ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

PREDICT AND ASSOCIATE TOURIST PREFERENCE PATTERNS, THE CASE OF MINISTRY OF CULTURE AND TOURISM OF ETHIOPIA

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Information Science

BY

DAWIT HAILEMICHAEL HAILU

June 2015
Addis Ababa

ADDIS ABABA UNIVERSITY
COLLEGE OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

PREDICT AND ASSOCIATE TOURIST PREFERENCE PATTERNS, THE CASE OF MINISTRY OF CULTURE AND TOURISM OF ETHIOPIA

BY

DAWIT HAILEMICHAEL HAILU

ADVISOR

SOLOMON TEFERRA ABATE (PhD)

June 2015
Addis Ababa

DEDICATION

This paper dedicated to my Mother: **DILLAY DESALEGN**

DECLARATION

I declare that this thesis is my original work and has not presented for a degree in any other university. Wherever contributions of others are involved, every effort is made to indicate this clearly with due reference to the literature, and acknowledgement of collaborative research.

_____

Dawit Hailemichael Hailu

June 2015

This thesis has been submit for examination with my approval as university advisors.

SOLOMON TEFERRA ABATE (PhD) _____ _____

Signature Date

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Dr. Solomon Teferra Abate for his devotion, excellent guidance, and constant support during my work. This research work would not materialized without his time wise and constructive comments. I would like to thanks Mr. Bahredin Mensur head of policy planning, monitoring and evaluation directorate in the ministry of culture and tourism of Ethiopia; for bringing a valuable data for this study. I would also want to extend my gratitude to Mr. Dawit Hailu, director of catering and tourism training institute (CTTI) who has given me the opportunity to attend the postgraduate program. I am equally grateful to all who helped me in completing this study.

Table of Contents

LIST OF TABLES AND FIGURES

ABBREVIATIONS

Age_group_1	Age equals to 25 and above
Age_group_2	Age between 26=<and>=45
Age_group_3	Age greater than and equals to 46
DM	Data Mining
J48	An algorithm which is the successor of ID3(Iterative Dichotomiser) C4.5
KDD	Knowledge Discovery in Databases
KDP	Knowledge Discovery Process
LoS_ Ex	Extended Length of Stay
LoS_Med	Medium Length of Stay
LoS_Short	Short Length of Stay
LoS_Very_Ex	Very extended Length of Stay
MoCT	Ministry of Culture and Tourism of Ethiopia
CRISP-DM	Cross Industry Standard Process for Data Mining
UNESCO	United Nations Educational, Scientific and Cultural Organization
UNCTAD	United nations Conference on trade and development
UNWTO	United Nations World Tourism Organization
WEKA	Waikato Environment for Knowledge Analysis

ABSTRACT

Tourism is one of the largest and rapidly growing industries in Ethiopia. It has a vital influence on economic development of a country. The degree of tourist attraction of Ethiopia to the others east African countries is, not strong enough to penetrate the tourism market. The sector faces immense challenges, because of its intangible nature and lack of understanding tourist preference.

This study attempts to identify the major determinants factors to predict tourist preference in Ethiopia. This research used the CRISP data mining methodology and 10484 tourist data were taken from Ministry of Culture and Tourism of Ethiopia for experiment that was collected from 2008 - 2012.

Experiments conducted using the J48 decision tree and Naïve Bayes algorithm for classification and the Apriori algorithm of association rule in Waikato environment for knowledge analysis (WEKA). J48 decision tree algorithm with the overall model accuracy of 94.8 % has offered interesting rules. The results of this study have showed that the data mining techniques are valuable for predicating tourist preference.

Key words: - Tourism, Data mining, Classification, and Association.

CHAPTER ONE

1. INTRODUCTION

The volume and diversity of data captured by companies today are staggering. This exponential growth in collecting data increases the demand for data analysis and the need to turn this information into business processes and actionable plans to make smarter strategic decisions.

The idea of using computers to search for relevant pieces of information popularized in the article "*As We May Think*" by Vannevar Bush in 1945. The first automated information retrieval systems introduced in the 1950s and 1960s. The new technology has radically changed our society and economy. Information storage and retrieval activities, technology have the potential to realize the ultimate dream of the information retrieval specialist to make information available to any person, when and where it is required. Over the last four decades, the use of computer technology has evolved from the gradual automation of certain business operations, such as reservation, accounting, and billing, in today's integrated computing environments, which offer end-to-end automation of all major business processes. Not only the computer technology has changed, but also how that technology viewed and how it is uses in business has changed (9).

Uncovering new, interesting, and useful knowledge of tourist data can be helpful for tourism organizations in order to identify tourists' activity patterns and their preferences. Tourism is one of the largest and rapidly growing industries in the world. According to the World Tourism Organization (92), there were 846 million international tourist arrivals in 2006 only, which showed an increase of 5.4% over 2005. However, the developed world is taking the lion's share of the market to Europe, North America, and East Asia claim 76.3% the international tourists in the same year (92).

Every day, millions of people travel around the globe for business, vacations, sightseeing, or other reasons. An excessive amount of money spent on tickets, accommodations, food, transportation, and entertainment. Tourism is an information-based business where there are two types of information flow. One flow of information is from the providers to the consumers

or tourists, which is information about goods that tourists consume such as tickets, hotel rooms, entertainments. The other flow of information follows a reverse direction. It consists of

aggregate information about tourists to service providers(42).When the aggregated data about the tourists is presented in the right way, analyzed by the correct algorithm, and put into the right hands, it could be translated into meaningful information for making vital decisions by tourism service providers to boost revenue and profits.

Data mining can be a very useful tool for analyzing tourism-related data. Data mining is the process of automatically discovering useful information in large data repositories (97). It uses machine learning and statistical visualization techniques to discover and present knowledge in a form that is easily comprehensible to humans. Discovering new, stimulating and valuable knowledge by using a variety of techniques such as; classification, clustering, and association rules are the main aim of data mining. In the meantime, the necessity of using data mining in tourism is inevitable.

Focusing on using a decision tree algorithm helps to analyze tourist markets from two different dimensions: including impact factors of tourist spending and influence factors of tourist destinations (104). The first step helps to focuses on total expenditure of tourists in the classification objects in the decision tree model. While the second step helps focuses on attributes and values including cost, traffic, shopping environment, dining etc.

Data mining is present on helpful occasions of every part of tourism such as the hotel industry; for instance, by offering tourist information kiosk to provide customer information during their stay and then analyzing their data (105).

Ethiopia has a large tourism potential owing natural, historical and cultural gifts. Its' potential is diversified that include some of the highest and lowest places in Africa along with immense wildlife, including some endemic ones. Very old and well preserved historical traditions, churches, and castles to witness that, an attractive cultural diversity of about 80 nations and nationalities; and various ceremonies and rituals of the Ethiopian Orthodox Church that open a window on the authentic world of the Old Testament (83).

As tourism revenue in Ethiopia nearly doubled in the last decade, the Ethiopian tourism industry expanded its capacity considerably by providing accurate, reliable, and timely information for tourists (83). Such increase of revenue in the tourism sector has led to overproduction of tour and travel organization and subsequent competition among them has increased.

Currently, there are many tourism and travel industries in Ethiopia. These organizations used a small portion of data (daily, weekly, monthly, and yearly statistical data report) for their activities like decision-making, revising the existing rules, policies and to induce new policies. Nevertheless, this has limited capacity to discover new and unforeseen pattern and relationships that is hide in conventional database. The effort of reducing and controlling such unmanageable tourist data by identifying tourists' preferences, take preventive measures on such factors and improve the quality of life manually and statistical method with a small portion of data is time consuming, doesn't give motivating result, it is error prone and a difficult task .

Data mining techniques have been recognize as powerful tools for predictive modeling tourist decision-making process **(**42). However, two practical yet important problems have not been resolve by the data miners in empirical tourism research. Firstly, comprehensibility-the role of the data mining should not only generate accurate predictions, but also provide insights why certain prediction made. Secondly, lack of using enough data for training samples and testing. Many data mining methods may not achieve satisfactory performance if learned on small data sets.

1.1 Statement of the Problem

Ethiopia's degree of tourist attraction to the others four African countries (Egypt, South Africa, Kenya and Tanzania), is not strong enough to penetrate the tourism market (62). However, as these all countries possess their own tourism supplies, Ethiopia also has its own supplies and yet the flow of international tourists to Ethiopia is very little. This claim comes out from the figure of World Tourism Organization international tourist's statistics report of 2012, which is register within Ethiopian tourists' arrival for the last fifteen years. It is by much less than other four major African countries.

Ethiopia has seen an annual growth rate of 10.1% from 1991 to 2008, which is higher than the African average in a similar period, implying Ethiopia is gaining a portion of the African market share (100).

Tourism is a service industry: therefore, there are inherent challenges with service marketing that affect how the tourism product is communicate to the consumer public (95). Major reason for unfulfilled potential lies in most tourism marketing is, focusing on the destination or outlet (in other words the products being offered) and lacking focus on the consumer which means lack of understanding tourist preference and experiences (64).

Data mining tools and techniques in the tourism industry can be used to forecasting expenditures of tourists, analyzing profiles of tourists, and forecasting number of tourist arrivals (42).

The reasons behind the sector's poor performance are lack of understanding the tourist preference (83). Therefore, the concern of this thesis is to identify and detect the pattern that helps to understanding tourist preference using data-mining techniques and tools.

In carrying out the study, an attempt made to address and seek to answer the following research questions:-

1. What data mining algorithms and models are more suitable for predicting tourist preference?
2. What are the main determinant factors of tourist arrival to Ethiopia?
3. Which tourists are likely to return to the same tourist attraction?

1.2 Objectives

1.2.1 General Objective

The general objective of this study is to predicate tourist preference and generate association rules using data mining methods and techniques.

1.2.2 Specific Objectives

In order to achieve the general objective, the following specific objectives were attempt in the present research:

1) Review different literatures that can support the study.
2) Select, collect, and prepare data set required for experiment.
3) Pre-processing data in order to have cleaned dataset, that is suitable for data mining algorithm.
4) Select data mining tool and algorithms to be use in this study.
5) Build model that predict tourist preference.
6) Evaluate the performance of the model.
7) Finally, interpreting results, and then coming up with conclusion and recommendations.

1.3 Scope and limitation of the study

This research conducted based on the data obtained from the Ministry of Culture and Tourism of Ethiopia, considering that it represents all tour and travel agencies, regional tourism bureaus and other hospitality industries in Ethiopia. The tourist data used for this study collected between the period of 2008 and 2012.

The scope of the thesis is limited predict foreign tourist preference.

1.4 Significance of the Study

The developed model can have many advantages for the tourism sector and other related organizations. Those are:

- The output of this study can be an input for further research in this and other related areas in the context of Ethiopia.

- This study can give hands on experience for the researcher for understanding studies in the future.
- The finding of this study can help the Ministry culture and tourism of Ethiopia to identifying major tourist preferences and reach at measurable and actionable recommendations in promotional planning, decision-making and inducing new policies.
- The tourism service provider could be identify the patterns of activities preferred by the tourists with the geographical area they are from; it is likely that they could create suitable packages for tourists from the given regions.
- Can help to create marketing strategies and maximize organizational profits.
- Travel agencies can maintain regular clients, classify their preferences, and leverage short-term service costs, which in turn optimize profitability.
- This study will give an optimal method in predicting tourist preferences
- In general, the society, the government, the researchers, domain experts, policy makers, National Tour and Travel Organization will get benefit from this research.

1.5 Methodology

Methodology is a process that mainly consists of intellectual activities. Usually only the end goal of the methodological process manifested as the product or result of the physical work (76). It is evident that a certain set of steps is usually required to accomplish a certain task. These sets of steps could guide which activity to do first and keep on doing in a chronological order. The choice of following the set of steps depends on how one is familiar with them and depending on the immense benefits they offer compared with others.

Data mining often described as the process of discovering correlations, patterns, trends, or relationships by searching through a large amount of data stored in repositories, corporate databases, and data warehouses. Humans, in that sense, are limit by information overload; thus, new tools and techniques are being develop to solve this problem through automation. Data mining uses a series of pattern recognition technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases.

1.5.1 Study Design

For conducting this research, the six-step process model of knowledge discovery in database (KDD) process model was implemented. This model developed by adopting the CRISP-DM to the needs of academic research community. The CRISP-DM consists of understanding the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge steps.

This research is an academic research and results of the research can possibly deployed to bring a solution.

Cross Industry Standard Process for Data mining (CRISP- DM) Steps

1. Business understanding – This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

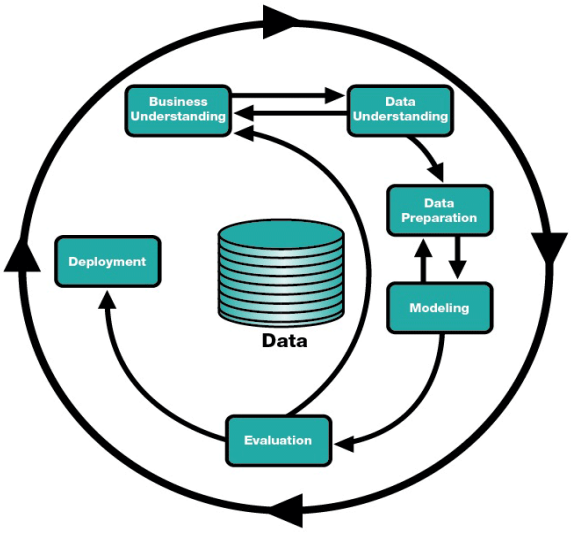


Figure 1: CRISP – DM

1. Data understanding – This step starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information. This step includes collecting sample data and deciding which data, including format and size. Data checked for completeness, redundancy, missing values, the plausibility of attribute values, etc. Finally, the step includes verification of the usefulness of the data with respect to the data mining goals.

   In order to understand the nature of the data and the attributes of tourist information details storage mechanism, brief explanation made by the record and data clerks officers. The information about each tourist is stored in MS-EXCEL, MS-WORD and hard copy in English language in 2 categories: foreign and non- foreign tourists. The data, which are tourists' data, filled and reported by the clerk. It contains full details of each tourist. That is, the data contain detailed information of a tourist such as name, gender, age, occupation, nationality, travel frequency, travel destination preferences, travel purpose, attraction preferences, a source of motivation, length of stay, payment methods, destinations to visit and travel arrangement.

2. Data preparation – Covers all activities to construct the final dataset from the initial raw data. The data that collected and stored in excel format passed through important steps of data pre-processing, data cleaning, data reduction, and data transformation. This step concern deciding which data used as input for data mining. It also involves data cleaning, which includes

checking the completeness of data records, removing or correcting for noise and missing values, etc. The cleaned data further processed by feature selection. In fact the ideal practice for variable selection is to take all the variables in the dataset, feed them to the data mining tool and let it find those which are the best predictors.

Feature Selection Method:-Feature selection is one of the dimensionality reduction technique used in data mining. It often used as data preprocessing method before applying to any classification algorithm. This reduces high dimension data by selecting useful attributes only for specific class. Redundant and irrelevant features omitted while doing feature selection. There are three standard approaches: Embedded, Filter and Wrapper. In the embedded approach algorithm, it decides which approaches are used. A wrapper approach uses target-learning algorithms to find relevant feature subset, while in filter approach features are select before applying a learning algorithm. In this research, wrapper and filter with best first search approaches used in experiments in this study.

Wrapper Method:- The wrapper is technique for selecting best subset of features using a specific classification algorithm (5).The difference between embedded and the wrapper approach is that the wrapper has internal cross validation while embedded is not having. It uses a target data-mining algorithm as a black box to select a best feature subset. This method takes into account feature dependencies while searching and building a model. The search was conduct using possible parameters. The goal of search method is finding the state, which is having maximum evaluation.

3. Modeling – In this phase, various modeling techniques are select and applied and their parameters calibrated to optimal values. At this point, that data mining models and tools used to interrogate the data and convert it into knowledge for decision-making. In addition, at this stage, selecting a particular data mining method that matches the goals of the data mining process defined in the first step. However, the details of building and training a model vary from technique to technique and hence there are no blue print procedures. For this reason prior to training and building a model, J48 decision tree and naïve Bayes of classification used. As well as Apriori for detecting association rule used in WEKA open source software.

4. Evaluation – At the stage the model (or models) obtained are more thoroughly evaluate and the steps executed to construct the model reviewed to be certain it properly achieves the business objectives.

   Evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge. Only approved models retained, and the entire process revisited to identify which alternative actions could have been take to improve the results. The fact that classification capability of any model depends on the learning ability of the datasets provided. The model with the highest true positives and less false negatives means there is a likelihood of classifying the data sets correctly, making the precision and the recall values to be maximum(76).

   Performance Measurements:- The performance of classification algorithms is usually examine by evaluating the accuracy of the classification. However, since classification is often a fuzzy problem, the correct answer may depend on the user. Traditional algorithms evaluation approaches such as determining the space and time overhead can be used. However, these approaches are usually secondary. Determining which best and excellent is depends on the interpretation of the problem by users. To compare j48 and Naïve Bayes the following performance measurement used in this study:-

   Classification accuracy**: -**usually calculated by determining the percentage of tuples placed in a correct class.

   ***ROC:* -** receiver operating characteristic curve or ROC (relative operating characteristic) curve shows the relationship between false positives and true positives. An operating characteristic curve originally used in communication area examined false alarm rates. It has also used in information retrieval to examine fall out which is the percentage of retrieved that is not relevant verses recall the percentage of the retrieve that are relevant (40).

   Area under ROC curve often used as a measure of the quality of the classification models. A random classifier has an area under the curve of 0.5, while Area under Curve (AUC) for a perfect classifier is equal to 1. An area under the ROC curve of 0.8, for example, means that a randomly selected case from the group with the target equals 1 has a score larger than that for a randomly chosen case from the group with the target equals 0 in 80% of the time. When a classifier cannot distinguish between the two groups, the area will be equal to 0.5 (the ROC

curve will coincide with the diagonal). When there is a perfect separation of the two groups, i.e., no overlapping of the distributions, the area under the ROC curve reaches to 1 (the ROC curve will reach the upper left corner of the plot).

Mean Absolute Error (MAE):- the MAE is lower, the performance is better.

TP Rate: rate of true positives (instances correctly classified as a given class)

FP Rate*:* rate of false positives (instances falsely classified as a given class)

Precision and Recall: **-**Precision and Recall are the most commonly used measures in information retrieval systems. Precision indicates the accuracy of recommendation system and recall measures the extent to which recommender system can recommend the items that are interesting to the user. Precision and Recall are defined by using the confusion matrix. Precision is a measure of exactness, determines the fraction of relevant items retrieved out of all items retrieved. Recall is a measure of completeness, determines the fraction of relevant items retrieved out of all relevant items.

5. Deployment – Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and present in a way that the customer can use it (17).

1.6 Organization of the Paper

This research divided into five chapters. The first chapter is an introduction part, which contains background to the research work, a statement of the problem addressed, objective of the research, and methodologies adopted for the study.

The second chapter mainly discussed the Literature reviewed on the different data mining models, techniques used.

The third chapter focuses on data understanding of the tourist. In addition, this chapter includes the activities undertaken to understand the business functionality of the Ministry of culture and tourism the data collection, data preprocessing and statistical summarizations of attributes under study. The Fourth chapter is the experiment, which provides discussions about the output of the models. Finally, in the fifth chapter conclusions and recommendations are present based on the research findings.

CHAPTER TWO

2. REVIEW OF LITERATURE

Recreational and educational travel already existed in the classical world and even earlier, in Egypt under the Pharaohs. In the latter, there is evidence of journeys emanating from a luxury lifestyle and the search for amusement, experience, and relaxation. The privileged groups of the population cultivated the first journeys for pleasure. They visited famous monuments and relics of ancient Egyptian culture, including, for example, the step pyramid of Sakkara, the Sphinx and the great pyramids of Gizeh – buildings that had been constructed a good thousand years earlier. The Greeks had similar traditions. They travelled to Delphi in order to question the Oracle, participated in the Python Games (musical and sporting competitions) or the early Olympic Games (34).

Classical Rome also gave impetus to travelling and particular forms of holiday. Holiday travel became increasingly important due to the development of infrastructure. Around 300 A.D., there existed a road network with 90,000 kilometers of major thoroughfares and 200,000 kilometers of smaller rural roads. These facilitated not only the transport of soldiers and goods, but also private travel. Above all, wealthy travelers seeking edification and pleasure benefited from this system. In the first century after Christ, there was a veritable touristic economy, which organized travel for individuals and groups, provided information, and dealt with both accommodation and meals. The well-off Romans sought relaxation in the seaside resorts in the South or passed time on the beaches of Egypt and Greece. The classical world did not only have the "bathing holiday", but also developed an early form of "summer health retreat" in swanky thermal baths and luxury locations visited by rich urban citizens during the hot months. Something that had its origins primarily in healthcare soon mutated into holidays for pleasure and entertainment, which could also include gambling and prostitution. The decline of the Roman Empire caused the degeneration of many roads. Travel became more difficult, more dangerous, and more complicated.

The mobility of mediaeval corporate society was shaped by its own forms and understandings of travel tailored to diverse groups, including merchants, students, soldiers, pilgrims, journeymen, beggars and robbers. From the twelfth century, the movement of errant scholars

became increasingly important. Journeys to famous educational institutions in France (Paris, Montpellier), England (Oxford), and Italy (Bologna) became both a custom and a component of education. The desire to experience the world emerged as an individual, unique guiding principle. Travelling tuned from a means into an end: now, one travelled in order to learn on the road and developed in doing so a love of travel and life that not infrequently crossed over into licentiousness and the abandonment of mores. With regard to the motivation to travel, one can see here an important process with long-term repercussions – travelling and wandering, since then, been seen as a means of confronting oneself and achieving self-realization.

Tourism has been consider as an export industry since foreign tourists who travel abroad purchase goods and services with money from their home countries. Tourism markets governess by national regulations. The liberalization of trade in tourism and travel-related services can also take place through the General Agreement on Trade in Services (GATS) of the World Trade Organization (WTO), at the multilateral level, as well as through regional trade agreements (RTAs) covering trade in services at the regional level. Regulatory commitments under such agreements can play a significant role in promoting tourism, including intra-regional tourism among developing countries. By reducing regulatory barriers through these agreements, countries can enhance the gains from tourism trade for firms, workers, and consumers (92).

One of the most crucial aspects of international tourism is the cross-border movement of consumers. This permits even unskilled workers in remote areas to become services exporters for instance, by selling craft items, performing in cultural shows, or working in a tourism lodge.(93)

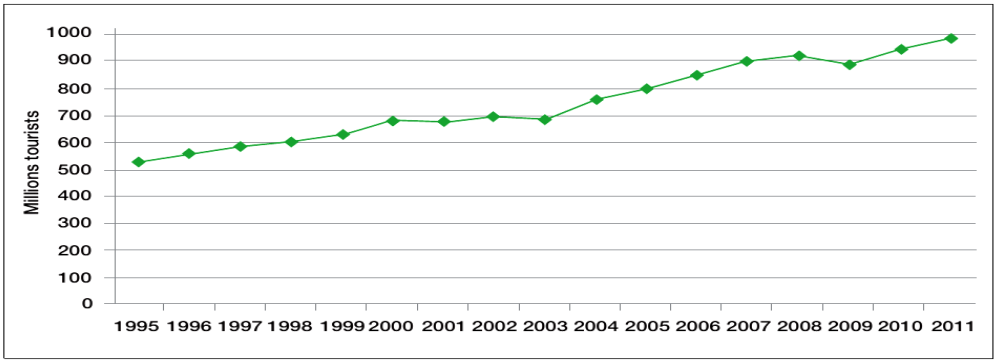


Figure 2:- Tourism Growth Source: UNWTO 2012

2.1 Precursors of Modern Tourism

An early form and precursor of modern tourism was the grand tour undertaken by young nobles between the 16th and 18th centuries. This possessed its own, new structures that clearly defined by corporate status: the original goal was to broaden one's education, mark the end of childhood and acquire and hone social graces; however, over time, leisure and pleasure became increasingly important. On the one hand, this created the differentiated paradigm of travel "as an art" (11). On the other, the search for amusement and enjoyment implied an element of travelling as an end in itself (70). The classic grand tour lasted between one and three years. Route, sequence, and contacts, not to mention the educational program, planned down to the last detail. The aristocrats travelled with an entourage of the equerries, tutors, mentors, protégés, domestic servants, coachmen, and other staff. These provided for safety, comfort, education, supervision, and pleasure in accordance with their specialized area of responsibility.

Ethiopia, the oldest independent nation in Africa, has a heritage dating back to the first century AD. Traders from Greece, Rome, Persia, and Egypt knew of the riches of what is now Ethiopia, and by the first century AD, Axum was the capital of a great Empire. This realm became one of the first Christian lands of Africa. Late in the 10th Century, Axum declined and a new Zagwe dynasty, centered what is now Lalibela, ruled the land. Axum, Lalibela and Gonder now provide our greatest historical legacy. It was in the 16th Century that the son of the great explorer Vasco DA Gama came to Ethiopia, but then found a land of many kingdoms and provinces beset by feuds and War (72).

Ethiopiahas almost all types of primary tourist products: historical attractions, national parks with endemic wildlife and cultural and religious festivals. UNESCO recognizes nine world heritage sites (as many as Morocco, South Africa and Tunisia and more than any other country in Africa): Axum’s obelisks, the monolithic churches of Lalibela, Gondar's castles, the Omo Valley, Hadar (where the skeleton of Lucy was discovered), Tia's carved standing stones, the Semien National Park, walled city of Harar and Meskel ceremony (101).

Tourism in Ethiopia dates back to the pre-Axumite period when the first illustrated travel guides to Ethiopia found in the friezes of the pyramids and ancient sites of Egypt. These depicted travels to the land of Punt, which the Egyptians knew was the source of the Nile, and where they traded for gold, incense, ivory, and slaves. The fourth century Persian historian Mani

described the Kingdom of Axum as being one of the four great empires of the world, ranking it alongside China, Persia and Rome.

Modern tourism in Ethiopia said to have started with the formation of a government body to develop and control it in 1961: the Ethiopian Tourist Organization.

Legend has it that Emperor Menelik I, the son of the Queen of Sheba and King Solomon, brought the Ark of the Covenant from Jerusalem to Axum, where he settled and established one of the world's longest known, uninterrupted monarchical dynasties.

This is only one example of Ethiopia's magnificent history, which encompasses legend and tradition, mystery and fact, from a powerful and religious ancient civilization. The well -trodden path through Ethiopia's famous and fascinating historic places takes through a scenically magnificent world of fairy -tale names, such as Lalibela, Gondar, Deber Damo and Bahar Dar (101).

2.2 Destination Drivers/Attributes

Destination drivers are those attributes of the destination that can be associated with the destination and that correspond with the values and the actual needs of prospective tourists and have the likelihood of evoking an image that will stimulate tourist's interest to visit such a destination.

In real life we are aware of such generally shared associations, like Mercedes=prestige, Volvo=safety, Kenya=safari, Porsche =wild driver, etc. It is important to determine the destination attributes that will build strong associations. A simple methodology proposed by Esu and Mdaze-Arrey on —"Branding of Cultural Festivals as Destination Attraction…" could be used to determine the destination attractions that will serve as connects with the destination. The method begins with the generation of destination attribute, selection of significant attributes using two statistical techniques: importance-performance analysis matrix and discriminate analysis to test significant associations. This is necessary because the attributes must agree with the tourist's value, needs, and motivations. A need is a state of disequilibrium that requires satisfaction. Motivation is the drive to satisfy the identified need. The concepts of push and pull factors often used to explain the concept of need and motivation. Push factors are the socio-psychological needs that will encourage a person to travel, while the pull factors is one in which the person is motivated, or arouse by the destination. The push factors are logical and temporary antecedents to pull factors. The destination must possess attributes that match the tourist's needs

before the tourist would respond positively to the promotional strategy. The attributes that show significant relationship with market segment are those ones that are conceptualize as destination drives.

Attributes helps to measure destination image (42)

- ✓ Scenery/natural attractions
- ✓ Hospitality/friendliness/receptiveness
- ✓ Climate
- ✓ Costs/price levels
- ✓ Nightlife/entertainment
- ✓ Sports facilities/activities
- ✓ Shopping facilities
- ✓ Personal safety Different cuisine/food/drink
- ✓ Restful/relaxing
- ✓ Historic sites/museums
- ✓ Accommodation facilities
- ✓ Different customs/culture
- ✓ Local people
- ✓ Tourist sites/activities
- ✓ Local infrastructure/ transportation
- ✓ National parks/wilderness areas
- ✓ Architecture/buildings
- ✓ Beaches
- ✓ Crowdedness
- ✓ Cleanliness
- ✓ Cities
- ✓ Accessibility
- ✓ Opportunity for adventure
- ✓ Facilities for information/tours
- ✓ Atmosphere (familiar versus exotic)
- ✓ Economic development/affluence
- ✓ Family or adult oriented
- ✓ Opportunity to increase knowledge
- ✓ Quality of service
- ✓ Fairs/exhibitions/festivals
- ✓ Extent of commercialization
- ✓ Political stability
- ✓ Fame/reputation/fashion
- ✓ Degree of urbanization
- ✓ Friends and relatives
- ✓ Wildlife
- ✓ Sophistication
- ✓ Interesting
- ✓ Busy/exciting

2.3 World Tourism Arrivals (2008-2012)

According to the UNWTO report, in 2009, worldwide international tourism recorded a 35 million decline down 3.7 percent from 2008. The uncertainty around the H1N1 and influenza pandemic, the air tariff disruption caused by the volcanic eruption in Iceland, and the economic uncertainty affecting the Euro zone turned 2009 into one of the toughest years for tourism, especially for Europe and America. .

International tourism continued to recover strongly in the year 2010 reaching 940 million that increase more than compensated the decline in the year before with an added 58 million arrivals, up 7.8% over the former highest recorded arrival year of 2008 (93). Asia was the first region to recover and the strongest growing region in 2010, with international tourist arrivals reaching a new record of 204 million but recovery was slower for the Americas and Europeans. Meanwhile Africa, which was the only region to show positive, figures in 2009. Maintained growth during 2010 hosting of events such as the FIFA World Cup in South Africa .In 2011, Worldwide, international tourist arrivals grew by 4.6 percent in 2011 to 983 million maintaining momentum, despite persistent economic turbulence in the euro zone, major political changes in the Middle East and North Africa, and the natural disaster in Japan (93).

Europe and Asia and the Pacific both were the fastest-growing regions (6 Percent) in terms of tourist arrivals in 2011. Popular uprisings in North Africa and the Middle East during 2011 took a toll on tourism in both regions: Africa recorded a 1 percent increase, only, due to the loss of visitors in North Africa, while the Middle East saw an 8 percent decline in arrivals.

Region		**Tourist Arrivals**			
		Year			
	2008	**2009**	**2010**	**2011**	**2012**
Africa	115,999	150,102	140,076	160311	180294
Europe	95,354	118,689	136,690	162,784	170653
Americas	59,240	77,826	95,203	96,246	117082
South Asia	11,195	12,404	15,366	20,746	19182
East Asia & The Pacific	19,477	26,448	33,393	28,884	46515
Middle East	25,228	37,428	42,301	47,583	53472
Oceania	3,664	4,340	5,221	5,874	5367
Total	**330,157**	**427,286**	**468,305**	**523,438**	**596341**

Table 1: International visitor's arrival of Ethiopia by region (2008-2012)

2.4 Major Sources of Tourist

USA exists as the leading visitor source to Ethiopia, accounting up to 16 percent. Below figure 5 reveals the top ten major sources of markets and shows that the ranking of markets remained virtually unchanged in the years under review for this research/2008 – 2012.

Of the top 10 inbound visitor markets, the United Kingdom, China, and Germany take second and third place respectively in the year 2012 with a relatively significant share of the market. Comparing the arrivals from the top 10 inbound visitor markets, France and Saudi Arabia were the only markets that experienced declines over the years, whereas, arrivals from the other seven visitor markets posted double-digit increases.

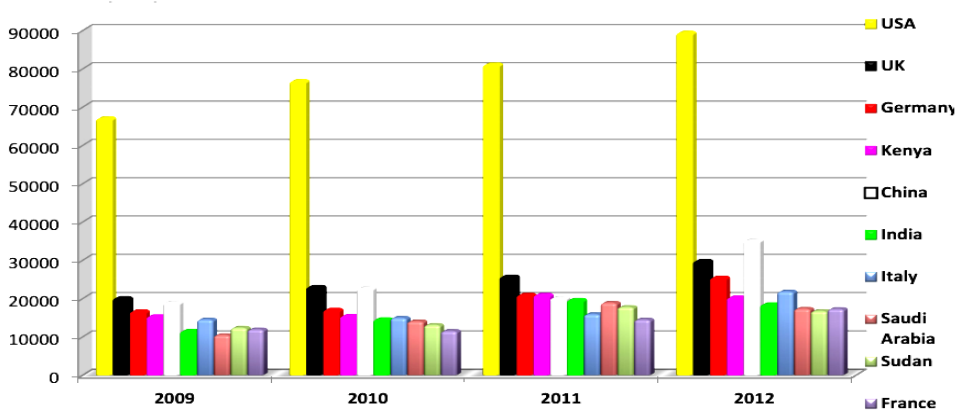


Figure 3: Top 10 inbound visitor /Source: MoCT, statistical bulletin 2013

Accommodation Capacity: The Ministry of Culture and Tourism has adopted a system of approving and classifying the hotels based on the facilities and services provided by them. The total numbers of registered tourist hotels are 595 with total 19998 rooms.

Tourist Flow by Travel Purpose: Despite the increases in the traffic, the pattern of reasons for tourist visiting Ethiopia remains more of less the same from 2008 -2012. The main motivations to do tourism in Ethiopia were pleasure. This reflects that Ethiopia chosen for leisure travel.

year	Purpose of visit						
	Business	Conference	Leisure & Holiday	Transit	Visiting Relatives& Friends	others	Total
2008	49209	15721	99394	77572	25482	62779	383399
2009	71374	47516	138070	81481	35593	53253	427286
2010	77816	36145	171414	84229	28672	70029	468305
2011	91,064	50,531	183,008	86,020	37,116	75,699	523,438
2012	130321	57475	191537	90990	45174	80844	596341

Table 2: Purpose of visit from 2008 -2012

Tourist Flow by sex categories

		Year				
		2008	2009	2010	2011	2012
		Count	Count	Count	Count	Count
Sex	Female	986520	119640	131125	146563	429366
	Male	250710	307646	337180	376875	166975
	Total	1237230	427,286	468,305	523,438	596341

Table 3:- Tourist Flow by sex categories

Tourist flow by age group

		Year			
		2009	2010	2011	2012
		Count	**Count**	**Count**	**Count**
Age Group	Unstated	31534	34561	38630	67088
	< 19	21492	23556	26329	108176
	20-40	161557	177066	197912	108176
	41-60	173649	190319	212725	130002
	61-84	38413	42101	47057	126901
	> 85	6409	7025	7852	31010
	Total	427,286	468,305	523,438	596341

Table 4:- Tourist flow by age group

2.5 Data Mining For Travel and Tourism

Travel and tourism industry is one of the main users of information technology. Progresses information technology affects the services and facilities offered and how they delivered and promoted. It is also affecting the organizational structure and interactions between customers and service providers. Travelers are increase used by internet and communication technology to find places that meet needs and expectation. According to Bualis (2002), integrated knowledge of tourist characteristics, images, attitudes, and preferred destination attributes should use to market destination more easily. Tourist sites and hotels can use data mining: to create a direct mailing campaign, plan seasonal promotions, plan the timing, placement of advertisement campaigns, create personalized advertisement, define which market segments are growingly most rapidly and determine the number of rooms to reserve for wholesale customers and travelers (19). In addition, there is a data explosion in the travel and tourism. Generation of centralized reservation and property management systems has resulted in big amount of data for tour agencies. At the same time, more access to more data.

That globalization has changed tourist consumer behavior. Globalization has the capacity to create impacts on cultural criteria (culture, subculture, and social class), social criteria (reference groups, family roles, and status), personal criteria (age and life cycle stage, occupation, economic circumstances, personality), and self-concept psychological criteria (motivation, perception, learning, beliefs and attitudes). All these indicated that factors impacted by globalization dynamic, psychological factor of the tourists considered most important as it directly involves tourist consumer behavior (82).

New generation travelers are more complex and highly demanding on the quality of products. These travelers have known very well about attractions and tourism products. They had many experiences to spend time and money to travel. The new travelers like to compare details of products and choose the suitable items for themselves and they use the internet to search for information by themselves more than asking agency for services (41). A new generation travelers' plan a travel trip by themselves and wanted to face the enjoyable and exciting situation which is more unpredictable rather than traveled following their plan, so searching travel information setup in customer's is necessary to be put into account before making a decision all the time. The new travelers are not just passive consumers anymore.

There are many classification algorithms in the Waikato Environment for Knowledge Analysis system. It is a group of machine learning algorithms and data processing tools implemented in Java in 1993. It developed as support for the whole process of experimental data mining such as preparation of input data, statistical evaluation of learning schemes, visualization of input data, and the result of learning and used for education, research, and applications. Its main features are 49 data preprocessing tools, 76 classifications/regression/MLP algorithms, 8 clustering algorithms, 15 attribute/subset evaluators + 10 search algorithms for feature selection, 3 algorithms for finding association rules, 3 graphical user interfaces such as explorer (exploratory data analysis), the experimenter (experimental environment) and knowledge flow (new process model interface).

2.5.1 Association rule

By using association technique mining, tourism organizations can identify different types of tourist profiling behavior (99).

This technique is a data mining technique, which introduced by Agrawal in 1993. He proposed the algorithm of Apriori in 1994. The main idea of Apriori is to find a relationship between different items of a database. The association technique is useful to find a relationship between two or more attributes and it attempts to find rules between attributes.

Association Rule is a powerful data analysis technique that appears often within the data mining subject area. The analysis of association rule used in a variety of ways and in many applications, including merchandise stocking, insurance fraud investigations, medical diagnosis, and even climate prediction. However, there is a distinct lack of growth of data mining applications within the tourism industry.

This research presents and emphasizes data mining and in particular the association rule of Apriori algorithm. The algorithm can be help to detect patterns for the tourists based on past tourist experiences and their motivation. It will be focus on recommending many attractions, and detecting tourist preference.

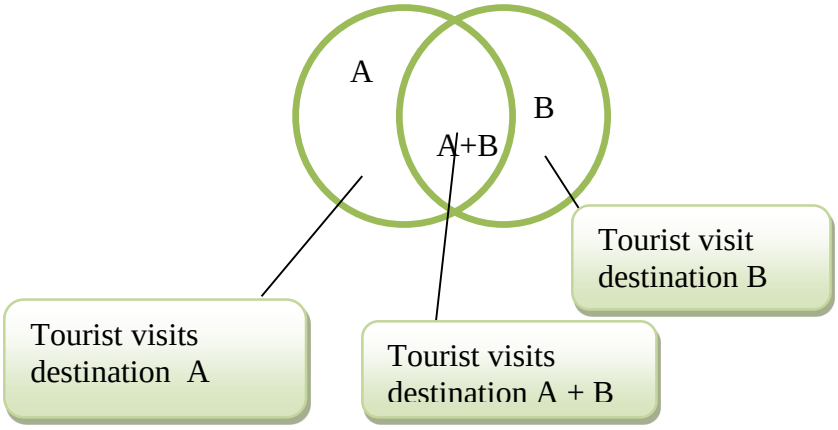


Figure 4: Association rule

Association relates to the market basket analysis of hotels, airlines and other services among visitors for the principle of partner selection and marketing alliances. A market analysis of the preferred products among possible visitors is an important analysis to carry out before an investment made (31). This would allow travel and tourism business for the possible understanding of travelers' interest and needs, then able to offer specially designed packages (103).

Apriori Algorithm: There are many association rule algorithms. These algorithms can be divide into two classes; the first one mainly focused on improving the analytical efficiency of the association rules. The other one pays more attention to the application of the association rule algorithm and how to deal with value type variables and promotes the association of the single concept layer to multiple concept layrs include and further reveals the inner structure of objects. Apriori algorithm is one of the classical association rule algorithms; Agrawal et al proposed the earliest Apriori algorithm. The algorithm mainly including two parts: producing frequent item sets and producing association rules according to the frequent item sets. The algorithm scans database, accumulates each item count, and collects the items that meet the minimum support (min_sup), finds out the frequent-one item sets, and named it. Then, the algorithm uses to find out the second frequent item sets and find out the frequent two-item sets and so on and keeps doing this until it cannot find out the frequent -item sets. In these frequent item sets, it will be define as a strong-association rule, if it reaches the minimum confidence (36).

2.5.2 Classification Model Techniques

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for being able to use the model to predict the class of objects whose class label is unknown.

J48 decision tree algorithm: One of the decision tree algorithms used for this study was the J48 decision tree algorithm which is the successor of ID3 (Iterative Dichotomiser,) C4.5. J48 decision tree is a popular utility that involves decision based classification and adaptive learning over a training set (Whitten and Frank). J48 algorithm of decision tree technique is one of classification and a prediction algorithm, which used for this study; support both numeric and nominal predicators and nominal class attribute values. The J48 algorithm is the WEKA implementation of the C4.5 top-down decision tree learner proposed by Quinlan in 1993. The algorithm uses the greedy technique and is a variant of ID3, which determines at each step the most predictive attribute, and splits a node based on this attribute. It deals with numeric attributes by determining where thresholds for decision splits should place.

An important feature of J48 is a facility of generating outputs both in tree form and rule sets. Graphically, it displays the classification process of a given input for giving output class labels. Rule sets are generally easier to understand since each rule describes a specific context associated with a class and also shows the hierarchy of the determinant factors or attributes. The decision tree algorithm takes inputs, data partition, D, which is a set of training tuples and their associated class labels, attribute list, the set of candidate attributes and attribute_ selection_ method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of splitting attribute and, possibly, either a split point or splitting subset. The output of the algorithm is a decision tree. Given training tuples, the algorithm followed by a decision tree is as followed attribute selection method specifies a heuristic procedure for selecting the attribute that best discriminates the given tuples according to class. The process of decision tree generation by repeatedly splitting on attributes is equivalent to partitioning the initial training set into smaller training sets repeatedly, until the entropy of each of these subsets is zero; i.e. each one has instances drawn from only a single class.

The entropy method of attribute selection is to choose to split on the attribute that gives the greatest reduction in (average) entropy, i.e. the one that maximizes the value of information gain. At any stage of this process, splitting on any attribute has the property that the average entropy of the result in subsets will be less than (or occasionally equal to) that of the previous training set. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simple) tree found. The expected information needed to classify a tuple in D given by:

$$Info\ (D) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Where pi is the probability that an arbitrary tuple in D; belongs to class chi and is estimated by |Ci, D /| D |. A log functions to the base 2used, because the information encoded in bits. Info (D) has been just the average amount of information needed to identify the class label of a tuple inD. At this point, the information we have only on the proportions of tuples of each class. Info (D) is also known as the entropy of D. Suppose we were to partition the tuples in database Don some attribute A having V distinct values, {a1, a2, . . ., av} as observed from the training data. If A is discrete-valued, these values correspond directly to the V outcomes of a test on A. Attribute A used to split D into v partitions or subsets, {D1, D2… DV}; where Dj contains those tuples in D that have outcome aj of A. These partitions would correspond to the branches grown from node N. Hypothetically; we would like this partitioning to produce an exact classification of the tuples. That is, we would like each partition to be pure. However, it is quite likely that the partitions will be impure (e.g., where a partition may contain a collection of tuples of different classes rather than from a single class). The amount of information we would still need (after the partitioning) in order to arrive at an exact classification measured by:

$$Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{|D|} * Info(D_j)$$

The term $\frac{|D_j|}{|D|}$ acts as the weight of the jth partition. InfoA (D) is the expected information required to classify a tuple from D based on the partitioning by A. The smaller the expected information (still) required, the greater the purity of the partitions. Information gain defined as the difference between the original information requirement (i.e., based on just the proportion of classes) and the new requirement (i.e., obtained after partitioning on A). That is,

Gain (A) = Info (D) – Info A (D)

Gain (A) tells us how much would be gain by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The attribute A with the highest information gain, (Gain (A)), chosen as the splitting attribute at node N. This is equivalent to saying that we want to partition on the attribute A that would do the "best classification," so that the amount of information still required to finish classifying the tuples is minimal (i.e., minimum_ Info A(D)).

Besides, of this, the pruned tree in decision tree has a hierarchy in that the most significant variable that used to discriminate the records is located at the top. It maximizes computational efficiency as well as classification accuracy. The process of pruning (post-pruning) traditionally begins from the bottom of the tree (at the child leaves), and propagates upwards. J48 algorithm recursively classifies until each leaf is pure, meaning that the data have been categorize as close to perfectly as possible.

The overlying principle of pruning is to compare the amount of error that a decision tree would suffer before and after each possible prune, and to decide accordingly maximally avoid error. The metric used to describe possible error, denoted error estimate (E), is calculated as

$$E= (e+1)/ (N + m)$$

Where 'E' is Error estimate, 'e' is misclassified examples at the given node, 'N' is examples that reach the given node, and 'm' is all training examples. Applying pruning methods to a tree typically results in reducing the size of the tree to avoid unnecessary complexity (produces fewer, more easily and interpretable results) and to avoid over-fitting of the data set when

classifying new data that means improve the prediction and classification accuracy of the algorithm by minimizing over-fitting. In WEKA J48 classifier, lowering the confidence factor decreases the amount of post-pruning since the effectiveness labeled by the confidence factor. Post-pruning in the C4.5 algorithm is the process of evaluating the decision error (estimated percent misclassifications) at each decision junction and propagating this error up the tree. At each junction, the algorithm compares the weighted error of each child node versus and misclassification error (if the child nodes deleted and the decision nodes were assign the class label of the majority class).

Bayesian classifier: A Bayesian classifier is, based on the idea that the role of a (natural) class is to predict the values of features for members of that class. This grouped in classes because they have common values of the features. Such classes often called natural kinds. Bayes' Theorem is about conditional probabilities. The independence of the naive Bayesian classifier embodied in a particular belief network where the features are the nodes, the target variable (the classification) has no parents, and the classification is the only parent of each input feature. This belief network requires the probability distributions P (Y) for the target feature Y and P (Xi|Y) for each input feature Xi. For each example, the prediction computed by conditioning on observed values of the input features and by querying the classification. A simplifying assumption (the "naive" part) is that the probability of the combined pieces of evidence, given this prediction, is simply the product of the probabilities of the individual pieces of evidence, given this prediction. The assumption is true when the pieces of evidence work independently of one another, without mutual interference. In other cases, the assumption merely approximates the true value. In practice, the approximation usually does not degrade the model’s predictive accuracy much, and it makes the difference between a computationally feasible algorithm and an intractable one.

Given an example with inputs X1=v1,...,Xk=vk, Bayes' rule is used to compute the posterior probability distribution of the example's classification, Y:

$$P(Y \mid X1=v1,...,Xk=vk) = (P(X1=v1,...,Xk=vk \mid Y) \times P(Y))/(P(X1=v1,...,Xk=vk))$$

$$= (P(X1=v1|Y)\times\cdots\times P(Xk=vk \mid Y)\times P(Y))/(\sum Y P(X1=v1|Y)\times\cdots\times P(Xk=vk \mid Y) \times P(Y))$$

Where the denominator is a normalizing constant to ensure the probabilities sum to 1. The denominator does not depend on the class and, therefore, it not needed to determine the most likely class.

Naive Bayes classifiers are highly scalable, requiring a number of parameters, linear in the number of variables (features/predictors) in a learning problem. Maximum likelihood training done by evaluated a closed-form expression. Which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. The Bayesian Learning Algorithms combine training data with a priori knowledge to get the posterior probability of a hypothesis. Therefore, it is possible to figure out the most probable hypothesis according to the training data. The basis for all Bayesian Learning Algorithms is the Bayes Rule.

P(h) = prior probability of hypothesis h

P(D) = prior probability of training data D

P(h|D) = probability of h given D

P(D|h) = probability of D given h One of the algorithms that assume the Bayesian conditional Probabilistic in predicting the class membership and used under this study is the Naive Bayes Algorithm.

Naive_Bayes_Learn (examples)

for each target value vj

estimate P(vj) for each attribute value ai of each attribute a

estimate P(ai | vj)

The question of how it looks like when brought into a formula is?

Let X be a set of instances xi = (a1,a2,…,an) and V be a set of classifications vj

for each target value vj

estimate P(vj) for each attribute value ai of each attribute a estimate P(ai | vj)

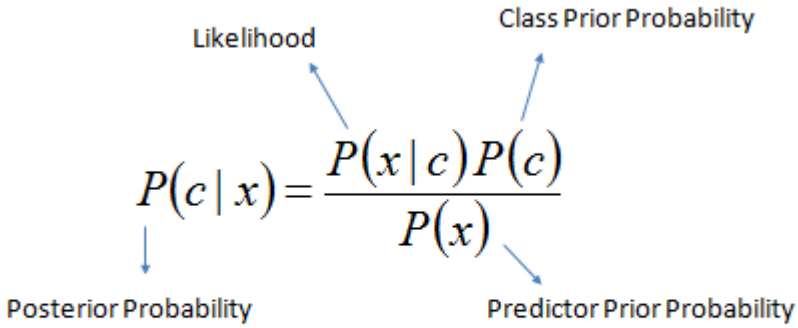
Classify_New_Instance (x)

$$v = \max_{vj \in V} P(v_j) \prod_{a_i \in x} P(a_i \mid v_j)$$

Implementation of Naïve Bayes Classifier in WEKA class uses estimator classes. Numeric estimator precision values chosen based on analysis of the training data. For this reason, the classifier is not an Updateable Classifier (which in typical usage initialized with zero training instances) -- if one needs the Updateable Classifier functionality, it is possible to use the Naïve Bayes Updateable classifier. The Naïve Bayes Updateable classifier uses a default precision of 0.1 for numeric attributes when build Classifier called with zero training instances.

Naïve Bayes Algorithm: The Naive Bayesian classifier algorithm is, based on Bayes' theorem with independence assumptions between the predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation, which makes it particularly useful for very large datasets. Despite its' simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption called class conditional independence.



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

2.6 Applications of Data Mining

Various fields of data mining application described below.

1. Retail/marketing – Data mining used in marketing/retail to identify buying patterns from customers, find associations among customer demographic characteristics, predict response to mailing campaigns and market basket analysis.

2. Banking – In this field data mining has the functions of detecting patterns of fraudulent credit card use, identify loyal customers, predict customers likely to change their credit card affiliation, determine credit card spending by customer groups, find hidden correlation between different financial indicators, and identify stock trading rules from historical market data.

3. Insurance and Health care–The potential applications of data mining in this area are claims analysis, i.e. which medical procedure claimed together, predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior.

4. Transportation –In this field, data mining can be used to determine the distribution schedule among outlets, analyze loading patterns, and analyze the performance of the engines of the airplane and to analyze road traffic accident.

5. Medicine-Data mining has also good applications in this field to characterize patient behavior to predict office visits and identify successful medical therapies for different illnesses.

6. Intrusion detection– It is a passive approach to security as it monitors information systems and raises alarms when security violations are detect. This process monitors and analyzes the events occurring in a computer system in order to detect signs of security problems.

Intrusion detection systems (IDSs) may be either host based or network based, according to the kind of input information they analyze.

2.7 Related works

1. "Applying Data Mining to Analyze Travel Pattern in Searching Travel Destination Choices "by Pairaya Juwattanasamran,Sarawut Supattranuwong and Sukree Sinthupinyo.This research focuses on the travelers who use mobile devices as a tool to search fortravel destination choices such as accommodation, tourist attraction, things to do, restaurant and souvenir shop. The researcher applied data mining method, association rules technique to analyze the relationship between travelers' profile and their transactions. They used Knowledge discovered from the database as a rule set, which provides travel information for travelers via mobile. The framework is design as a knowledge incremental learning. In addition, the paper demonstrates that applying data mining with tourism sector can increase opportunities for the competitive operations of tourism firm to respond the travelers' demand effectively.
2. "A Modified Approach towards Tourism Recommendation System with Collaborative Filtering and Association Rule Mining", by Monali Gandhi, Khushali Mistry and Mukesh Patel. This research provides a tourism recommendation system;it is based on similarities of user opinions like rating or likes and dislikes. Therefore, the recommendation provided by collaborative cannot be consider as quality recommendation. Recommendation after association rule mining is having high support and confidence level. So that will be consider as a strong recommendation. The research tries to hybrid both collaborative filtering and association rule mining to produce strong and quality recommendation even when sufficient data are not available.
3. "Critical Factors of Customer Satisfaction in Ethiopian Service Sector", by Rajasekhara Mouly Potluri and Mangnal. The study tried to explore Ethiopian service sector and customer satisfaction levels. The study measured customer satisfaction levels with recalled service encounters and the method of data collection was convenience type. The findings of the analysis showed that 36% customers in the Ethiopian service sector were dissatisfied with employees' interaction skills. Furthermore, another 47% of the customers were also disappointed with the service delivery system and 61% customers were not pleased with the service recovery process and complaint handling procedure,

respectively. Moreover, 49% of the customers expressed overall dissatisfaction with the services provided by Ethiopian service sector.

4. “An empirical study of tourist preferences using conjoint analysis,” by Shalini N. Tripathi, 2010. This research tried to address the irregularity and the various underlying factors responsible for tourist preference. The objective being the determination of customer preferences for multi attributes hybrid services like tourism and helping it to create a sustainable competitive advantage, leading to greater customer satisfaction and positive word of mouth. The researchers have been using conjoint analysis, which estimates the structure of a consumer‘s preferences, given his/her overall evaluations of a set of alternatives that are pre-specified in terms of levels of different attributes.
5. “A hybrid recommendation approach for a tourism system”, by Joel P. Lucas a, Nuno Luz b, María Moreno, Ricardo Anacleto, Ana Almeida Figueiredo and, Constantino Martins. In this thes is the researchers try to develop a recommender system for tourism, where classification based on association is applied. Classification based on association methods, also named associative classification methods, consist of an alternative data mining technique, which combines concepts of classification and association in order to allow association rules to be employ in a prediction context. The proposed methodology evaluated in some case studies, but the performance of the system was varying in different cases.

CHAPTER THREE

3. BUSINESS AND DATA UNDERSTANDING

One of the phases in the knowledge discovery process is, understands the business domain. Without a keen understanding of the business domain, no matter what tools used or how good techniques followed, may not provide useful result. Having an in-depth knowledge in the business domain enables data analysts clearly set the objectives and attempts to make to attain the defined goals (65). This initial phase focuses on understanding the objectives of the study and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.

In data mining process, there are two possibilities to be consider before using any data for experiment; in some cases when the problem is, known and correct data is available as well, there is an attempt to find the models or tools, which will be use. On the other hand, some problems might occur because of duplicate, missing, incorrect, outlier values and there is a need to make some statistical methods. To solve these kinds of problems the data mining process passes through the following five steps:

- ✓ Defining the data mining problem
- ✓ Collecting the data for mining
- ✓ Detecting and correcting the data
- ✓ Building the model
- ✓ Model validation

Defining the data mining problem: The Ministry of Culture and Tourism is responsible for formulating national policies and programs for the development, and promotion of domestic and international tourism and responsible for compilation, tabulation, and dissemination of information on various aspects of tourism in Ethiopia.

The ministry of culture and tourism stores data in two categories: inbound tourist, which involves nonresidents travelling in a country and the domestic tourist, which involves all residence tourist in Ethiopia. As we have mentioned in chapter one the scope of this study is limited to access inbound tourist only.

The objective of this study is predicting tourist preference using data mining tools and techniques. Accordingly identify the main determinant factors of tourist arrival and identifying

frequent visitor helps to predicate tourist preference. The J48 and Naïve Bayes implemented to classify tourist: as frequent and non-frequent tourist, with different length of stay. This would help to identify the determinant factors of tourist arrival.

To find the possibility of tourist arrival return to the same tourist attraction, the Apriori algorithm of association rule were deploy. The association rules implemented to find factors that a tourist impose to visit frequently and to find the possibility of visiting once again.

Collecting Data form mining: This process mainly focuses on the collection of data from different sources and locations. There are two methods, which used to collect the data mining data. These are - Internal data usually collected from existing databases, data warehouses, and OLAP. Actual transactions recorded by individuals are the richest source of information. External data: In addition to data shared within a company, data items can also collected from demographics, psychographics, and web graphics. For this study, data was collected from hard copy and soft copy of operational data which collected from 2008 -2012.

Detecting and Correcting the Data: Real-world databases usually confronted with noise, missing, and inconsistent data due to their typically huge size. Data preprocessing commonly uses as a preliminary data mining practice. It transforms the data into a format that has been easily and effectively processes by data mining algorithms (65). There are numbers of data pre-processing techniques applied for this study:-

Data cleaning: this can be apply to remove noise and correct inconsistencies, outliers, and missing values.

Data integration: - Merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube.

Data transformations: It is the process, which improves the accuracy and efficiency of mining algorithms involving distance measurements.

Data reduction: It is the reduction of the data size by aggregating, eliminating redundant features.

Building the Model: The process of estimating and building the model includes the following four parts: select data mining task, select data mining method, select suitable algorithm and, extract knowledge

Select data mining task (s): Selecting which task to use depends on the model whether it is predictive or descriptive (61). Predictive models predict the values of data using known results

and/or information found in large dataset. Classification, regression, time series analysis, prediction, and estimation are tasks for predictive model. On the other hand, descriptive model identifies patterns or relationships in data and serves as a way to explore the properties of the data examined. The importance of prediction for particular data mining applications can be varying. That means selecting which task to use depends on the model whether it is predictive or descriptive. Both descriptive and predicative used for this study.

Select data mining method (s): After selecting the task, the next step is choosing the method of the data mining. Two methods used for modeling including decision trees and association rules.

Select suitable algorithm: The next step is to select a suitable specific algorithm that implements the general methods. J48, Naïve Bayes, and Apriori used for building a model.

Extracting knowledge: This is the last step in building the model which is the result (the answers for the problem solved in data mining) after making the simulation for the algorithm.

Model Validation: In all cases, the function of the data mining models is to assist users in decision-making. However, the model by itself does not help users to give decisions; hence, there is a need to use experts to evaluate the models.

3.1 Data Source

The data employed in this research collected from the Minister of culture and tourism of Ethiopia, statistic directorate. Moreover, the data gathered from incoming passengers through the Bole International Airport and border passing.

The data were tourists' data, which filled and reported by the clerk. It contains full details for each tourist. The data contain detailed information about gender, age, occupation, departure date, full address, travel frequency, nationality, mode of transport, travel destination preferences (region to visit), travel purpose, length of stay, travel origin and travel arrangement. The data stored using MS-excel, MS word and hard copy forms in the English language, reported by the tourism statistic directorate from 2008 to 2012.

3.2 Data Understanding

Tourism is different from travel. In order to consider tourism, there must be a displacement: an individual has to travel, using any type of means of transportation. Travel refers to the activity of travelers. A tourist is someone who moves between different geographic locations, for any purpose and any duration. The visitor is a particular type of traveler and consequently tourism is a subset of travel.

In order to prevent confusion to define "Tourism", UNWTO defined it as: "Tourism comprises the activities of persons traveling to and staying in places outside their usual environment for not more than one consecutive year for the purpose of seeking new experiences, leisure, business, having fun, entertaining, sports, seeing cultural & historical places etc."

The following attribute used in the ministry of culture and tourism excel and hard copy record to keep track of the tourist detail.

1. Name
2. Age
3. Gender
4. Address
5. Travel Frequency
6. Travel Purpose
7. Attraction Preferences
8. Payment Methods
9. Source of motive
10. Length of stay
11. Travel arrangement
12. Mode of Transport
13. Nationality
14. Travel origin

Original Data	
Year	Total records
2008	41921
2009	87938
2010	98008
2011	99316
2012	114828
Total records	**442011**
Format	**. XLSX**
size	**1.14**

Table 5: Original date before preprocessing

3.3 Definition of attributes

- A business visitor is a visitor whose main purpose for a tourism trip corresponds to the business and professional category.
- Nationality is a legal concept in which a tourist holds citizenship of a certain state. It is possible to have more than one citizenship.
- Length of stay(duration) refers to the terms: “long-term” used as the equivalent of a year, “short-term” as less than one week, “Medium” less than three months, “Extended” above three months and less than one year. The final demand-side dimension that needs to be examine is how long people stay in Ethiopia. Assessing and influencing (lengthening) the length of stay of tourists in a destination is a critical role for tourism planning, particularly marketing strategies and can be one of the most influential aspects of the economic impact of tourism in a destination. Ethiopia offers to visitor an impressive collection of historic sites. While different sites of Ethiopia have large distances between them and the excessive traveling may be tedious for tourists, it is positive that it guaranties a longer stay in the country and therefore generates higher rates of expenditure.
- Attraction Preferences of tourism trip defined as the place visited that is central to the decision to take the trip.
- Cultural Life is people's customs, clothing, food, houses, language, dancing, music, drama, literature, and religion.
- Source of motivation: - the global integrating network of biological and cultural forces which gives value and direction to travel choices, behavior and experience (27).
- Travel Purpose:- The categories of “travel purpose” established by UNWTO and measured throughout the developing world on arrival declaration cards are, recreation, tourism, vacationing, research travel for the gathering of information, holiday, visit people, volunteer travel for charity, migration to begin life somewhere else, religious pilgrimages and mission trips, business travel, trade, commuting, and transit through one country to another.
- Mode of Transport: Travel may occur by human-powered transport such as walking or bicycling, or with vehicles, such as public transport, automobiles, trains and airplanes. Almost all visitors arrive in Ethiopia by flying into Bole International Airport. The airport

handles over a million passengers a year and up-graded in 2000 to a level that is more than adequate for current demand. Of the 400 or so scheduled flights into Addis Ababa each week, 290 (72%) are by Ethiopian Airline (64).

- Travel Origin: Travel origin can be transpose to a different geographical level using the term "place" a region or other subnational geographic location from where the tourist arrived. Regardless of the nationality, visitors can become from different country, especially for those who visit East Africa such as Kenya, Ethiopia, and Sudan as one package tour. Therefore, the travel origin defines that from ware tourist or visitor come to Ethiopia.
- Payment methods: The way that a tourist chooses to compensate the service of tourism and goods that is also acceptable to the service provider and seller. Ethiopia allows only three type of payment method for tourist namely: Cash Payment, bank transfers and traveler checks.
- Travel Frequency: This is simply to count the number of arrivals of specific tourist.
- Occupations: - Tourists' usual or principal work or business, especially as a means of earning income.

3.4 Data Preparation

The Ministry of Culture and tourism keeps the tourist data in flat file, MS excel and Word. The collected data for this research was not clean and suitable for analysis as it is. Therefore, the compulsory step to exercise data mining activity on data, data preprocessing took most of the research time. To do so, different data preprocessing techniques used to prepare suitable dataset for the experiment.

Copying the hard copy data to a computer took more time in the data preprocessing, In addition, most important tasks in data mining is preparing the data in a way that is suitable for the specific data-mining tool or software package. Usually, the real world dataset contains incomplete, noisy, and inconsistent data and such unclean data may cause confusion in the data mining process (97). Thus, data cleaning has become necessary in order to improve the quality of data so as to increase the accuracy and efficiency of the data mining models.

3.4.1 Data Preprocessing

The purpose of data preprocessing in the knowledge discovery process is to cleanse the data and to transform it into a form that is suitable to the subsequent steps. The data pre-processing includes a number of tasks. Common tasks presented and discussed as follows.

Data converted to attribute relationship file format (arff), to take advantage of easier data manipulation and compatible interaction with the selected tool. The data that was collect and stored in excel format passed through important steps of data preprocessing such as data cleaning, data reduction and data transformation.

3.4.2 Data Cleaning

Data cleaning is a routine work to "clean" the data by filling in missing values, smoothing, and noisy data, identifying or removing outliers, and resolving inconsistencies.

12 missing values or records that had unknown values due to the tourists and clerk language barrier were filled using WEKA. WEKA able to predict a missing value by using machine-learning algorithm and filled the missing value with the same data type. As a result, all records are prepared for data mining.

3.4.3 Data Reduction

There are a number of strategies for data reduction, these include data aggregation, dimension reduction (removing irrelevant attributes, data compression using encoding schema such as minimum length encoding and sampling).

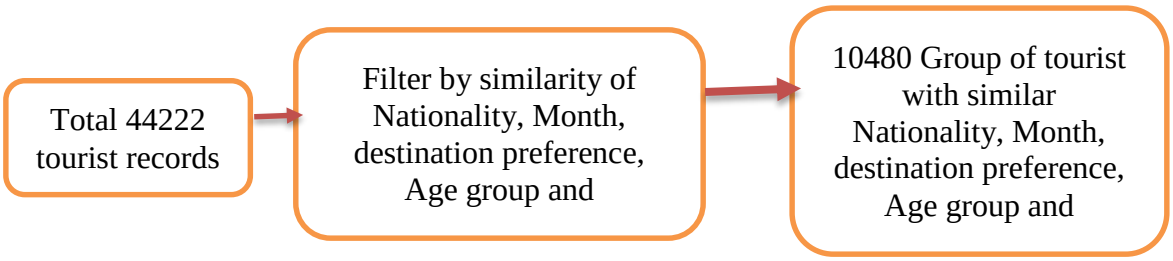


Figure 5: Grouping tourist with similar preferences

To put up a number tourist group with similar Nationality, Month, destination preference, and age group, it is important to create one attribute. Accordingly one attribute namely: "Number_ of_ Tourist" was created.

The dataset taken contains 14 attributes; some of them are irrelevant to the mining task according to the research objective specified. The research focused on tourist preferences, which used to generate patterns related to tourist preference. Name, Address, payment _ method, and mode of transport removed from the tourist dataset. Because these attributes are irrelevant to the specified objective. Therefore, due to data reduction process, the remaining 11 attributes are selected for data mining.

3.4.4 Data Transformation

Data is transfer or consolidate into forms appropriate for mining process. Two methods use during transformation the data: discretization, and concept hierarchy.

Discretization techniques used to reduce the number of values for a given continuous attribute, by dividing the attribute into a range of intervals. Interval value labels used to replace actual data values. Concept hierarchies used to reduce the data collecting and replacing to low-level concepts.

WEKA contains filters for discretization, normalization, re-sampling, attribute selection, transformation, and combination of attributes. To discretize instances in WEKA we can go through; Filters > unsupervised > Attribute > Discretize > apply > save.

Discretize instances were applied for **number_ of_ visitor**, **Age**, and **length_ of_ stay**, then replaced the encrypted values with nominal values in the word file as shown in Table 6.

Travel_ Frequency: - This attribute also contains data like one time, two times, three times and more than three. Therefore, summary or aggregation operations in this field were apply. More than two times visitors aggregated to frequent_ visitors and those who visit for the first time labeled as first_ time_ visitor.

Number_ of_ Visitors: -This attribute contains a continuous number value so by using discretization method the number of visitors are categorized into 4 different categories, such as Low_ Number of Visitors, Med_ Number _ of_ Visitors, High_ Number_ of_ Visitors , and Very_ High_ Number_ of Visitors.

Month: - This attribute generalized to higher-level concepts, which is the 12 months generalized to four seasons. This is a common categorical system in the tour and travel agencies like January, February, and March to Season_ 1, April, May, June to Season_ 2, July, August, September to season_ 3 and October, November, December are to season 4. Many hierarchies

for nominal attributes are implicit within the database schema and can be automatically define at the schema definition level.

All data transformation process result given in the following table 6.

Attribute name = Travel_ Frequency	
Old_Value	Transformed value
1	First_ Time_ Visitor
≥2	Frequent_ Visitor
Attribute name = Number_ of_ Visitors	
Old_Value	Transformed value
≥ 200	Low_No_ Visitors
≥ 500	Med_ No_Visitors
≥ 1500	High_ No_Visitors
≥ 2000	Very_ High_No_ Visitors
Attribute name = Month	
Old_Value	Transformed value
January, February, March	Season_1
April, May, June	Season_2
July, August, September	Season_3
October, November, December	Season_4
Attribute name = Age	
Old_Value	Transformed value
>=25	Age_group_1
26=< age >=45	Age_group_2
>=46	Age_group_3
Attribute name = Length of stay	
Old_Value	Transformed value
1-3 days	LoS_Short
4-7 days	LoS_Med
8 – 14 days	LoS_ Ex
15 >= days	LoS_Very_Ex

Table 6:- shows data transformation values

The data pre-processing activities in this research done by consider all this point and facts. First by understanding the problem statement clearly and considering the research objective and through an in-depth literature review as a result the following attributes taken as a candidate for the mining to reach the main objective of the research.

S.no	Attribute	Description	Type
1.	Age	Age of the tourist	Nominal
2.	Gender	Male or Female	Nominal
3.	Nationality	Nationality of the tourist	Nominal
4.	Region to visit	Destination preference	Nominal
5.	Travel Frequency	Identify first time visitor or Frequent visitor	Nominal
6.	Source of motivation	A source of information about Ethiopia	Nominal
7.	Travel purpose	Reasons for visiting Ethiopia	Nominal
8.	Length of Stay	Length of stay of the tourist in Ethiopia	Nominal
9.	Season	To identify low and high season	Nominal
10.	Number of visitors	Number of tourists within a season	Nominal
11.	Occupation	Tourist Profession currently	Nominal

Table 7:-shows selected attributes for mining

After the data preprocessing was completed the final dataset used for mining had 10481 records described by eleven attributes. The attribute selected using their weight of information gain to predict the target class.

CHAPTER FOUR

4. EXPERIMENT AND ANALYSIS

As described in the methodology section WEKA data-mining tool of version 3.7 used to experiment in this study. WEKA supports input documents, Arff (Attribute relation format file), and CSV (Comma Separated Value) format. The Arff format is the standard format for WEKA input and it mainly used in the experiment. For the experimentation purpose, the .xlsx file was converted to .CV (comma delimited) format.

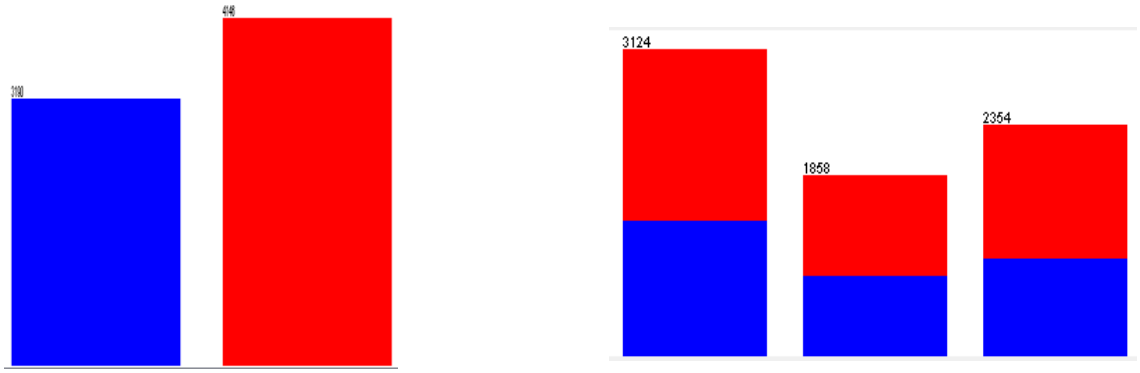
Usually two types of machine learning activities are common in tourism: - association learning and classification learning (42). In association learning, the learning method searches for associations or relationships between features of tourist preference. For example, the algorithm may try to find out if tourists who are interested to visit a historical site or natural and wildlife as well as prefer to stay for a long time in Ethiopia. That is, there is no specific target variable in this type of data mining, and so this popularly known as unsupervised learning. A second style of machine learning is classification learning. This learning scheme takes a set of classified examples from which it discovers a way of classifying unseen examples. This is a form of supervised learning, in which there is a specific target variable. For example, by using classification analysts may be interested to classify tourists into two groups: - high length of stay and low length of stay in the tourist destination area. Based on a set of demographic and other variables the classification algorithm will establish the specific attributes of a tourist that qualify them as a high or a low length of stay.

Two data mining techniques used in this study: - classification and association. Classification helps to predict the category of tourist using source of motivation, purpose of visit and other attributes. The association can be identify the relationship between the demographic patterns of the tourist and the destination areas. Therefore, useful and interesting patterns of the tourist's preference will be identify.

4.1 Issue of Class Imbalance Problems

A dataset is called imbalance if at least one of the class is represent by a significantly less number of instances than the others (29). In imbalanced data classification, the class boundary learned by the standard machine learning algorithms can be severely skew toward the positive

class. Thus, the false-negative rate can be excessively high. One major task to overcome the class imbalance problem is to resample the original training dataset, either by oversampling the minority class and/or under-sampling the majority classes until the classes represented in a more balanced way. In this research class, imbalance problem is not an issue because there is no class imbalance problem as we can see in figure-5 below.



Visitor frequency

Age group

Figure 6:- class imbalance

4.2 Building Classification Models

Four circumstances intended to build. These are J48 tree with and without pruning and Bayes naïve with display mode in old format as true and false. Trying those four scenarios with full attributes by changing the different parameters in the object editor of WEKA 3.72 was performed.

In the classification model building, measuring classifier accuracy on unseen data is a very important issue. WEKA provides three validation methods for model accuracy problems, namely; cross validation, percentage split and the distinct training and test data sets. Percentage split and splitting training and test data sets options have been use for this research.

In the percentage split option, before applying splitting, first, I used Randomize the dataset (Unsupervised→ Instance), so that random permutation created. Out of the total data, 70% (7666) records were employ in model building and the remaining 30% (3144) records used for validation set or testing.

4.3 EXPERIMENTATION AND ANALYSIS OF RESULTS

Analysis of the J48 and Bayes naïve models made in terms detailed accuracy of the classifier on the training dataset as tested on the test data based on a confusion matrix of each model result. The confusion matrix is a valuable tool for analyzing how well our classifier can recognize tuples of different classes. Confusion matrix shows four important numerical quantities (true positive, true negative, false positive and false negative).

Using the experimental result, comparison were done to each other in terms of different performance metrics values, accuracies, number of leaves, and size of the tree generated, ROC curves and execution time. The models are also compares with regard to the patterns/ knowledge discovered. The scenarios for the aforementioned classifications and experimented in this research are as listed below:

a) Decision Tree with pruning
b) Decision Tree without pruning
c) Naïve Bayes with DispalyModeInOldFormat False
d) Naïve Bayes with DispalyModeInOldFormat True

The results of each model analyzed, compared, and finally selected the best model based on the criteria of evaluation.

4.3.1 J48 experimental result analysis of Visitor_ frequency

To predict frequent visitors of tourist using WEKA 3.72 provides the options of using MinNumObj (The minimum number of instances per leaf) with default value 2 and can be flexibly changed to increase the number of leaves under a given node and minimize successive tree branching. Furthermore, it also gives several options related to tree pruning. In J48, employs two pruning methods. The first one is, known as a sub tree replacement. This means that nodes in a decision tree may be replace with a leaf reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The second type of pruning used in J48 is termed sub tree raising. In this case, a node may be move upwards towards the root of the tree, replacing other nodes along the way. Error rates used to make actual decisions about which parts of the tree to replace or raise. There are multiple ways to do this. The simplest is to reserve a portion of the training data to test on

the decision tree. Then, the reserved portion used as test data for the decision tree, helping to overcome potential over-fitting. This approach known as reduced-error pruning. In order to assess the effects of MinNumObj and the confidence for pruning and un-pruning 18 experiments were conducted.

By using the feature selection method, Nationality, Travel purpose, Duration_ of_ Stay, No_ of_ Visitors and source of motivation are selected as best predictor for Visitor_ frequent class.

Experiments	Parameters		
	Pruned	Confidence Factor	The minimum number of instances per leaf. (minNumObj)
Experiment 1	True	0.25	2
Experiment 2	True	0.25	5
Experiment 3	True	0.25	10
Experiment 4	True	0.50	2
Experiment 5	True	0.30	5
Experiment 6	True	0.30	10
Experiment 7	True	0.30	2
Experiment 8	True	0.30	5
Experiment 9	True	0.50	10
Experiment 10	False	0.25	2
Experiment 11	False	0.25	5
Experiment 12	False	0.25	10
Experiment 13	False	0.30	2
Experiment 14	False	0.30	5
Experiment 15	False	0.30	10
Experiment 16	False	0.50	2
Experiment 17	False	0.50	5
Experiment 18	False	0.50	10

Table 8:-J48 experiment trails with different setting

As it can be observe from the above table-8, the experimentation of J48 algorithm conducted in 18 different cases by changing the parameters. Each trial of the experimentation result presented below in table 9.

Experiments	UnPruned	Confidence Factor	Min Num Obj	Correctly Classified Instances (%)	Mean absolute Error	noof leaves	Size of tree	Time taken to build (sec)	AV. TP Rate	AV. FP Rate	AV. Precision	AV. Recall	AV. ROC Area
1	true	0.25	2	93.7	0.0931	243	271	0.03	0.937	0.072	0.937	0.937	0.987
2	true	0.25	5	93.5	0.0968	564	611	0.03	0.935	0.073	0.935	0.935	0.986
3	true	0.25	10	91.5	0.1185	463	489	0.02	0.915	0.086	0.916	0.915	0.978
4	true	0.3	2	93.9	0.0908	803	859	0.03	0.939	0.069	0.939	0.939	0.988
5	true	0.3	5	93.6	0.0946	574	623	0.03	0.936	0.07	0.936	0.936	0.987
6	true	0.3	10	91.7	0.1163	473	501	0.03	0.917	0.093	0.917	0.917	0.98
7	true	0.5	2	94.5	0.0778	990	1057	0.04	0.945	0.052	0.946	0.945	0.991
8	true	0.5	5	94.1	0.0833	672	727	0.03	0.941	0.055	0.942	0.941	0.99
9	true	0.5	10	91.9	0.1077	556	587	0.03	0.92	0.081	0.92	0.92	0.983
10	**False**	**0.25**	**2**	**94.8**	**0.0684**	**1253**	**1333**	**0.02**	**0.948**	**0.05**	**0.948**	**0.948**	**0.993**
11	False	0.25	5	94.1	0.0774	838	899	0.02	0.941	0.053	0.943	0.941	0.991
12	False	0.25	10	91.9	0.1032	679	644	0.03	0.922	0.076	0.923	0.922	0.984
13	False	0.3	2	94.1	0.0774	838	899	0.02	0.941	0.053	0.943	0.941	0.991
14	False	0.3	5	91.9	0.1032	679	644	0.03	0.922	0.076	0.923	0.922	0.984
15	False	0.3	10	92.0	0.1032	679	644	0.02	0.922	0.076	0.923	0.922	0.984
19	False	0.5	2	94.7	0.0684	1253	1333	0.02	0.948	0.05	0.948	0.948	0.993
17	False	0.5	5	94.1	0.0774	838	899	0.02	0.941	0.053	0.943	0.941	0.991
18	False	0.5	10	91.9	0.1015	644	679	0.02	0.92	0.08	0.921	0.92	0.983

Table 9:- J48 experiment trials result with different setting

As we have seen on the table-9 pruned has high score than un-pruned, un-pruned is usually experimented if the pruned J48 experimentation results with a small tree and leaf size that doesn't generate further rule in the form of if…then. However, the experimentation shows on the above table 8 revealed quite a complex tree size and implies no further experimentation of un-pruned J48.

Accordingly, a thorough review of the experimented results indicates trial #10 with better model performance chosen for analysis.

As we have seen below in the confusion matrix, it makes sense to observe the detailed accuracy measurement of the experiment for calculating accuracy measures and performance as presented above Class TP Rate FP Rate Precision Recall F-Measure ROC Area. The numbers of true positives in this confusion matrix are 3050 records. Those records, which predicted as 'True' class by the classifier also happened true when tested on the test data. A number of the records, which classified to the "False' class by the classifier and they are actually False as tested on the test data (True Negative Rate) are 3904. Totally, the model has an accuracy of 94.8 %.

Experiment 1: J48 pruned with confidence factor 0.25 two MinNUM of objects (default value)

This is an experiment result wich has high performance than other experiment, so it is chosen for analysis.

```
=== Run information ===
Scheme:       WEKA.classifiers.trees.J48 -U -M 2
Relation:     DATA MINING 4 P a r t 3-WEKA.filters.supervised.attribute.Discretize-Rfirst-last-
WEKA.filters.unsupervised.instance.RemovePercentage-P70.0-V-
WEKA.filters.supervised.attribute.AttributeSelection-EWEKA.attributeSelection.WrapperSubsetEval -B
WEKA.classifiers.trees.J48 -F 5 -T 0.01 -R 1 -- -C 0.25 -M 2-SWEKA.attributeSelection.BestFirst -D 1 -N 5
Instances:    7336
Attributes:   7
              Nationality
              Travel Purpose
              Duration_of_Stay
              No_of_Visitors
              Motive
              Age Group
              Travel_Frequancy
Test mode:    evaluate on training data
```

- The classifier run on test data with options.
- Internal name of the data set.
- Total instances in the arff file
- Total number of attributes selected for this class

```
=== Classifier model (full training set) ===
J48 unpruned tree
------------------
Number of Leaves  :             1253
Size of the tree :      1333
Time taken to build model: 0.03 seconds
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances        6954               94.7928 %
Incorrectly Classified Instances       382                5.2072 %
Kappa statistic                          0.8944
Mean absolute error                      0.0684
Root mean squared error                  0.1849
Relative absolute error                 13.9093 %
Root relative squared error             37.2953 %
Coverage of cases (0.95 level)          99.9046 %
Mean rel. region size (0.95 level)      58.949  %
Total Number of Instances             7336
=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.956    0.058      0.926      0.956     0.941      0.993    Frequent_Visitors
                   0.942    0.044      0.965      0.942     0.953      0.993    First_Time_Visitor
 Weighted Avg.     0.948    0.05       0.948      0.948     0.948      0.993
=== Confusion Matrix ===
    a    b   <-- classified as
 3050    140 |    a = Frequent_Visitors
  242   3904 |    b = First_Time_Visitor
```

The total number of the records which were correctly classified in true and false classes of the frequency of visitor is 6954. J48 Tree with pruning has generated model with better performance though its tree structure. It has more leaf nodes as well as it is long lengthy. However, its ability in correctly classifying records into both 'True' and 'False' classes is 94.8%, which is very good, and Its ROC area is also above 0.5, which is the minimum possible acceptable value for ROC curve. If drawn the ROC Curve 0.993 is above the diagonal. ROC area is plotted from True positive Rate (TPR) on the y axis against the False Positive Rate (FPR) on the x-axis. With 0.993 of ROC values, the model has a very good and hopeful accuracy.

J48 experimental result analysis of source of motivation

trials	Un-Pruned	Confidence Factor	Min Num Obj	Correctly Classified Instances (%)	Mean absolute Error	nos of leaves	Size of tree	Time taken to build(sec)	AV. TP Rate	AV. FP Rate	AV. Precision	AV. Recall	AV. ROC Area
1	False	0.25	2	99.4	0.0052	267	279	0.02	0.995	0.003	0.995	0.995	1
2	False	0.25	5	99.3	0.0062	261	272	0.43	0.993	0.003	0.993	0.993	1
3	False	0.25	10	99.3	0.0078	249	257	0.02	0.993	0.004	0.993	0.993	0.999
4	**False**	**0.3**	**2**	**99.5**	**0.0052**	**267**	**279**	**0.02**	**0.995**	**0.003**	**0.995**	**0.995**	**1**
5	False	0.3	5	99.3	0.0062	261	272	0.02	0.993	0.003	0.9963	0.993	1
6	False	0.3	10	99.2	0.0078	249	257	0.02	0.993	0.004	0.993	0.993	0.999
7	**False**	**0.5**	**2**	**99.5**	**0.0052**	**267**	**279**	**0.02**	**0.995**	**0.003**	**0.995**	**0.995**	**1**
8	False	0.5	5	99.3	0.0062	261	272	0.02	0.993	0.003	0.993	0.993	1
9	False	0.5	10	99.2	0.0078	249	257	0.02	0.993	0.004	0.993	0.993	0.999

Table 10:- J48 experimental result analysis of motive

As we can see from the table, trail number 4 and 7 are equal with the different confidence factors, so trail number 4 with 0.3 confidence factor taken for analysis.

```
Number of Leaves  : 267
Size of the tree :      279
Time taken to build model: 0.02 seconds
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances        7312               99.6728 %
Incorrectly Classified Instances        24               0.3272 %
Kappa statistic                          0.995
Mean absolute error                      0.0037
Root mean squared error                  0.0429
Relative absolute error                  0.849  %
Root relative squared error              9.2144 %
Coverage of cases (0.95 level)          99.8637 %
Mean rel. region size (0.95 level)      33.5969 %
Total Number of Instances             7336
=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.997     0.001      0.998     0.997     0.998      1        Cultural_Life
                 0.995     0.001      0.997     0.995     0.996      1        Natural_and_Wildlife
```

0.997 0.002 0.995 0.997 0.996 1 History
Weighted Avg. 0.997 0.002 0.997 0.997 0.997 1
=== Confusion Matrix ===
a b c <-- classified as
3115 6 3 | a = Cultural_Life
0 1849 9 | b = Natural_and_Wildlife
6 0 2348| c = History
=== Re-evaluation on test set ===

J48 experimental result analysis of Length of stay

trials	Un-Pruned	Confidence Factor	Min Num Obj	Correctly Classified Instances (%)	Mean absolute Error	No of leaves	Size of tree	Time taken to build (sec)	AV. TP Rate	AV. FP Rate	AV. Precision	AV. Recall	AV. ROC Area
1	False	0.25	2	96	0.031	349	381	0.04	0.961	0.022	0.958	0.961	0.991
2	False	0.25	5	95.9	0.0316	328	355	0.04	0.959	0.022	0.957	0.959	0.991
3	False	0.25	10	95.2	0.036	166	187	0.04	0.04	0.022	0.95	0.952	0.991
4	**False**	**0.3**	**2**	**96.3**	**0.0293**	**506**	**545**	**0.06**	**0.963**	**0.021**	**0.961**	**0.963**	**0.992**
5	False	0.3	5	96.1	0.0302	479	509	0.04	0.962	0.021	0.959	0.962	0.992
6	False	0.3	10	95.4	0.0346	317	341	0.03	0.955	0.02	0.953	0.955	0.991

Table 11:- J48 experimental result analysis of Length of stay

4.3.2 Naive Bayes/Bayesian Classifiers

As we have discussed in the literature review one of the advantages of Naive Bayes is, it requires a small amount of training data to estimate the parameters. It assumes that the effect of an attribute value of a given class is independent of the values of the other attributes. This assumption called class conditional independence. The same percentage as J48 of the data set for training and test used for Naïve Bayes to building model.

A Bayesian classifier is, based on the idea that the role of a (natural) class is to predict the values of features for members of that class. Examples are grouped in classes, because they have common values of the features. Such classes are often called natural kinds.

The object editor under naïve Bayesian has fundamental parameters such as DisplayModeInOldFormat. This parameter is uses depending on the number of classes and attributes we have. The old format is better when there are many class values and the new format is ideal when there are fewer class and many attributes.

Naïve Bayes was experimented into two states, which is a Naïve Bayes with DisplayModelInOldFormat False and DisplayModelInOldFormat True.

Experimentation of Naïve Bayes for Visitor Frequency

	DisplayModelInOldFormat	
	True	False
Correctly classified	86.873	86.873
Time taken to build	0.01	0.01
Mean absolute error	0.2113	0.2113
AV.TP.Rate	0.869	0.869
AV.FP.Rate	0.141	0.141
AV.Precision	0.869	0.869
AV.Recall	0.869	0.869
AV.ROC.Arae	0.922	0.922

Table 12:-Experimentation result of Naïve Bayes for frequent visits

```
Time taken to build model: 0.01 seconds
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances        6373               86.873  %
Incorrectly Classified Instances       963               13.127  %
Kappa statistic                          0.7316
Mean absolute error                      0.2113
Root mean squared error                  0.3257
Relative absolute error                 42.99   %
Root relative squared error             65.7068 %
Coverage of cases (0.95 level)          98.8686 %
Mean rel. region size (0.95 level)      84.2966 %
Total Number of Instances            7336
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
0.827     0.099      0.865     0.827     0.846      0.922    Frequent_Visitors
 0.901     0.173      0.871     0.901     0.886      0.922    First_Time_Visitor
Weighted Avg.0.869     0.141      0.869     0.869     0.868      0.922
=== Confusion Matrix ===
a    b   <-- classified as
2638  552 |    a = Frequent_Visitors
 411 3735 |    b = First_Time_Visitor
```

Experimentation of Naïve Bayes for length of Stay

	DisplayModelInOldFormat	
	True	False
Correctly classified	91.7939	91.7939
Time taken to build	0	0
Mean absolute error	0.0587	0.0587
AV.TP.Rate	0.918	0.918
AV.FP.Rate	0.052	0.052
AV.Precision	0.906	0.906
AV.Recall	0.918	0.918
AV.ROC.Arae	0.982	0.982

Table 13:- Experimentation result of Naïve Bayes for length of Stay

As we have seen in on the above tables-13 and 12, DisplayModelInOldFormat with the value of True and False is equal. So no need an experiment with both true and false.

=== Re-evaluation on test set ===

```
User supplied test set
Relation:     DATA MINING-WEKA.filters.supervised.attribute.Discretize-Rfirst-last-
WEKA.filters.unsupervised.instance.RemovePercentage-P30.0-V
Instances:     unknown (yet). Reading incrementally
Attributes:   9
=== Summary ===
Correctly Classified Instances        2886               91.7939 %
Incorrectly Classified Instances       258                8.2061 %
Kappa statistic                          0.849
Mean absolute error                      0.0587
Root mean squared error                  0.1848
Coverage of cases (0.95 level)          97.9326 %
Total Number of Instances             3144
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.967     0.077      0.891     0.967     0.927      0.983  LoS_Very_Ex
                 0.292     0.019      0.545     0.292     0.38       0.906    LoS_Ex
                 0.966     0.038      0.967     0.966     0.966      0.991    LoS_Med
                 0.902     0.029      0.9       0.909     0.902      0.962    LoS_Short
   Weighted Avg 0.918     0.052      0.906     0.918     0.909      0.982

=== Confusion Matrix ===
a    b    c    d   <-- classified as
 1196    0   41    0 |    a = LoS_Very_Ex
  145   66   15    0 |    b = LoS_Ex
    2   55 1624    0 |    c = LoS_Med
    0    0   23   64 |    d = LoS_Short
```

Source of Motivation	
	DisplayModelInOldFormat= False
Correctly classified	95.3926
Time taken to build	0
Mean absolute error	0.0403
AV.TP.Rate	0.954
AV.FP.Rate	0.022
AV.Precision	0.954
AV.Recall	0.954
AV.ROC.Arae	0.989

Table 14:- Experimentation result of Naïve Bayes for Source of Motive

In generated the result of Naïve Bayes by changing the parameters of the display mode in old and new formats are the same. Except that there is a difference in the time taken to build the model; both the formats have the same learning capacity for the datasets trained and tested. So this imposed comparing the result of J48 and Naïve Bayes.

4.3.3 Comparison of J48 and Naïve Bayes models

Comparison of the two models is made in terms of detailed accuracy of class looking general model accuracy, such as the precision, ROC Area, recall and the rules generated for interpretation. The following table 15 gives the relative comparison between the two models.

Sno	Classifier	Class Model build for	Correctly Classified Instances(%)	Mean absolute Error	Time taken tobuild (sec)	AV. TP Rate	AV. FP Rate	AV. Precision	AV. Recall	AV. ROC Area
1	Naïve Bayes	Visitor frequency	86.873	0.2113	0.01	0.869	0.141	0.869	0.869	0.922
	J48	Visitor frequency	94	0.0684	0.02	0.948	0.05	0.948	0.948	0.993
2	Naïve Bayes	Source of Motivation	95.3926	0.0403	0	0.954	0.022	0.954	0.954	0.989
	J48	Source of Motivation	99.5	0.0052	0.02	0.995	0.003	0.995	0.995	1
3	Naïve Bayes	Length of stay	91.7939	0.0587	0	0.918	0.052	0.906	0.918	0.982
	J48	Length of stay	96.3	0.0293	0.06	0.963	0.021	0.961	0.963	0.992
4	Naïve Bayes	Tourist destination by region	81.1069	0.0546	0	0.811	0.05	0.82	0.811	0.975
	J48	Tourist destination by region	100	0.0028	0.02	1	0	1	1	1

Table 15:- Comparison of J48 and Naïve Bayes models

The above Table-15 shows there is a relative good model prediction in all J48 and naïve-Bayes. The ROC Area for Naïve Bayes on class model "Visitor frequency "indicates 0.922. It is almost the same with j48 0.993 .This signifies the number of correctly classified datasets are higher in the model built by J48 than Naïve-Bayes.

Naive Bayes has a better prediction if the attributes are conditionally independent to each other. For the given data under study J48 has shown better accuracy and the rules generated by this model used for interpretation.

4.3.4 Comparison between classifiers' and experts' judgments

The expert and classifier judgments may differ in predicting a record to a certain class label. This shows that the classifier to another class may label the record that labeled by expert to one class. This kind of problem can be reducing the performance of the system. Therefore, to evaluate the performance of the system in terms of how correctly the model classifies records into different labeled classes is more important. When we see the result under class model "Visitors_ Frequency", 94.8% are correctly classified, it is a good result, but it shows 5.2 %

exits unclassified of the total records. Therefore, it is important to investigate why the 5.2% of the total misclassified.

The classifier predicts the records into a certain class as there are similar attributes that lie in the same class boundary. However, the attribute that determines the class boundary of the given record is hide because of the attribute similarity trend applied by the classifiers.

In this study, the expert and classifier vary in classifying a certain records, when we see the tables 16, 17, and 18 below it show a record that classifier and expert judgments variation.

To compare the expert and the classifier, sample record taken from the J48 output. The records intentionally selected from the output with majority similar value of the attribute to show the significance of the remaining attributes during classification. Five attributes four of them has equal value, but they are different in one attribute which is "Travel_ Purpose" were selected.

The algorithm looks for similarity of a majority of the attributes. Age group 2, medium number of visitor's, nationality USA and business travel purpose is correctly classified with both the expert and the classifier as frequent visitors, but purpose of the visit = conference is not the same the classifier and expert. The model takes the similarity of all the attributes disregarding the difference in the travel purpose, and the j48 algorithm classifies both to be frequent visitors.

As the classifier, age group_3 with extended duration of stay (above 15 days), high _ Number of _Visitors and, purpose of visit = business come from Eastern Europe are frequent visitors, but as the expert judges they are not frequent visitors. This problem is repeatedly shows in all the tables; 16, 17 and 18 that taken from the experimental result.

Therefore, we can say that the misclassification occurred from the underestimating of a single attribute's value taking the similarity of the other attributes as the predominant predictive values.

Age group	Number of visitors	Nationality	Travel Purpose	Frequent_Visitors	
				Expert	Classifier
Age Group_2	Medium _No_Visitors	USA	Business	Yes	Yes
Age Group_2	Medium _No_Visitors	USA	Conference	No	Yes
Age Group_2	Medium _No_Visitors	USA	Leisure_and_Holiday	No	Yes
Age Group_2	Medium _No_Visitors	USA	Archeology	Yes	Yes
Age Group_2	Medium _No_Visitors	USA	Transit	Yes	No
Age Group_2	Medium _No_Visitors	USA	Visiting_Relatives_ &_ Friends	Yes	No

Table 16:- Comparison result between classifiers' and experts' judgments

Age group	Duration of Stay	Number of visitors	Travel Purpose	Nationality	Frequent _ Visitors	
					Expert	Classifier
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Conference	USA	Yes	yes
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Conference	Australia	No	Yes
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Conference	China	Yes	Yes
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Conference	East Africa	Yes	No
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Conference	East Asia & The Pacific	No	No
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Conference	East Europe	Yes	Yes

Table 17:- Comparison result between classifiers' and experts' judgments

Age group	Duration of Stay	Number of visitors	Travel Purpose	Nationality	Frequent_Visitors	
					Expert	Classifier
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Business	USA	Yes	Yes
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Business	Eretria	No	No
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Business	East Europe	Yes	No
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Business	Other Oceania	No	No
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Business	Philippines	No	No
Age Group_3	Extended duration of stay	High_Nomber of_Visitors	Business	Republic of Korea:	No	No

Table 18:- Comparison result between classifiers' and experts' judgments

Some Rules generated by J48 for class Visitors frequency

What makes the J48 algorithm a choice of data mining practitioners is, that it provides rules in the form of if….then, which is easy to understand and be familiar with most important rules, obtained. Under this study, results generated by J48 algorithm and that are interested in the classification of the records are presented below for discussion.

1) If Age between 26 and 45 and visitors are coming from USA and Archeology then frequent visitors.
2) If Age between 26 and 45 and visitors are coming from china then first time visitors.
3) If Age greater than 46 and very extended stay and for business purpose, thena first time visitor.
4) If Age greater than 46 and very extended stay and for leisure and holiday purpose and USA frequent visitors

5) If Age greater than 46 and very extended stay and for leisure and holiday purpose and east Africa then frequent visitors
6) If Age greater than 46 and very extended stay and for leisure and holiday purpose and East Asia & the Pacific then frequent visitors.
7) If Age greater than 46 and very extended stay and for leisure and holiday purpose and Europe: then Frequent_Visitors
8) Age group less than and 25 and nationality Australia and purpose of visit leisure _and_Holiday and Visiting_Relatives_ &_ Friends then first time visitors.

4.3.5 Discussions of Results generated by j48 with different classes

In order to reach a common conclusion form about the significance of the above rules and the attributes used to create those rules, the relation of the attributes with the predicted class predicted by the rules evaluated based on suggestions offered by domain experts and results of previous research works. As we can see from the rules, the model with class "visitor frequency" has generated predictions for first time visitors and frequent visitors. The model shows that the majority of the tourist comes to different regions of Ethiopia are first time visitors. However, there are some cases that tourist enables to visit frequently. As the model result shows, tourists 'age between 26 and 45 from USA, East Asia & the Pacific are frequent visitors. In addition, tourists coming for the purpose of leisure and holiday are frequent visitors.

Tourist visited Ethiopia have different source of motivation. As the source of motivation class models, shows, source of motivation is a key factor for length of stay and destination preference. Tourist spent for more than 15 days and coming for the purpose of Leisure_ and_ Holiday and Visiting_ Relatives_ &_ Friends was interested to visit the cultural life of Ethiopia. A tourist spent 4 to 7 days was interested in natural and wildlife, but a tourist spent a short time of period was visited historical site. Domain experts agree with the result; that is foreign tourist need tangible promotion through different media. Currently there are only few dedicated tourism-Medias, in Ethiopia, the Medias does not emphasize on Ethiopian history. Domain experts further stated that such as coffee and endemic animals promoted than history. In addition, Ethiopian Diasporas are a good promoter of cultural life. Therefore, the result of study has also confirmed that cultural life and nature and wild lives are attracting more tourists to Ethiopia than history.

Tourists from the USA, Australia, Belgium, Canada, Central Africa, and the Middle East are more interested in the cultural life of Ethiopia. As well as tourist interested in cultural life, visits

Addis Ababa, Amhara, Debub and Oromia. Moreover, for Nature and wildlife purpose the Tourist prefers Debub and Oromia.

Central Africa, USA, and Europe are potential tourist for the purpose of conference. Domain experts also agreed with this output, they stated that conference in Ethiopia grows because of the African Union, which is in Addis Ababa.

Potential visitors of Ethiopia are identified as age above 45, from USA, Middle East, central Europe, and Africa. Africa comes as one of the potential source of visitors to Ethiopia in cultural life and conference.

4.3.6 Apriori / Association Rule Experiment Result

Association rules are a data mining technique that seeks to find frequent connections between attributes in a data set (60). The objective of using the association rule is to detect the relationship between the attributes. It helps to identify the possibility of visiting the destination areas with one route package. This may help to discover new route packages.

The experiment was done with default value of the Apriori sitting in WEKA except the confidence metric which changes to 0.5 and the number of rules =100. This was done because 100 rules with minimum confidence of 50 % will be generated. The assumption of making 100 rules is, the algorithm might be generating interesting rules more from the dataset.

Some rules generating by Aporiari

1) Duration_of_Stay=LoS_Very_Ex Region _Visited=Addis_Ababa 2104 ==> Motive=Cultural_Life 2104 <conf:(1)> lift:(2.49) lev:(0.12) [1259] conv:(1259.99)
2) Duration_of_Stay=LoS_Med Travel_Frequancy=Frequent_Visitors 2767 ==> Age Group=Group_3 2757 <conf:(1)> lift:(2.07) lev:(0.14) [1426] conv:(130.57)
3) Age Group=Group_2 3082 ==> Travel_Frequancy=First_Time_Visitor 2774 <conf:(0.9)> lift:(1.69) lev:(0.11) [1128] conv:(4.65)
4) Motive=Natural_and_Wildlife 3095 ==> Duration_of_Stay=LoS_Med 2690 <conf:(0.87)> lift:(1.55) lev:(0.09) [956] conv:(3.35)
5) Duration_of_Stay=LoS_Med Age Group=Group_3 3589 ==> Travel_Frequancy=Frequent_Visitors 2757<conf:(0.77)> lift:(1.65) lev:(0.1) [1084] conv:(2.3)

```
6) Travel_Frequancy=Frequent_Visitors Age Group=Group_3 3702 ==>
   Duration_of_Stay=LoS_Med 2757 <conf:(0.74)> lift:(1.33) lev:(0.07) [683]
   conv:(1.72)
7) Age Group=Group_3 5040 ==> Travel_Frequancy=Frequent_Visitors 3702
   <conf:(0.73)> lift:(1.58) lev:(0.13) [1352] conv:(2.01)
8) Travel_Frequancy=First_Time_Visitor Motive=Cultural_Life 2244 ==> Region
   _Visited=Addis_Ababa 1554 <conf:(0.69)> lift:(1.78) lev:(0.06) [678] conv:(1.98)
9) Duration_of_Stay=LoS_Very_Ex Age Group=Group_1 1555 ==> Region
   _Visited=Addis_Ababa 1076 <conf:(0.69)> lift:(1.77) lev:(0.04) [469] conv:(1.98)
```

As we have seen in the above result, the Apriari algorithm generates interesting relationships.

4.3.7 Discussions of Results generated by Apriori

Tourists who visit Addis Ababa stays for more than 15 days are more interested in cultural Life. The domain experts agree with this result, they explain this is due to the city (Addis Ababa) have different cultural center, nation and nationality music and dance show, ceremonies and cultural food and drink. That is why most of the tourists are interested to visit the different cultural life of Ethiopia within a city. In addition, the experiment shows that 97 % of extended length of stay was record due to Cultural Life.

Age above 45 visitors have 76 % possibility to come again as a repeated visitor to Ethiopia. As well, as age between 25 and 45 of first time visitors have 71% possibility to stay one week.

In tourism market, there are pick season and low seasons. Peak season means when the vacationer flow is increased, and low season is a time interval were number of tourists flow get decrease. This study shows, season 4 (October, November, and December) is a peak season were tourist flow increases to visit cultural life. In this season tourist have 72% possibility to stay more than 15 days.

Season _1 (January, February, and March) also a second peak season, where tourists flow increased to visit nature and wildlife. The domain expert also agrees with this result. The expert further explains that in Ethiopia only the rainy season is the low season. The other seasons are peak seasons, this is because starting from celebrating the Ethiopian New Year, “MESKEL” (the finding of the true cross), and “GENA” (x-mass) and “TIMIKET” (epiphany) are more tourists preferred events.

Small numbers of first time visitors' are interested in history and stay for a short period, so this shows tourist, which visited historical sites, is decreasing. Accordingly, this might need further research.

4.4 Experiment summary

Number of researchers conducted studies to assess the application of data mining in the different sectors like Airlines, Banking, HealthCare, and Customs. The main intension of all researchers is to investigate how much data mining tools and techniques are help to predicte customers' preference in the above-mentioned sectors. Most researchers used clustering and classification techniques with k-Means and decision tree algorithms.

Even though all researchers focused on predicting customer preference, identifying the association between the customers and environment was forgotten. This study uses association rule algorithm to detect the possibility of visiting Ethiopia once more. This had an advantage in identify the most determinant factors that enable the tourist to decide which destination or in what situation to be visit.

One practical yet important problems have not been resolve by the data miners in experimental tourism research (22). That is lack of using enough data for training samples and testing. Many data mining methods may not achieve satisfactory performance if learned on small data sets. However, this study use total 442,011 tourists' records in experiment and it over-comes the problem that rise because of using small data.

The result obtained in this research has proved the association and classification techniques of data mining helps tourism industry for detecting and predicting tourist preferences. Moreover, it achieved the objective of the study.

CHAPTER FIVE

5. CONCLUSION AND RECOMMENDATIONS

In this study, we have been discussed the different types of machine learning techniques and explain how they have been analyze data related to tourism. Two types of machine learning activities were used: - association learning and classification learning. In association learning we find that the possibility of visiting Ethiopia frequently. We find out tourists interested in cultural life prefer to stay for long period. A second technique of machine learning which used in this study is classification learning. This learning scheme takes a set of classified pattern from which it discovers a way of classifying unseen patterns. By using classification analysts, we classify tourists into two groups—first time visitor and frequent visitor.

The objective of this study was to explore the tourist preference in the Ministry of culture and tourism of Ethiopia and by identifying the key determinant factor of tourist that could help in to maximize the tourist stay, maximize expenditure, create new packages route, and maintain tourist flow throughout the year. The study has also sought to know what data mining algorithms and models are more suitable for predicting tourist preference.

In addition, the study required an answer for, which tourist frequently returns to the same tourist attraction as repeated visitors, this helps to update the service providers of hospitality industry such as hotel, tour and travel agency to accommodate frequent tourists.

5.1 CONCLUSION

10481 records with eleven attributes used in experiments to create models with different classification class. In addition, Apriori algorithm of association used to generate rule. Wrapper evaluator used to feature selection in model building for each class.

Models built by implementing the J48 decision tree and Naïve Byes classifiers. Experimental results show that J48 decision tree classifier using the pruned technique with default confidence factor of 0.25 and minimum number of instances per leaf 2, performed best with accuracy of 94%.

Both decision tree and rule induction approach result is an encouraging output, especially simplicity of the output of both techniques to explain to the end user. The application of other data mining techniques like neural networks and Bayes techniques can be also apply and tested if they could be more applicable to this problem domain.

While, J48 classifier achieved promising and interesting results, some records misclassified due to the underestimating of a single attribute's value taking the similarity of the other attributes as the predominant predictive values or we can say that the data driven classification trend of the algorithm.

The tourist preferences are a very complex problem domain. The effects or contributions of different component such as shopping habit of the tourists should be see and study additionally to observe and understand its effects.

In this study variables like travel purpose, source of motivation, Nationality, and age found to be important variables to predict the tourist preference.

The results obtained in this research work have proved that the main motivation of tourist to visit Ethiopia is cultural life.

Addis Ababa identified as a center of cultural life and more extended stay attain in Addis Ababa, this shows different cultural center within a city is more preferable by tourists.

Ages above 45 visitors are more interested in culture, leisure, and holiday. Tourists above this age, have76 % possibility to visit Ethiopia once more. Nature and wildlife also has encouraging result next to the leisure and holiday. The least favorite of travel purpose is history.

The study shows that the majority of the tourist comes to Ethiopia are first time visitors. Especially age below 46 classified as a first time visitor. Tourists from USA, East Asia, & the Pacific classified as frequent visitors.

Tourist from USA, Australia, Belgium, Canada, Central Africa, and the Middle East are more interested in the cultural life of Ethiopia. Africa comes as a potential visitor for the purpose of conference.

5.2 RECOMMENDATIONS

Encouraging results obtained by employing both decision tree and association rule techniques as generated by J48, naïve Bayes and Apriori algorithms. The outputs of the algorithms are much understandable to explain the predicted outcome easily.

- The result obtained in this research has proved the advantage of using data mining to predicting tourist preferences. More specifically, it is supportive to construct, evaluate, and update policies and strategies.
- The records used in this study taken from the operational data of the Ministry of Culture and Tourism. However, substantial data needed for data mining tasks, further researches that integrate the operational and nonoperational data would rather come up with interesting results.
- The Ministry of Culture and Tourism, statistics directorate keeps tourist profile in different format such as in flat form, MS-word, and MS-excel, which is being difficult and time consuming. So it would be important to have a centralized data management through the distributed system so that raw data representing the Ministry of Culture and Tourism. Moreover, analysis would be representative of the Ministry of Culture and Tourism.
- It observed that, there are inconsistencies in recording tourists' data caused by lack of data integrity constraint and resulting in the data preprocessing taking extended time. It is recommend that the database redesign and take into account all the problems.
- Visiting historical sites become decreasing in number of tourists and duration of stay in the historical sites. It is recommend conducting studies on the historical motivation and promotional advertisements concerning the historical sites. Because of Ethiopia have a number of ancient historical destinations that will be visit by tourists.
- Conference become an emerging travel purpose which shows encouraging result in attracting tourist to Ethiopia but, with short duration of stay. Therefore, it is important to do research on how to increase their length of stay.
- The available tour packages are the same route traditionally the north, Debub, eastern. However, as the research shows us, new tour packages mainly focused on cultural life, nature and wildlife is important.

- Such as nationality, travel purpose and age found to be important variables to predict the tourist preference. Therefore, the ministry of culture and tourism and Tour and Travel Organization should focus their attention on these factors during promotional, advertising activities and construction of rules and policies.

5.3 References

1) A. Salguero, F. Araque, R. Carrasco, A. de Campos and L. Martínez 2010. Using Fuzzy Data Mining for finding preferences in adventure tourism.
2) Adrian Popescu, Gregory Grefenstette and Pierre Moëllic. 2008. Mining Tourist Information from User-Supplied Collections.
3) Alan Rea. 1995. *Data mining* and knowledge discovery in real-world *applications* of knowledge discovery London.
4) Alistair Williams. 2006. Tourism and hospitality marketing fantasy, feeling and fun, International Journal of Contemporary Hospitality Management, vol 18.
5) Altidor W. Khoshgoftaar T.M., and Van Hulse J. 2009. An Empirical Study on Wrapper Based Feature Ranking, 21st International Conference On tools With Artificial Intelligence, pp.75-82, IEEE.
6) Anagaw S. 2002. Application of data mining technology to predict child mortality patterns: the case of butajira rural health project (brhp). Master's thesis. Addiss Ababa University,Addis Ababa,Ethiopia.
7) Annika Hinze and Saijai Junmanee, 2007.Travel Recommendations in a Mobile Tourist Information System.
8) Anshul Goyal and Rajni Mehta. 2012. Performance Comparison of Naïve Bayes and J48 Classification Algorithms, IJAER, Vol. 7, No.11,
9) Bigus J. 2001. Data Mining with Neural Networks: Solving Business Problems- from Application Development to Decision Support, McGraw-Hill: New York.
10) Bigus JP. 2006. Data mining with Neural Networks: Solving Business problems from application Development to decision Support. McGraw-Hill, New York.
11) Brialle Attilio and Reiseneine Kunst.1997. Vom Beginn des modernen Tourismus/ Grand Tour /.
12) Buhalis, D. 2000. Tourism in an Era of Information Technology. In B. Faulkner, G. Moscardo and E. Laws (Eds.) Tourism in the Twenty-first Century: Lessons from xperience,London: Continuum.Rasmussen and K. Williams. 2006. Gaussian Processes for Machine Learning. MIT.
13) Cabena P. 2000. Discovering Data Mining - From concept to Implementation, Printice Hall, and New Jersy.

14) Cabena P. 2008. Discovering Data Mining: From concept to Implementation, Printice Hall, and New Jersy.

15) Cabena, P., Hadjinian, Stadler, R., Verhees, and Zanasi, A. 1997. Discovering Data Mining: From Concept to Implementation.Prentice Hall.

16) Chapman P. 1999. CRISP-DM: Step-by-step data mining guide SPSS Inc.

17) Chen, H and Cooper M. 2001, Using Clustering Techniques to Detect Usage Patterns in a Web-Based Information System. Journal of the American Society for Information Science and Technology, Vol.52

18) Chiang WY. 2012. Applying a New Model of Customer Value on International Air Passengers' Market in Taiwan. International Journal of Tourism Research.

19) Cios, Pedrycz , Swiniarski, and Kurgan. 2007. Data mining a knowledge discovery approach.

20) Crompton JL.1979. Motivations for pleasure vacations. Annals of Tourism Research.

21) Crouch, G. I.; Mazanec, J. A.; Oppermann, M.; Sakai, M. Y.Consumer psychology of tourism, hospitality and leisure. (pp 177-191).UK: CABI.

22) D.Lavanya1 and K. Rani. 2013. A Hybrid Approach to Improve Classification. International Journal of Application or Innovation in Engineering & Management (IJAIEM), Vol 2, No 1.

23) David Carson and Mário Ascençãoc. 2007. Sustainable tourism marketing at a World Heritage site.

24) David Hand. 2011. Principles of Data Mining. MIT Press.

25) Dhingra N. 2002. Screening Donated Blood for Transfusion- Transmissible Infections:World Health Organization available at: < http://www. who.int/bloodsafety/ makingsafe bloodavailable in africastatement.pdf>.

26) Dipali Bhosale and Patil. 2014. Feature Selection based Classification using Naïve Bayes, J48 and Support Vector Machine. International Journal of Computer Applications Vol 99 no.16.

27) Echtner, C. M., & Ritchie, J. R. B. (1991). The meaning and measurement of destination image.The Journal of Tourism Studies.

28) Fayyad, Piatetsky- Shapiro, Smyth and Uthurusamy, Advances in Knowledge Discovery and Data Mining. Cambridge, MIT Press/AAAI Press, 1996.

29) Fayyad U, Piatetsky-Shapiro G , Smyth P and Uthurusamy R. 1996. Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, Menlo Park.

30) Fisher R.A. 1996. The Design of Experiments, Edinburgh: Oliver and Boyd.

31) G. Manikandan, N. Sairam, C. Saranya and S. Jayashree . 2013. A Hybrid Privacy Preserving Approach in Data Mining.School Of Computing, SASTRA University, India.

32) G. Shaw and M. Williams. 2009. Knowledge Transfer and Management in Tourism Organizations: An emerging research agenda.

33) Gyr, Ueli. 2010. The history of tourism: Structures on the path to modernity. European History Online (EHO). Retrieved from http://ieg-ego.eu/en/threads/europe-on-the-road/the-history-of-tourism.

34) Haftom Gebregziabher. 2011. Application of data mining technology in predicting the seroprevalence of hbv,hcv,hiv; the case of the national blood bank of addis ababa. A thesis submitted to the school of graduate studies of Addis Ababa University. Addis Ababa.

35) Han J. and Kamber M. 2001. Data Mining: Concepts and Techniques. Morgan Kauf-mann Publishers.

36) Han, J. and M. Kamber. 2011. Estimating joint default probability by efficient importance sampling with applications from bottom up.

37) Han, Jiawei and Kamber. 2001. Data Mining: concepts and Techniques. San Fransisco; Morgan Kufman Publishers.

38) Helen T. 2003. Application of data mining technology to identify significant patterns in census or survey data. Masters Thesis Addis Ababa University, Addis Ababa, Ethiopia.

39) Huang, J., Huang, C. & Wu S. 1996. National character and response to unsatisfactory hotel service. International Journal of Hospitality Management

40) Hyde KF. 2000. A hedonic perspective on independent vacation planning, decision-making and behavior.

41) Indranil Bose. 2009. Data Mining in Tourism:The University of Hong Kong, Hong Kong,

42) Indranil Bose. 2009. Managing a Big Data Project: The Case of Ramco.

43) J. Van Hulse M. Khoshgoftaar, and A. Napolitano, 2007. Experimental perspectives on learning from imbalanced data. in Proc. 24th Int. Conf. Mach. Learn. , Corvallis.

44) Japkowicz N. 2000. The Class Imbalance Problem: Significance and Strategies. International Conference on Artificial Intelligence (IC-AI), Las Vegas, Nevada.

45) Jiawei Han and Micheline Kamber. 2006. Data Mining Concepts and Techniques, Second Edition. Elsevier Inc.

46) Jiong Yan. 2008. Cultural and attitudinal influences on destination choice: preferences of Chinese domestic tourists, Dissertation to at the Faculty of Agricultural Sciences, Georg-August-University Göttingen, Germany.

47) Jiong Yan. 2008. Cultural and attitudinal influences on destination choice: preferences of Chinese domestic tourists, Dissertation. Georg-August-University Göttingen, Germany.

48) Joel P. Lucas a, Nuno Luz b, and María N. 2000. A hybrid recommendation approach for a tourism system.

49) Kamani, Samarasinghe, Saluka, Roshan and D. Yapa. 2013. Data Mining and Service Customization in Leisure and Hospitality. International Journal of Soft Computing and Engineering (IJSCE) vol 3 no 5.

50) Kapil Sharma, Sheveta Vashisht, Heena Sharma, Richa Dhiman, and Jasreena Kaur Bains .2013. A Hybrid Approach Based On Association Rule: Mining and Rule Induction in Data Mining. International Journal of Soft Computing and Engineering (IJSCE) Vol3 no.1.

51) Khalid Sheikh. 2003. Advanced Data Mining and Applications: Second International Conference, ADMA. London.

52) King, Brian and Hyde, G. 1989. Tourism Marketing in Australia, Hospitality Press, Melbourne.

53) Kristina Ba and Ulf Johanssonb. 2012. Creating and consuming experiences in retail store environments: Comparing retailer and consumer perspectives, Lund University, Sweden Journal of Retailing and Consumer Services.

54) Lansing P. and Vries P. 2007. Sustainable tourism: Ethical alternative or marketing ploy.

55) Leila Etaati and David Sundaram. 2014. Adaptive tourist recommendation system: conceptual frameworks and implementations.

56) M Danubianu, S Pentiuc, O Schipor, and M Nestor. 2008. Distributed intelligent system for personalized therapy of speech disorders. Ungureanu Computing in the Global Information Technology. The Third Edition ICCGI.

57) Margaret H., Danham and Sridhar. 2006. Data mining, Introductory and Advanced Topics, Person education.

58) Masud Karim, Rashedur M. and Rahman. 2013. Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing, Journal of Software Engineering and Applications.

59) Matthew North. 2012. Data Mining for the Masses. A Global Text Project Book. Utah State University

60) Michael Hall and Williams Allan.2006. Product improvement or innovation: what is the key to success in tourism', in innovation and Growth.

61) Michael J and Linoff. 2011. Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. Wiley Publishing.

62) Mitchell and Coles. 2006. Tourism business and the local economy: Increasing Impact through a linkage approach. ODI Briefing Paper, March, London.

63) Moct Bulletin.2013.tourism Statistics 2008-2012 of Ethiopia. Addis Ababa, Ethiopia.

64) Monali Gandhi, Khushali Mistry and Mukesh Patel. 2014. A Modified Approach towards tourism Recommendation System with Collaborative Filtering and Association Rule Mining, International Journal of Computer Applications .Volume 91 – No.6.

65) N Morgan and A Pritchard. 2002. Destination branding: creating the unique destination proposition.

66) N Morgan. 1998. Tourism promotion and power: creating images, creating identities tourism promotion and power: creating images and creating identities.

67) Ngay J. 2002. On Discriminative vs. Generative Classifiers: A comparison of Logistic Regression and Naive Bayes, Neural Information Processing Systems.

68) Olsen M and Connolly. 1999. Research on information technology in the hospitality industry. International Journal of Hospitality Management, 23(5):473-484,2004

69) Opaschowski H. 2001. Tourism in 21 century. Hamburg GmbH.

70) Pairaya J., Sarawut S. and Sukree S. 2013. Applying Data Mining to Analyze Travel Pattern in Searching Travel Destination Choices. The International Journal of Engineering and Science (IJES) vol 2 no 4.

71) Pankhurst. 1990. Resettlement Policy and Practice in Ethiopia: Options for Rural Development. Zed Books London.

72) Philip Kotler. 2000. Marketing Management, Millennium Edition.

73) Poon, A. (1993). “Tourism, Technology and Competitive Strategies”. New York: Cabi Publishing

74) Prather J. 2011. Medical Data Mining: Knowledge Discovery in a clinical Data Warehouse. Available at URL:<http://www.amia.org/pubs/symposia.[Access date April 04, 2015].

75) R. J. Schalk. 2001, Artificial Neural Networks , McGraw-Hill.

76) R. Pal and I. Bose. 2009. An Optimization Based Approach for Deployment of Roadway Incident Response Vehicles.

77) Raghavan VV. 1998. A Perspective on Data Mining. Journal of the American Society for Information Science

78) Reddy, Adilakshmi and Swathi. 2014. A novel association rule mining and clustering based hybrid method for music recommendation system. International Journal of Research in Engineering and Technology vol 03 no 5. Available at http://www.ijret.org. [Accessed on 24 May 2015].

79) Shalini and Tripathi. 2012. An empirical study of tourist preferences using conjoint analysis. International journal of business science and applied management, vol5 no2.

80) Shalini N. Tripathi. 2010. An empirical study of tourist preferences using conjoint analysis. Int. Journal of Business Science and Applied Management, Volume 5, Issue 2.

81) Smith, Valene. 1977. Hosts and Guests: The Anthropology of Tourism, University of Pennsylvania Press, Philadelphia.

82) Tariku Atomsa. 2007. Tourism in Ethiopia: available athttp://www.eeaecon.org/ Papers% 20presented%20final/Tariku%20Atomsa%20%20Tourism%20in%20Ethiopia2_Quo_ Vadis.html.

83) Tesfaye Hintsay. 2002. Predictive Modeling Using Data Mining Techniques In Support to Insurance Risk Assessment. Masters Thesis Addis Ababa University, Addis Ababa,Ethiopia.

84) Thearling K. 1995. An Overview of Data Mining at Dun and Bradstreet. DIG White aper.

85) Tina R. and Patil S. 2013. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal Of Computer Science And Applications vol. 6, no.2. Amravati University.

86) Trybula WJ. 1997. Data Mining and Knowledge Discovery. Annual Review of Information Science and Technology (ARIST).

87) Tzu Ching Lin. 2012. Enhancing Tourism Intermediaries with the Data Mining Process, International Conference on Information and Knowledge Management. IACSIT Press, Singapore.

88) Tzu Ching Lin. 2012. Enhancing Tourism Intermediaries with the Data Mining Process, International Conference on Information and Knowledge Management (ICIKM), Singapore.

89) Umair Shafique and Haseeb Qaiser. 2014. A comparative study of data mining process models, KDD, CRISP-DM and SEMM. University of Gujrat, Pakistan, International Journal of Innovation and Scientific Research vol 12.

90) UNCTAD. 2010. Policy and practice for global tourism. Retrieved from http://www2.unwto.org/ publication/policy-and-practice-global-tourism

91) UNWTO. 2007. Annual Report of the World Tourism Organization on Development Assistance Activities, New York.

92) UNWTO. 2012. World tourism highlights 2012. Retrieved from http://dtxtq4w60xqpw. cloudfront. net/sites/all/files/pdf/ unwto_ highlights14_en.pdf

93) Vannevar Bush. 1945. As We May Think. Available at http://www.theatlantic.com /magazine /archive/1945.[accessed on February 20 2015].

94) Williams and Baláž. 2013. Tourism risk tolerance and competences.

95) Wirth and Hipp, R. (2000). "Towards a Standard Process Model for Data Mining". In Proceeding of the 4th International Conference on the Practical Applications of knowledge Discovery and Data Mining, Pages 29-39, Nanchester,UK.

96) Witten, Ian H. and Frank, Eibe. 2000. Practical Machine Learning Tools and Techniques with Java Implementations.Academic Press USA.

97) Wong J, Chen H. Chung, and P. Kao. 2012. Identifying valuable travelers and their next foreign destination by the application of data mining techniques. Asia Pacific Journal of Tourism Reserach, 11(4): 355-373.

98) Wong, J., Chen, H., Chung, P., & Kao, N. (2006). "Identifying Valuable Travelers and their Next Foreign Destination by the Application of Data Mining Techniques". Asia Pacific Journal of Tourism Research, 11(4), 355-373.

99) World Bank.2006.World Development Report of 2006: equity and Development.

100) Yabebal Muluneh. 2010. Tourist Flows and Its Determinants in Ethiopia. Ethiopian Development Research Institute, Addis Ababa.

101) Yipeng Zhou, Juan Hu, and Junping Du. 2009. Cross-media topic analysis and information retrieval.

102) Yo H. 2003. Sensitivity analysis for data mining: in Proceedings of the 22nd International Conference of the North American Fuzzy Information Processing Society NAFIPS. Chicago, Illinois.

103) Zou Guoxia. 2009. The Application of Data Mining in Tourism Information, International Conference on Environmental Science and Information Application Technology(ESIAT),

A p p e n d I x

J48 output for travel frequency class model

```
Age_Group = Group_2
|  No_of_Visitors = Med_No_Visitors
|  |  Nationality = USA
|  |  |  Travel_Purpose = Business: Frequent_Visitors (9.0/3.0)
|  |  |  Travel_Purpose = Conference: First_Time_Visitor (13.0/2.0)
|  |  |  |  |  Travel_Purpose = Visiting_Relatives_ &_ Friends: Frequent_Visitors (28.0/13.0)
|  |  Nationality = Middle East
|  |  |  Duration_of_Stay = LoS_Very_Ex: Frequent_Visitors (37.0/13.0)
|  |  |  Duration_of_Stay = LoS_Ex: First_Time_Visitor (6.0/1.0)
|  |  Nationality = North Africa: First_Time_Visitor (3.0)
|  |  Nationality = North America
|  |  |  Travel_Purpose = Business: First_Time_Visitor (4.0)
|  |  |  Travel_Purpose = Visiting_Relatives_ &_ Friends: Frequent_Visitors (19.0)
|  |  Nationality = North-East Asia
|  |  |  Travel_Purpose = Business: First_Time_Visitor (16.0)
|  |  |  Travel_Purpose = Conference
|  |  |  |  |  Duration_of_Stay = LoS_Very_Ex: Frequent_Visitors (2.0)
|  |  |  Travel_Purpose = Transit: Frequent_Visitors (9.0/4.0)
|  |  Nationality = Northern Europe
|  |  |  Travel_Purpose = Business: First_Time_Visitor (15.0/1.0)
|  |  |  Travel_Purpose = Conference: Frequent_Visitors (6.0/2.0)
|  |  |  Travel_Purpose = Leisure_and_Holiday
|  |  |  |  |  Duration_of_Stay = LoS_Very_Ex: First_Time_Visitor (14.0/1.0)
|  |  Nationality = Other Africa
|  |  |  Travel_Purpose = Business: First_Time_Visitor (10.0)
|  |  |  Travel_Purpose = Leisure_and_Holiday: First_Time_Visitor (22.0/6.0)
|  |  |  Travel_Purpose = Transit: Frequent_Visitors (18.0)
|  |  Nationality = Saudi Arabia: First_Time_Visitor (2.0)
|  |  Nationality = Tanzania: First_Time_Visitor (0.0)
Age_Group = Group_3
|  Nationality = USA
|  |  Travel_Purpose = Business: Frequent_Visitors (16.0/1.0)
|  |  Travel_Purpose = Conference: First_Time_Visitor (9.0/2.0)
|  |  Travel_Purpose = Leisure_and_Holiday: Frequent_Visitors (21.0)
|  |  Travel_Purpose = Visiting_Relatives_ &_ Friends
|  |  |  No_of_Visitors = High_No_Visitors: First_Time_Visitor (12.0/3.0)
|  |  |  No_of_Visitors = Low_No_Visitors: Frequent_Visitors (7.0)
|  |  |  No_of_Visitors = very_High_No_Visitors: First_Time_Visitor (8.0/1.0)
|  |  |  No_of_Visitors = Very_Low_No_Visitors: First_Time_Visitor (2.0)
|  Nationality = Greece: Frequent_Visitors (22.0/1.0)
|  Nationality = India: Frequent_Visitors (43.0/1.0)
|  Nationality = Israel: Frequent_Visitors (31.0)
|  Nationality = Italy: Frequent_Visitors (27.0/2.0)
|  |  No_of_Visitors = High_No_Visitors: First_Time_Visitor (15.0)
|  |  No_of_Visitors = Low_No_Visitors: Frequent_Visitors (29.0)
|  |  No_of_Visitors = very_High_No_Visitors: First_Time_Visitor (4.0)
|  |  No_of_Visitors = Very_Low_No_Visitors: Frequent_Visitors (7.0)
|  Nationality = Netherlands: Frequent_Visitors (42.0)
|  |  Travel_Purpose = Business: First_Time_Visitor (19.0/1.0)
|  |  Travel_Purpose = