paper found that there are still enterprises that do not use big data mining in Malaysia's tourism industry today (4). Data mining has found typical patterns and influencing factors in the data, and it is difficult for managers to find these typical patterns and influencing factors (3).

From the perspective of business needs, the lack of big data applications, such as investors' lack of business income data form different tourist destinations, may lead them to make inaccurate investments in Malaysian tourist destinations. If the business managers of the tourist destinations lack information where the tourists come from, it is impossible to judge the source of tourists, so that the service quality provided for tourist cannot be improved in a targeted manner, resulting in the loss of passenger traffic. Therefore, from the analysis of these two aspects, the lack of data will affect the income of the industry.

From the perspective of customer needs, if the service level of a tourist destination fails to satisfy them, they may not recommend the tourist destination to friends in social media, resulting in a decrease in passenger flow at the tourist destination. From the analysis of these two aspects, the lack of data will affect the income of the industry.

Objective

In response to these two problems, this paper applies data mining technology to analyze and study the simulated business income data and the simulated data of the tourist's hometown, and obtain the highest or lowest income tourist destinations and the place where the tourists with the largest proportion of different tourist destinations belong.

From the perspective of business needs, it can help investors make accurate investments in different tourist destinations, and at the same time allow business managers to improve service levels in a targeted manner. At the same time, the most effective data mining classifier is obtained through experiments in this article, which is convenient for tourism enterprises.

From the perspective of customer needs, after the targeted improvement of the service quality of the tourist destination, they may recommend this tourist destination to friends around them to increase the passenger flow of the tourist destination. Therefore, the use of data mining technology can meet both business needs and customer needs.

Related Work

This section represents several related types of research on application data mining in tourism. All the related works were using different techniques in classification and the best method in getting the best result will be mentioned**.**

Algur et al. used the number of travelers from 2002 to 2013 to classify Historical Monument places. The location data set is preprocessed and allocated with different class labels such as low, medium, and high according to the number of visitors every year. There are several classification methods under a decision tree with 10 cross-validation folds is used such as Random Tree, REP Tree, Random Forest, and J47 algorithms. Those results showed Random Forest is the best among other classifiers by analyzing their performance metrics (5).

Irawan et al. selected a place that can be developed based on public and tourists to access tourist site which is more helpful to develop. Their experiment outcomes showed by using C4.5 shown that Nature Tourism object in Simalungun district can be developed in a level of recall of 83.33% and accuracy of 90%. C4.5 can provide better results on tourist location compare to other methods. Irawan et. al mentioned using C4.5 algorithm with 10 rules as a reference in the design and development of the application's GUI in classification for recommending tourist attraction which is a good method (6).

Srivihok et al. mentioned market segmentation is an important tool for dividing markets into smaller groups for comprised of individuals and they proposed a market segmentation method for travelers who visit Thailand for business. The technique is to evaluation unsupervised learning techniques such as SOM neural network, K-means and Hierarchical clustering by the number of the average Silhouette index and comparing the performance of supervised machine learning techniques such as J48, One R, Decision Table, MLP and Naïve Bayes. The classes of data (segments of tourist) used in supervised learning method are provided by the unsupervised learning method. The results indicated by Naïve Bayes performance are better compared to others to forecast the segments of new business tourists as part of the production from clustering method (7).

Urgessa et al. applied the information gain-based attribute selection method and construct a model for the compared algorithms after and before

selection of the attributes. Their research was framed by classification models which constructed using the after and before selected algorithms based on information gain to compare the performance of each situation. The methods selected by them are Decision Tree (J48, Random Forest, PART) and Support Vector Machine (SMO) (8). Their models were constructed on the tourism data to find out the noise-tolerant classification algorithm in the domain to recognize user behavior, improve the service, and business chances. The best performing algorithm is identified and the result showed Random Forest and SVM are more noise-tolerant as showed better performance (8).

