Prediction Models Based on Flight Tickets and Hotel Rooms Data Sales for Recommendation System in Online Travel Agent Business

Handito Muhammad Septiadi[1], Citrananda Ariandika[2], Andry Alamsyah[3]
[1,2,3]Faculty of Economics and Business
Telkom University
Bandung, Indonesia
[1] handito.mseptiadi@hotmail.com, [2]ariandika.citra@gmail.com, [3]andrya@telkomuniversity.ac.id

Abstract

Indonesia as one of the favorite vacation destinations of domestic and foreign travelers made the value of investment in the tourism industry continued to grow significantly. This was created more Online Travel Agent business in recent years. However, it made a lot of business travel and Umrah travel in Indonesia is threatened with bankruptcy, after the online travel business activity is rampant in enterprise business market ticket sales and travel tours. The case study of this research was different from the Online Travel Agent business in general because it worked in real-time using flight tickets and hotel rooms sales data. Data mining, extraction of hidden predictive information from large databases, was a powerful technology with great potential to help companies focus on the most important information in their data warehouse. By using the method of classification in data mining, the objectives of this paper was able to create predictive models from flight tickets and hotel rooms sales data using the decision tree classification approach. The result of this paper is beneficial for business that can be used as basic algorithm for programming in Online Travel Agent recommendation feature.

Keywords: Data Mining; Decision Tree; Classification; CHAID; C5.0.

1. Introduction

Over the years, the increase of interest in travelling by people of Indonesia has been shown by the increasing number of passengers and cargo departures in airports of Indonesia (Badan Pusat Statistik, 2013). The entry of low-cost carrier into Indonesia early 2000s might be one of many factors that causing this phenomenon. Correlated with travel, hotel occupancy rate by local travelers has also gone up as well (Badan Pusat Statistik, 2014). The Ministry of Tourism Indonesia recorded that the development of the Travel Agency in 2007-2011 in Indonesia continued to increase from 655 to 1.120 businesses (Kementerian Pariwisata, 2012). Considering these increasing numbers, local online intermediary industry player such as Traveloka, PegiPegi and WeGo are competing against each other to get most local and international travelers.

Among those intermediary industry players, they provide a common feature, which is customers must input their departure location, destination, airline, hotel, and other aspects manually without any recommendation. If the customer wants to get the cheapest airline and route available, he/she must manually search through many dates and routes available. This old workflow can consume a long time for the customers.

This research aims to reduce time consumed for searching cheapest flights or hotel rooms by changing the conventional workflow. The new process requires the customers to input their desired budget, and the algorithm will display lower or the same flights and hotel room rates. The benefit of this new workflow is the customers can easily manage their travelling budget. To support this new workflow, we created a predictive model that can be used for flight and hotel room recommendation algorithm and a competitive advantage as a feature for competing in intermediary industry. This feature is for those who want to travel but still don't know where to go with their budget. This feature will adjust customers' destination to become suitable to customers' budget preferences.

2. Theoretical Background

Methodology for analysing flight tickets and hotel rooms sales data is data mining classification. Data mining is the process of finding interesting patterns and knowledge from large amounts of data. Data sources include databases, data warehouses, web, other information repositories, or data that flows into a dynamic system (Han et al., 2012). Classification technique is controlled learning technique that classifies the data items into a predetermined class label. This is one of the most useful technique in data mining to build a classification model from a set of input data (Al-Radaideh and Al Nagi, 2012). Classification is used to estimate by using categorical rather than numeric variables (Larose, 2014).

We used decision tree because it is a self-explanatory model, can process numeric, nominals, and also discrete values (Rokach and Maimon, 2015). With its ability to breakdown the complex decision-making process becomes much simpler. Decision trees can also find hidden relationships between a number of potential input variables with a target variable. In addition a decision tree can combine data exploration and modelling, so it is good as a first step in the modelling process (Azmi and Dahria, 2013).

Chi-squared Automatic Interaction Detection (CHAID) and C5.0 are classification algorithms. CHAID finds the pair of values in V_i that is least significantly different with respect to the target attribute. The significant difference is measured by the p value obtained from a statistical test. The statistical test used depends on the type of target attribute. An F test used if the target attribute is continuous, a Pearson chi-squared test if it is nominal, and a likelihood ratio test if it is ordinal (Rokach and Maimon, 2015a). There are several steps to make decision tree using CHAID algorithm (H. P. and Alamsyah, 2015):

1. Merging
   Category merging can be done on independent variable that has more than two category that are related.
2. Splitting
   In this part independent variable which used as the best split node. Splitting conducted with p-value on each independent variable.
3. Stopping
   Decision tree should be terminated by the rules. If there is no a significant independent variable or if a tree reaches a maximum value limit of the tree defined specifications.

C4.5 was replaced in 1997 by a commercial version of the system C5.0 Rulequest Research, Inc. C5.0 greatly improve the scalability of decision trees (Kantardzic, 2011). C5.0 claimed is more efficient than C4.5 in terms of memory and computation time. In certain cases, C5.0 presents the acceleration of an hour and a half (which is required for the algorithm C4.5) to only 3.5 seconds. Besides supporting the boosting procedure that can improve predictive performance (Rokach and Maimon, 2015b).

3. Methodology and Experimentation

This research uses flight tickets and hotel rooms data sales as our primary data source. The ticket sales data consists of: date of transaction, airline, route, price, payment methods, and gender of the customer. While hotel sales data consists of: guest gender, month of check in, day of check in, hotel names, hotel-star, amount of rooms and days, room rates per room and total, and payment method. All variables data type we convert into discrete. The ticket sales data contains 8853 data, and the hotel sales data contains 4912 data.

	Guest	MonthCheckIn	DayIn	Hotel	Stars	Qty	Days	Price	Total	Payment	MonthRsv
2	M	Dec	Minggu	PROMENADE	3	<=1.50	3	(1599000.00, 2080500.00]	(2078000.00, 2851000.00]	credit	JANUARI
3	M	Dec	Rabu	PROMENADE	3	(1.50, 2.50]	1	(639750.00, 799500.00]	<=531500.00	credit	JANUARI
4	M	Dec	Jumat	HORISON BANDUNG	4	(1.50, 2.50]	2	(531500.00, 639750.00]	(2078000.00, 2851000.00]	cash	JANUARI
5	M	Dec	Rabu	TRINITI	3	<=1.50	1	(1263000.00, 1599000.00]	>4556500.00	transfer	JANUARI
6	M	Dec	Sabtu	IBIS TRANS STUDIO	3	<=1.50	1	(1018750.00, 1263000.00]	(531500.00, 639750.00]	credit	JANUARI
7	F	Dec	Sabtu	HARRIS BANDUNG	4	<=1.50	1	(799500.00, 1018750.00]	<=531500.00	credit	JANUARI
8	M	Dec	Senin	GRAND PREANGER	5	<=1.50	2	(1599000.00, 2080500.00]	(1263000.00, 1599000.00]	credit	JANUARI
9	M	Dec	Sabtu	V HOTEL	3	<=1.50	3	(1263000.00, 1599000.00]	(1263000.00, 1599000.00]	credit	JANUARI
10	M	Dec	Kamis	SURYA INDAH	2	(1.50, 2.50]	1	(531500.00, 639750.00]	(531500.00, 639750.00]	transfer	JANUARI
11	M	Dec	Kamis	THE AKMANI	4	(3.50, 4.50]	3	(799500.00, 1018750.00]	>4556500.00	credit	FEBRUARI
12	F	Dec	Jumat	ASTON BRAGA HOTEL	4	(1.50, 2.50]	2	(1018750.00, 1263000.00]	>4556500.00	cash	JANUARI
13	M	Dec	Jumat	SANTIKA BANDUNG HOTEL	3	<=1.50	1	(799500.00, 1018750.00]	(799500.00, 1018750.00]	credit	JANUARI
14	M	Dec	Jumat	PROMENADE	3	<=1.50	1	(1599000.00, 2080500.00]	<=531500.00	transfer	JANUARI
15	M	Dec	Sabtu	TRINITI	3	(1.50, 2.50]	2	(1018750.00, 1263000.00]	(1599000.00, 2078000.00]	transfer	JANUARI
16	M	Dec	Sabtu	GOLDEN FLOWER	4	(1.50, 2.50]	1	(1018750.00, 1263000.00]	(2078000.00, 2851000.00]	credit	JANUARI
17	F	Dec	Senin	JAYAKARTA	4	<=1.50	4	>4556500.00	(2851000.00, 4556500.00]	credit	JANUARI
18	M	Dec	Selasa	The Travelhotel Cipaganti	3	<=1.50	1	(2080500.00, 2851000.00]	(799500.00, 1018750.00]	transfer	JANUARI
19	F	Dec	Selasa	HYPER INN	3	<=1.50	2	(531500.00, 639750.00]	(2078000.00, 2851000.00]	credit	JANUARI
20	M	Dec	Selasa	The Travelhotel Cipaganti	3	<=1.50	1	(1263000.00, 1599000.00]	(531500.00, 639750.00]	credit	JANUARI
21	M	Dec	Selasa	The Travelhotel Cipaganti	3	<=1.50	1	(799500.00, 1018750.00]	(639750.00, 799500.00]	credit	JANUARI

Figure. 1. The Result of Preprocess Hotel Rooms Sales Data

Issued	Monthly	Payment	Airlines	Departure	Destination	Departure City	Destination City	Route	Price Class	Price	Price Intervals	Gender
2/1/2014	January	cash	JT	DPS	BDO	Denpasar	Bandung	DPS BDO	1062850 - 1284600	1204000	Price 09	M
2/1/2014	January	cash	JT	DPS	BDO	Denpasar	Bandung	DPS BDO	1062850 - 1284600	1204000	Price 09	F
2/1/2014	January	cash	XN	PDG	CGK	Padang	Jakarta	PDG CGK	749150 - 819850	750000	Price 06	F
2/1/2014	January	cash	XN	PDG	CGK	Padang	Jakarta	PDG CGK	749150 - 819850	750000	Price 06	F
2/1/2014	January	cash	XN	PDG	CGK	Padang	Jakarta	PDG CGK	749150 - 819850	750000	Price 06	F
2/1/2014	January	cash	XN	PDG	CGK	Padang	Jakarta	PDG CGK	749150 - 819850	750000	Price 06	M
2/1/2014	January	cash	XN	PDG	CGK	Padang	Jakarta	PDG CGK	749150 - 819850	750000	Price 06	F
2/1/2014	January	cash	XN	PDG	CGK	Padang	Jakarta	PDG CGK	749150 - 819850	750000	Price 06	F
2/1/2014	January	cash	XN	BDO	PDG	Bandung	Padang	BDO PDG	1062850 - 1284600	1250000	Price 09	F
2/1/2014	January	cash	JT	CGK	PNK	Jakarta	Pontianak	CGK PNK	532750 - 609450	555000	Price 03	M
2/1/2014	January	cash	JT	CGK	PNK	Jakarta	Pontianak	CGK PNK	532750 - 609450	555000	Price 03	F
2/1/2014	January	cash	JT	BDO	BTH	Bandung	Batam	BDO BTH	672850 - 749150	698000	Price 05	F
2/1/2014	January	cash	JT	BDO	DPS	Bandung	Denpasar	BDO DPS	< 455450	379000	Price 01	M
2/1/2014	January	cash	JT	BDO	DPS	Bandung	Denpasar	BDO DPS	< 455450	379000	Price 01	M
2/1/2014	January	cash	JT	BDO	DPS	Bandung	Denpasar	BDO DPS	< 455450	379000	Price 01	M
2/1/2014	January	cash	SJ	CGK	TNJ	Jakarta	Tanjung Pinang	CGK TNJ	455450 - 532750	522000	Price 02	F
2/1/2014	January	cash	SJ	CGK	TNJ	Jakarta	Tanjung Pinang	CGK TNJ	< 455450	56700	Price 01	M
2/1/2014	January	cash	JT	CGK	PNK	Jakarta	Pontianak	CGK PNK	609450 - 672850	610000	Price 04	M
2/1/2014	January	cash	JT	CGK	PDG	Jakarta	Padang	CGK PDG	455450 - 532750	472500	Price 02	M
2/1/2014	January	cash	JT	PDG	CGK	Padang	Jakarta	PDG CGK	455450 - 532750	527500	Price 02	M

Figure. 2. The Result of Preprocess Flight Sales Data

Our research workflow is as shown on Figure 3 (Marbán et al., 2009). The first process is business/research understanding phase; we aim to reduce time consumed for searching cheapest flights or hotel rooms, and for that we need a budget-based recommendation algorithm. The second step is data understanding phase; we gather all of our data and evaluate the quality of the data. The third step is data preparation phase; we choose which variable we want to analyze, and also preprocessing data. Fourth step is modelling phase; we implement which data mining method that's suitable for our objective, which is classification. Fifth step is evaluation phase; we evaluate the model that have been produced, is the model have reached the objective of our research or not. And the last step is deployment phase. These steps are reversible, when the researchers want to make some adjustments in earlier phase (changes in research aim) he/she can do so. The data of this research is shown in Table 1.

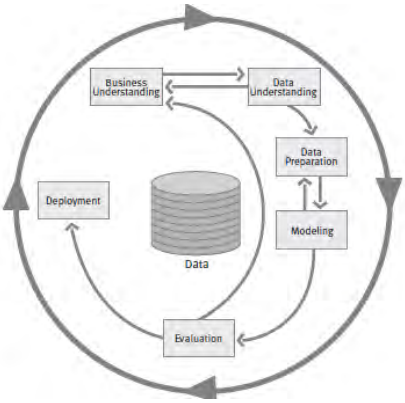

Fig. 3. CRISP-DM Process

The preprocessing step is done by deleting missing values and deleting round trip input for flight tickets data. But the hotel rooms data has none missing values. After the preprocessing phase, we continue to modelling phase which is process the data with data mining applications. As mentioned before, we use classification method for the prediction models. The flight tickets data using SPSS application with CHAID algorithm because the attributes are almost discrete and CHAID is known to process discrete attributes as target variable. Whereas the hotel rooms data are processed using Orange Canvas application with C5.0 algorithm concept.

Table 1. The Amounts of Data

	Airplane Ticket	Hotel
Number of Raw Data	8853	4912
After Preprocessing	7834	4912

The flight tickets model validation method, we used the default SPSS' method which is Split Sample; a method which splits 75% of the data into training dataset, and the rest 25% goes into test dataset. In order to declare the validation of prediction model, both of the models (training and test) must be identical in their model's roots. While the hotel rooms data we use

classification tree widget scheme in Orange Canvas.

4. Result and Analysis

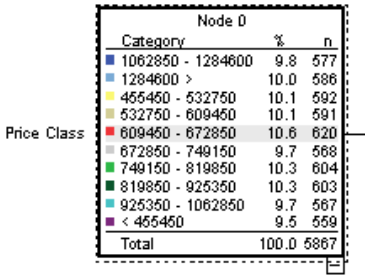

Fig. 4. Price Ranges of the Airplane Ticket Data

Airplane ticket data consists of four variables: Month of book, routes, price range and airlines. As seen as Figure 4, for the training dataset the algorithm uses 5867 data in total, complete with the percentages of each price range. According to the result in Figure 5, the most influential variable to price range is the month of book, then route. The result is so much bigger than appeared in between figure 4 through 6 and we can't include all of them in this paper. As shown in Figure 5, the first root after the first node of the decision tree is "Monthly" which the month of ticket booking is. As seen as Figure 5 which is July month of book, from the majority of flights in July are above Rp1.284.600 with the percentage of 24.4%, the same explanation is also applied for the rest of the price categories.

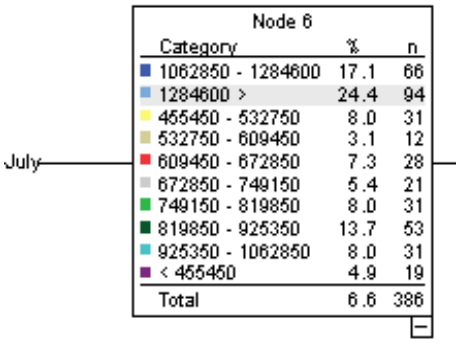

Fig. 5. July Month of Book

The second root is "Route" which is self-explanatory. Here in Figure 6 we can see 2 of 5 Route nodes, the interpretation is if the customers looking for flights in July with price range of under Rp1.062.850, the routes will be available is CGK PNK, BDO JOG and so on. Though there is 10.5% possibility for the price if the customer But for month April and May the second root is "Airline", and the third root is "Route". This happens because on April and May in the airline variable has much more varieties than other months in 2014.

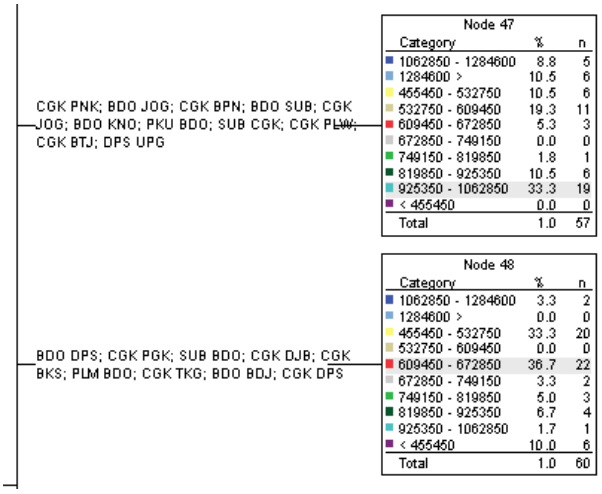

Fig. 6. April and May Month of Book

From preprocessing hotel rooms sales data in Figure 1, we choose four variables/attributes to construct prediction models: hotel names, day stay, the month of stay, and room rates. According to the result in Figure 7 and Figure 8, the decision trees are too long so we shrink them in order to fit in this paper.

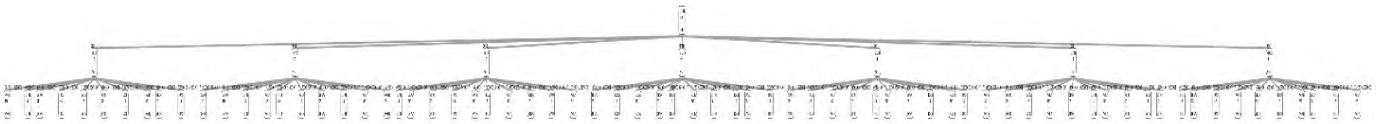

Fig. 7. Prediction Model Based pn Choosing Four-Star Hotel in Bandung, Day Stay and Hotel Room Rates

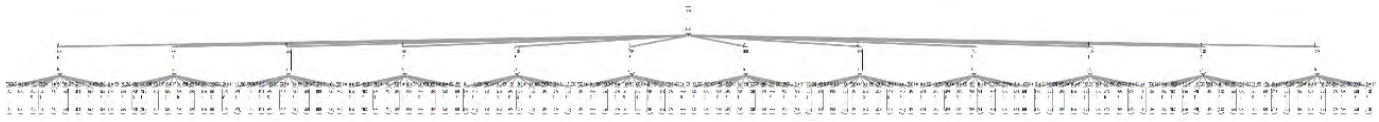

Fig. 8. Prediction Model Based on Choosing Three-Star Hotel in Bandung, Hotel Room Rates, and Month Stay

Model in Figure 7 generates rules which facilitate us to read the decision tree easily. If we zoom out Figure 7, it will show as Figure 9 that describes when a customer wants to stay at four-star hotels in Bandung on Sunday (Minggu) at a budget cost per room for Rp1.018.750 - Rp1.263.000, customers will be recommended to De Java Hotel.

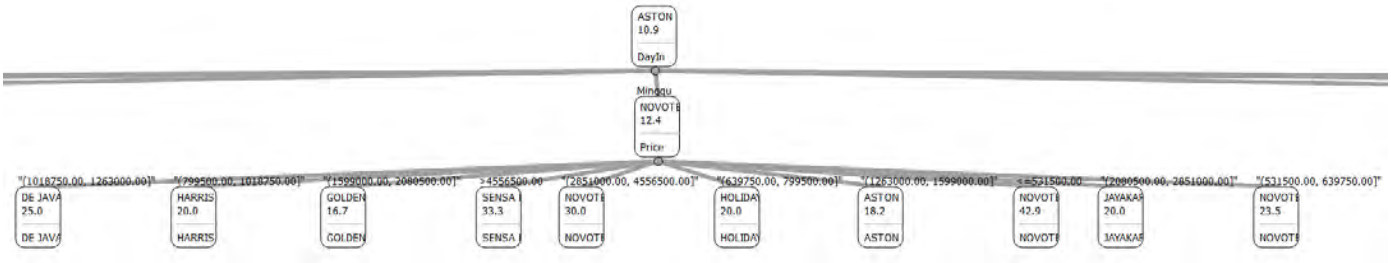

Fig. 9. The detail of Decision Tree in Figure 5

It can be seen on selected hotel in the rules of Figure 8 predictive model and also decision tree in Figure 10 that describe if a customer wants to stay in November at a budget cost per room for Rp2.851.000 - Rp4.556.500 at three-star hotels in Bandung, the customer will be recommended to the Hotel Panghegar. In addition, at a budget cost per room which is under Rp513.500 and about Rp1.263.000 - Rp1.599.000 customers is also recommended to the Hotel Panghegar.

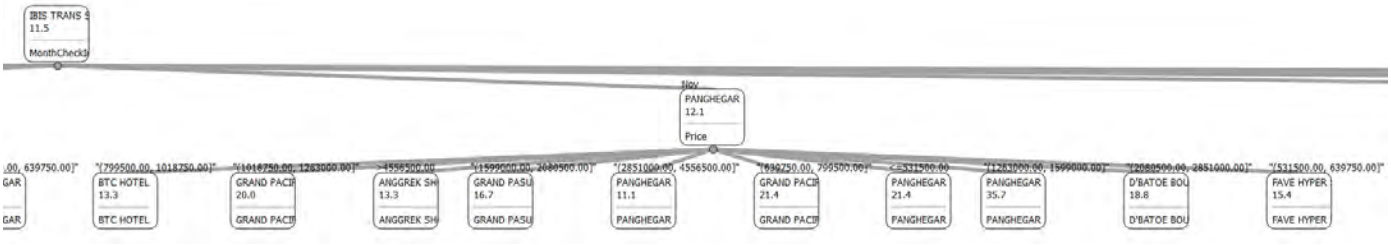

Fig. 10. The detail of Decision Tree in Figure 5

5. Conclusion

Classification methods are used to provide a recommendation for selecting airline and hotel based on customer's budget and others customer needs in the search engine features of Online Travel Agent. The feature is expected to be a business excellence. If the sales data has many variables/attributes of the data, the more variation of predictive models that can be generated. Recommendations can be seen from each of the rules generated by the predictive models.

In the future, we suggest to perform processing flight tickets and hotel rooms data by using others data mining techniques and algorithms for problem solving. And also recommend Online Travel Agent to provide organize and complete sales data in their record to facilitate the processing of data using data mining techniques and gain information.

References

Al-Radaideh, Q. A., & Al Nagi, E. (2012). 'Using Data Mining Techniques to Build A Classification Model for Predicting Employee Performance.' *International Journal of Advanced Computer Science and Applications (IJACSA),* Vol. 3 No. 2, pp.144-151.

Azmi, Z., & Dahria, M. (2013). 'Decision Tree Berbasis Algoritma untuk Pengambilan Keputusan.' *Jurnal SAINTIKOM,* Vol. 12 No. 3, pp.157-164.

Badan Pusat Statistik. (2013). *Jumlah Keberangkatan Penumpang dan Barang di Bandara Indonesia Tahun 1999 - 2013* [online]. http://www.bps.go.id/linkTabelStatis/view/id/1404 (Accessed 26 July 2015)

Badan Pusat Statistik (2014). *Jumlah Tamu Indonesia pada Hotel Bintang Menurut Provinsi Tahun 2003 – 2014* [Online]. Available: https://www.bps.go.id/linkTabelStatis/view/id/1377 (Accessed 16 March 2016)

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Massachusetts, USA: Morgan Kaufmann.

H. P., Raden Johannes, & Alamsyah, Andry. (2015). 'Sales Prediction Model Using Classification Decision Tree Approach for Small Medium Enterprise Based on Indonesia E-Commerce Data.' *Seminar & Conference on Business & Technology in ICT Industry.* [online] https://www.academia.edu/16500888/Sales_Prediction_Model_Using_Classification_Decision_Tree_Approach_For_Small_Medium_Enterprise_Based_on_Indonesian_E_Commerce_Data

Kantardzic, M. 2011. *Data Mining Concepts, Models, Methods, and Algorithms.* New Jersey: John Wiley & Sons, Inc.

Kementerian Pariwisata. (2012). Perkembangan BPW (Biro Perjalanan Wisata) Berskala Menengah dan Besar Menurut Provinsi Tahun 2007-2011. Retrieved September 21, 2015, from Statistik Jasa Perjalanan: http://www.kemenpar.go.id/userfiles/file/BPW%20TK%202007%20-%202011.pdf

Larose, Daniel T., dan Larose, Chantal D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining Second Edition*. New Jersey: John Wiley & Sons Inc.

Marbán, Óscar, Mariscal, Gonzalo, dan Segovia, Javier. (2009). A Data Mining & Knowledge Discovery Process Model. *Data Mining and Knowledge Discovery in Real Life Applications, pp. 438,* Ponce, Julio., dan Karahoca, Adem. Vienna: I-Tech Education and Publishing. Available at: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/a_data_mining__amp__knowledge_discovery_process_model (Accessed 30 Agustus 2015)

Rokach, Lior., dan Maimon, Oded. (2015). *Data Mining with Decision Trees: Theory and Applications 2nd Edition.* Singapore: World Scientific Publishing.