Wang et al. mentioned that travel agencies cannot identify valuable travelers and tourist next destination. In their study which used the RFM model to describe valuable travelers. C4.5 decision tree was used to segment the valuable traveler for effectively proposing the promotion strategies for travel company by forecasting the destination and package tour cross-selling promotion to increase profits (9). Their research used Taiwanese travelers as mining samples with the applied decision tree to find valuable tourist, decision making behaviors, and demographics. The research is focused on using the mining process to segmenting valuable travelers and analyzing travel destination correlation to create a mining procedure for travel company to do better database marketing (9).

Methodology

This section is describing the detailed of the methodology applied in this paper. The steps of methodology are shown in Fig 1. This research model consists of several components which are Dataset and preprocess, Classification process, and result analysis and KDD process.

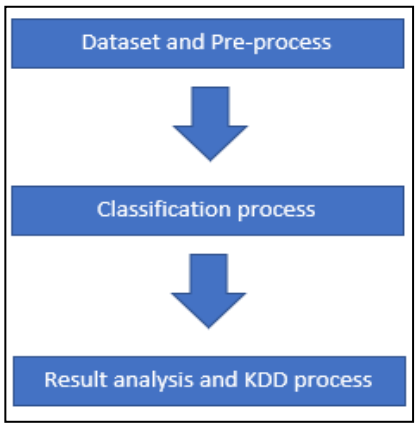


Figure 1. Research model

Data mining is a process that uses data as input and produces knowledge as output. The input is the tourism data set and the output is the rules, performance matrix, and the accuracy of the results of tourist's data. Data mining used algorithmic step in data mining process which known as Knowledge Discovery in Databases process (KDD) (10). Data mining required in the use of potentially large and diverse data set which may need for preprocessing to transformed into a representation suitable for data mining algorithm to remove missing and irrelevant data or attribute to tourism. The data mining software Waikato Environment for Knowledge Analysis (Weka) are using in this research as the tool of classification and analysis the results.

The dataset of tourist is collected from the year 2011 to 2012 from the online dataset. The data file is converted from excel (.csv) to Weka file (.arff). The dataset contains information of 8 types of business income in numeric in USD (art galleries, dance clubs, juice bars, restaurants, museums, resorts, parks or picnic spots, and beaches), periods, and 1 class is nominal which is the region of tourist (Africa, America, Asia, Europe, Oceania, and unstated). There are 924 instances in this dataset. The tourist's datasets are shown in Fig. 2.

Relation: tourism_dm2

No.	1: art_galleries	2: dance_clubs	3: juice_bars	4: restaurants	5: museums	6: resorts	7: parks/picnic_spots	8: beaches	9: period	10: class
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Date	Nominal
1	1.0	2.0	2.0	0.0	0.0	0.0	0.0	14.0	2011-..	Unstat..
2	1003.0	337.0	474.0	22.0	99.0	0.0	0.0	123.0	2011-..	Europe
3	1835.0	511.0	627.0	46.0	62.0	0.0	0.0	129.0	2011-..	Europe
4	517.0	144.0	137.0	26.0	10.0	0.0	0.0	56.0	2011-..	Europe
5	531.0	278.0	211.0	11.0	15.0	0.0	0.0	61.0	2011-..	Europe
6	228.0	114.0	108.0	5.0	4.0	0.0	0.0	19.0	2011-..	Europe

Figure 2. Tourist's Dataset

In Data mining, preprocessing is very important as it decides the quality of the result and exploit predictive data mining algorithms in knowledge discovery process (11). The effective preprocessing is needed to make the dataset be clean and consistent before used in the classification process. The tourist dataset does not contain any missing value in all the numeric attributes. The data are not going to convert into range by discretization as the concept of hierarchy in binary are only consist of the amount of income is more than an equal (income >= a) and less than (income < a) are going to be used in this research. The period is going to remove and class is set as a class attribute in Weka as the next process is classification. Figure 3 shows the concept of hierarchy.

Baghdad Science Journal
P-ISSN: 2078-8665
E-ISSN: 2411-7986

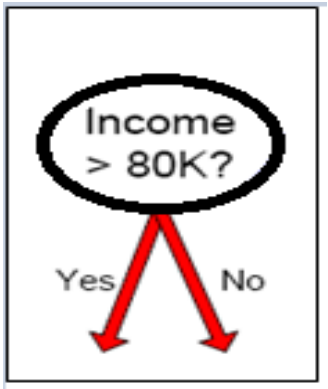


Figure 3. Concept of hierarchy

The classification process is learning a function that maps or classifies the data object into one of the predefined classes (12). For example, the tourist from going to juice bar spend more than 100 will class as Asian people in this research, there are 4 models of classification used which are JRIP, Random Tree, J48, and REP Tree. The function of each model is discussed in below section.

JRIP is a propositional rule learner that repeated incremental pruning for RIPPER. JRIP is constructed using WEKA and the classification rules. It will start with an empty set for the less prevalent to the more frequent value. JRIP consist of building and optimization stage. During the building stage, it will be repeated on grow (adding conditions) and prune (incrementally prune every rule) until the error rate >= 50% description length. Optimization is computing the original rule for a final representative of ruleset, if there are still residual positive. Then more rules are generated based on the residual positive and repeated in the build stage.

Random Tree randomly constructs decision trees. Random Tree is constructed using WEKA and the tree is represented by classification rules. Construction of each tree, algorithm picks a feature randomly at each node without any purity function check. If the categorical feature such as “Asia” has not been chosen before from the root of tree to the present node. It is useless to choose the repeat feature once more on the similar decision path as the pattern in the same path will have the same value but continuous feature such as “juice bar” can be picked more than once in the similar decision path. The tree stops growing if no more examples split in the current node or the depth of tree goes too deep.

J48 can be considered as C4.5 classification. J48 produces a classification-decision tree for the tourist dataset by recursive partitioning the tuples. J48 Tree classifier is constructed using WEKA and the built tree is represented by classification rules shown in Table 1. The depth-first strategy is used to build the decision tree. J48 considers all the possible tests to split the tourist dataset and selects the best information gain. The information gain of the binary partition point is based on distinct value and sub trees are built accordingly. This process is repeated for all attributes.

REP Tree Classification Models also is called fast decision tree learner. REP Tree is built using WEKA and the decision tree is represented by classification rules shown in Table 2. REP Tree builds a decision tree using prunes and information gained by reduced-error pruning. The REP Tree Classification sorts values for numeric attributes only one time.

The results of the 4 different models will be evaluated using performance evaluation metrics proved by Weka which are incorrectly classified instances, correctly classified instances, FP rate, TP rate, Precision, Recall and others. All the results will be compared for the knowledge discovery in the discussion section. Table 1 and 2 show the classification rules of the J48 and REP Tree classifiers.

Table 1. J48 Classification Rules

J48 pruned tree
----
beaches <= 571
| art_galleries <= 30
| | art_galleries <= 4: Unstated (12.0)
| | art_galleries > 4
| | | museums <= 5
| | | | beaches <= 15: America (10.0)
| | | | beaches > 15
| | | | | juice_bars <= 14: Oceania (8.0/2.0)
.
.
.
| | | | museums <= 40: Asia (2.0)
| | | | museums > 40: Europe (2.0)
| | | juice_bars > 469: Asia (224.0)
| | juice_bars > 6914
| | | restaurants <= 1089: America (24.0)
| | | restaurants > 1089: Asia (12.0)
Number of Leaves : 57
Size of the tree : 113

Table 2. REPTree Classification Rules

REPTree

```
============
beaches < 796
|  dance_clubs < 71.5
|  |  art_galleries < 30
|  |  |  art_galleries < 4.5 : Unstated (7/0) [5/0]
|  |  |  art_galleries >= 4.5 : Oceania (29/12) [9/3]
|  |  art_galleries >= 30
|  |  |  juice_bars < 26.5
.
.
.
|  |  juice_bars < 6544.5 : Asia (139/0) [78/0]
|  |  juice_bars >= 6544.5
|  |  |  dance_clubs < 42112 : America (12/0) [12/0]
|  |  |  dance_clubs >= 42112 : Asia (10/0) [4/0]
Size of the tree : 67
```

Discussion:

The data set used in this paper is about tourists from different regions visit Malaysia and income in USD of different places. This data set is built by 4 different methods in classification which are JRIP, Random Tree, J48, and Random Tree in WEKA with 10-fold cross-validation with the 924 tourist instances. The classification for 4 different types of classification is using "If…then" rule which is shown in Table 3 with its explanation. There are 20 rules applied in JRIP.

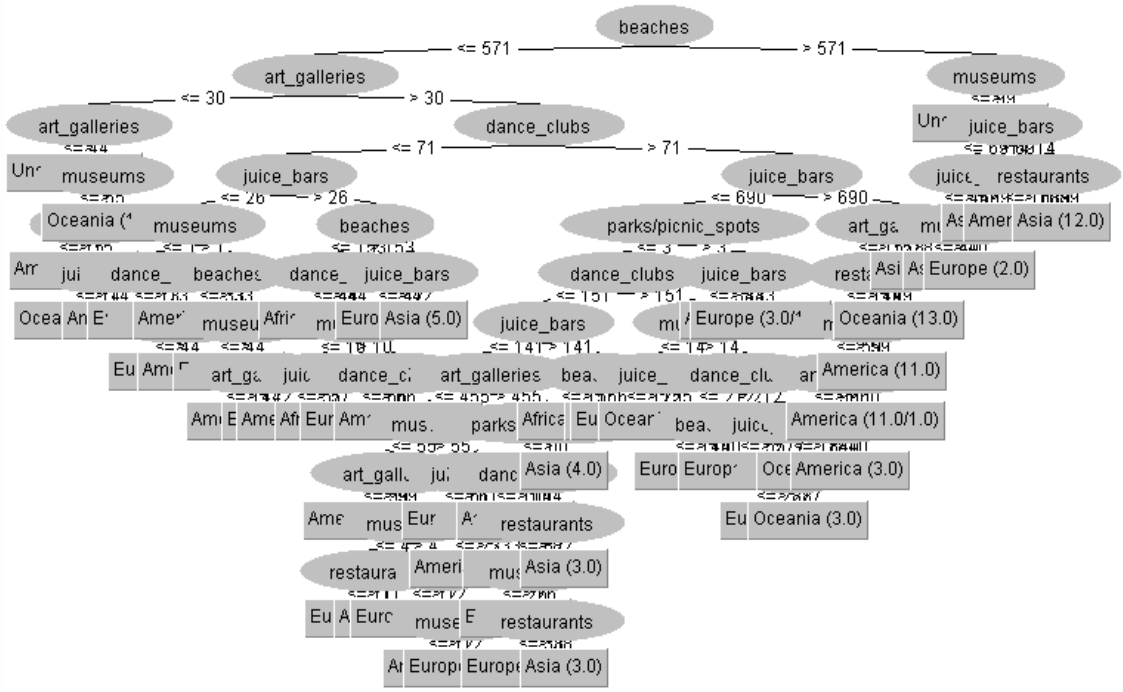
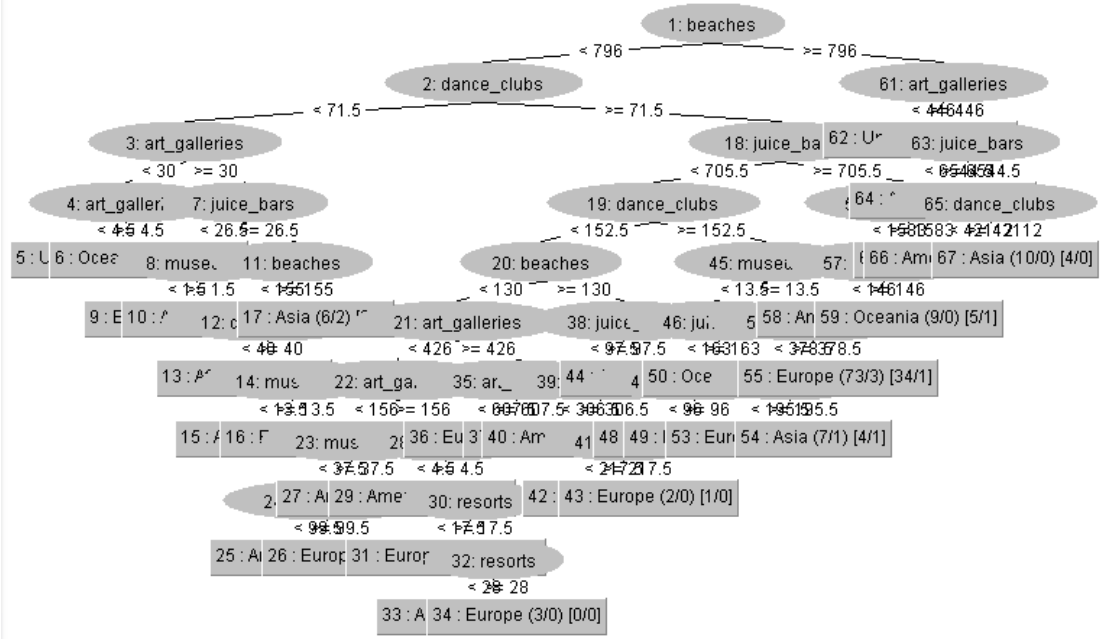
There are 227 of tree sizes for Random Tree. There are 57 leaves and 113 of tree for J48. There are 67 of tree for REP Tree. By comparing those rules generate by the 4 types of classifier J48 applied the most on the dataset so that the result will be more accurate compare to other 3 classifiers. Although Random Tree having more rule it split the dataset to deep.

Table 3. Explanation of rules with if...then

Classifier	Describtion
JRIP	**Decision Tree** None **Rule** (art_galleries <= 5) => class=Unstated (24.0/0.0) (juice_bars >= 27) and (dance_clubs <= 50) => class=Africa (53.0/15.0) **Explanation** • If art galleries equal of less than 5 USD, then is tourist come from region that unstated. There are 24/924 are unstated and 0 are wrongly classified • If juice bars equal of more than 27 USD and dance clubs equal or less than 50 USD, then is tourist come from Africa. There are 53/924 are Africa and 15 are wrongly classified
Random Tree	**Decision Tree** beaches < 585 | juice_bars < 25.5 | | art_galleries < 29 | | | art_galleries < 4.5 : Unstated (12/0) **Rule** (beaches < 585, juice bars < 25.5, art galleries < 29, art galleries < 4.5) => Unstated (12/0) **Explanation** • If beaches less than 585 USD, then if juice bars less than 25.5 USD, then if art galleries less than 29 USD, then if art galleries less than 4.5 USD, then is tourist from unstated. There are 12/924 are tourist from unstated and 0 are wrongly classified
J48	**Decision Tree** beaches <= 571 | art_galleries <= 30 | | art_galleries <= 4: Unstated (12.0) **Rule** (beaches<= 571, art galleries <= 30, art galleries < 4) => Unstated (12/0) **Explanation** • If beaches equal or less than 571 USD, then if art galleries equal or less than 30 USD, then if art galleries equal or less than 4 USD, then is tourist from unstated. There are 12/924 are tourist from unstated and 0 are wrongly classified
REP Tree	**Decision Tree** beaches < 796 | dance_clubs < 71.5 | | art_galleries < 30 | | | art_galleries < 4.5 : Unstated (7/0) [5/0] **Rule** (beaches < 796, dance clubs < 71.5, art galleries < 30, art galleries < 4.5) => Unstated (7/0) [5/0] **Explanation** • If beaches less than 796 USD, then if dance clubs less than 71.5 USD, then if art galleries less than 30 USD, then if art galleries less than 4.5 USD, then is tourist from unstated. For growing set there are 7/924 are tourist from unstated and 0 are wrongly classified. For pruning set there are 5/924 are tourist from unstated and 0 are wrongly classified.

The decision tree built from the tourist dataset by WEKA shown in Table 4 for J48, REP Tree, but JRIP does not have hierarchy tree as it is rule-based. All the decision tree is binary as it only contains 2 types of meaning which are more or equal to (>=) and less than (<). The decision tree can be explained by converting them into rules such as using the if…the rules in Table 4. The "oval" shape in Table 4 represents to the attributes and "square" shape in Table 4 represent as class. Table 4 shows value inside "square" shape or attributes of hierarchy tree represent to number of classified object and follow by number of incorrect classified object. The most complicated tree is Random Tree then follow by J48.

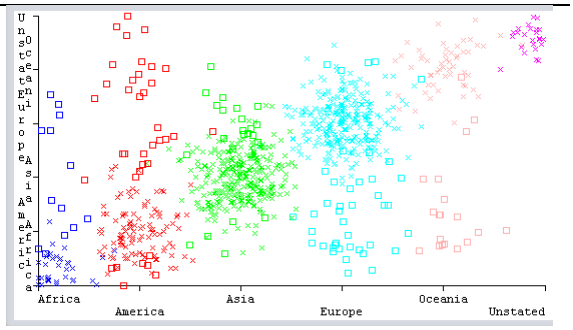
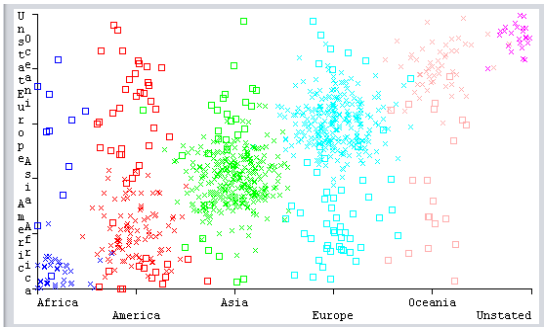
Table 4. Decision Tree

Classifier	Hierarchy Tree
JRIP	Rule base so there is no decision tree
J48	
REP Tree	

Besides that, WEKA also shows the error in the scatter plot the square is incorrect classified and x is correctly classified. There are 6 different colors used to represent to a different class. For the scatter

plot which can clearly visualize most of the American are wrongly classified and most Asian are correctly classified. Table 5 shows the classifier error generate by WEKA.

Table 5. Classifier Error

Classifier	Graph
J48	
REP Tree	

The results describe performance evaluation metrics on the correctly classified all the tourist's instances on their percentage of correctly classified, incorrectly classified, Kappa statistic, Mean absolute, TP rate, FP rate, Precision, Recall, F-Measure. All the measurement results are shown in Table 6 and 7. For JRIP the percentage of correctly classified is 84.09%, incorrectly classified is 15.91%, Kappa statistic is 0.78, Mean absolute is 0.066, TP rate is 0.841, FP rate is 0.058, Precision is 0.838, Recall is 0.841, F-Measure is 0.837. For Random Tree the percentage of correctly classified is 83.66%, incorrectly classified is 16.34%, Kappa statistic is 0.775, Mean absolute is 0.055, TP rate is 0.837, FP rate is 0.053, Precision is 0.834, Recall is 0.837, F-Measure is 0.835. For J48 the percentage of correctly classified is 85.50% (highest), incorrectly classified is 14.50% (lowest), Kappa statistic is 0.801 (highest), Mean absolute is 0.054 (lowest), TP rate is 0.855 (highest), FP rate is 0.045 (lowest), Precision is 0.857 (highest), Recall is 0.855 (highest), F-Measure is 0.856 (highest). For REPTree the percentage of correctly classified is 82.47% (lowest), incorrectly classified is 17.53% (highest), Kappa statistic is 0.761 (lowest), Mean absolute is 0.075 (highest), TP rate is 0.825 (lowest), FP rate is 0.051, Precision is 0.829 (lowest), Recall is 0.825 (lowest), F-Measure is 0.826 (lowest). The overall result after comparing the most accurate classifier is J48 and the worst is REPTree. The sequences of classifiers from the most to the less efficiency are J48, JRIP, Random Tree, REPTree. J48 is the best classifier applied for this tourist dataset.

Table 6. Results of 4 methods

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Kappa Statistic	Mean Absolute Error	Ranking among 4 methods
JRIP	84.09	15.91	0.780	0.066	2
Random Tree	83.66	16.34	0.775	0.055	3
J48	85.50	14.50	0.801	0.054	1
REP Tree	82.47	17.53	0.761	0.075	4

Table 7. Accuracy by class of 4 methods

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	Ranking among 4 methods
JRIP	0.841	0.058	0.838	0.841	0.837	2
Random Tree	0.837	0.053	0.834	0.837	0.835	3
J48	0.855	0.045	0.857	0.855	0.856	1
REP Tree	0.825	0.051	0.829	0.825	0.826	4

The TP (True Positive) rate and FP (False Positive) rate of the 4 classifiers in-depth as the results show in the form of confusion matrix with 6 x 6. Table 8 shown the Confusion Matrix of J48, and REP Tree. The confusion matrices that show "a" in the row and column is representing to the region of tourist come from is Africa, "b" in the row and column is representing to the region of tourist come from is America, "c" in the row and column is representing to the region of tourist come from is Asia, "d" in the row and column is representing to the region of tourist come from is Europe, "e" in the row and column is representing to the region of tourist come from is Oceania, and "f" in the row and column is representing to the region of tourist come from is unstated. The green dotted line in the confusion matrix in Table 8 represents the correct classified instances.

There are total 324 instances are originally classified as "c" by using J48 there are 301 instances correctly classified and 23 instances are incorrectly classified. The 23 instances should classify in class "c" but there are 3 incorrectly classified in "a", 4 incorrectly classified in "b", 16 incorrectly classified in "d", and 2 incorrectly classified in "e". There are total 312 instances are originally classified as "d" by using REP Tree there are 259 instances correctly classified and 53 instances are incorrectly classified. The 53 instances should be classified in class "d" but there are 6 incorrectly classified in "a", 26 incorrectly classified in "b", 13 incorrectly classified in "c", 6 incorrectly classified in "e", and 2 incorrectly classified in "f".

Table 8. Confusion Matrix

Classifier	Matrix (6x6)
J48	=== Confusion Matrix === a b c d e f <-- classified as 35 6 1 6 0 0 | a = Africa 7 107 6 15 9 0 | b = America 1 4 301 16 2 0 | c = Asia 2 24 11 270 5 0 | d = Europe 0 16 0 3 58 0 | e = Oceania 0 0 0 0 0 24 | f = Unstated
REP Tree	=== Confusion Matrix === a b c d e f <-- classified as 37 3 0 6 2 0 | a = Africa 15 95 4 19 11 0 | b = America 3 4 295 19 2 1 | c = Asia 6 26 13 259 6 2 | d = Europe 1 11 0 8 52 0 | e = Oceania 0 0 0 0 0 24 | f = Unstated

Conclusion:

This paper conducted experiments on the use of WEKA to implement 4 data mining and classification technologies on data from the Malaysian tourism industry, including JRIP (84.09%), Random tree (76.62%), J48 (85.39%) and REP Tree (83.44%). Extract important information from the data set about the income data of tourist destinations and the places where tourists belong, this paper finds the best performing algorithm in the comparison results to improve business opportunities in the travel-related fields. It provides information for investors to make accurate investments in different tourist destinations, and also helps managers to accurately judge the region of tourists come from in different tourist destinations. The most effective method is to use the J48 classifier for analysis, and the least effective is to use REP Tree for data mining analysis.

However, it should be noted that the performance of the data mining process directly depends on the number of available cases (instances) that can be used. Its use does not guarantee the best business results, but it can greatly reduce the risk of making wrong decisions. The results show that no one optimal algorithm can beat other algorithms in all cases (3).

Finally, this paper also provides a list of data mining resources and tools for those who wish to obtain more information on this topic.

Authors' declaration:

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are mine ours. Besides, the Figures and images, which are not mine ours, have been given the permission for re-publication attached with the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee in University Sains Malaysia.

References:

1.Yonya International Company. Malaysia is ninth most visited in the world in UNWTO list[Internet]. 2014 September 7 [cited 2020 Jun 30]. Available from http://travel-to-malaysia.com/malaysia-is-ninth-most-visited-in-the-world-in-unwto-list/.

2.WORDISK. Tourism in Malayisa [Internet]. 2020. [cited 2020 Jun 30] . Available from https://worddisk.com/wiki/Tourism_in_Malaysia/#cite_note-2.

3.Jovanović V. THE SELECTION OF OPTIMAL DATA MINING METHOD FOR SMALL-SIZED HOTELS. In Synthesis 2015-International Scientific Conference of IT and Business-Related Research. Singidunum University; 2015. p. 519-524.

4.Hirschmann R. Travel and tourism in Malaysia – Statistics & Facts [Internet]. 2020 January 7 [cited 2020 Jun 30] . Available from https://www.statista.com/topics /5741/travel-and-tourism-in-Malaysia/.

5.Algur SP, Bhat P, Hiremath PS. Application of data mining in the classification of historical monument places. INT J INTELL SYST APP. 2016 Aug 1;8(8):58.

6.Irawan E, Gunawan I, Tambunan HS, Qurniawan H. Application of Classification Algorithm C4. 5 in Recommendations for Natural Tourism Development in District Simalungun. J PHYS CONF SER. IOP Publishing. 2019. Aug;1255(1): p. 012078.

7.Srivihok A, Yotsawat W. Market segmentation of inbound business tourists to Thailand by binding of unsupervised and supervised learning techniques. J SOFTW. 2014 May 1;9(5).

8.Urgessa T, Maeng W, Lee JS. Application of Data Mining Techniques for Tourism Knowledge Discovery. INT J COMPUT INF SCI ENG. 2017 Feb 3;11(1):119-124.

9.Wong JY, Chen HJ, Chung PH, Kao NC. Identifying valuable travelers and their next foreign destination by the application of data mining techniques. ASIA PAC J TOUR RES. 2006 Dec 1;11(4):355-373.

10.Weiss GM, Davison BD. Data mining. In TO APPEAR IN THE HANDBOOK OF TECHNOLOGY MANAGEMENT, H. BIDGOLI (ED.); 2010.

11. Alexandropoulos SA, Kotsiantis SB, Vrahatis MN. Data preprocessing in predictive data mining. KNOWL ENG REV. 2019;34.

12.Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to discovery knowledge in databases. AI MAG. 1996;3(17): p. 37-54.

تطبيق تقنيات التنقيب عن البيانات على النفقات السياحية في ماليزيا

كاي مياو[1] * **تان شي آن[2]**

[1]جامعة سينز ماليزيا ، الصين.
[2]جامعة سينز ماليزيا ، ماليزيا.

الخلاصة:

تلعب السياحة دورًا مهمًا في التنمية الاقتصادية لماليزيا حيث يمكنها تعزيز فرص العمل في الاقتصاد المحيط بها. من خلال تطبيق استخراج البيانات على بيانات السياحة للتنبؤ بمجال الفرص التجارية وهذا يعد اختيارًا جيدًا. استخراج البيانات هو العملية التي تأخذ البيانات كمدخلات وتنتج معرفة المخرجات. بسبب ازدياد عدد السكان الذين يسافرون في بلد آسيا في هذه السنوات القليلة. يبدأ العديد من رواد الأعمال أعمالهم الخاصة ولكن هناك بعض المشاكل مثل الاستثمار الخاطئ في مجالات الأعمال الخدمات السيئة التي أثرت على دخل أعمالهم. الهدف من هذه البحث هو استخدام تقنية استخراج البيانات لتلبية احتياجات العمل واحتياجات العملاء للمؤسسات السياحية والعثور على تكنولوجيا استخراج البيانات الأكثر فعالية. بالإضافة إلى ذلك ، تم تجربة تنفيذ هذا البحث لأربع تقنيات تصنيف استخراج البيانات لاستخراج رؤى مهمة من مجموعة البيانات السياحية. كانت الأهداف هي معرفة أفضل الخوارزمية أداءً من بين النتائج المقارنة لتحسين فرص العمل في المجالات المتعلقة بالسياحة. كانت نتائج المصنفات الأربعة الصحيحة هي(84.09 ٪ JRIP (، 83.66) Random Tree٪ (، 85.50) J48٪ (، وشجرة82.47) REP ٪ .(سيتم تحليل جميع النتائج ومناقشتها في البحث.

الكلمات المفتاحية: التصنيف، التنقيب في البيانات، J48 ، شجرة REP ، السياحة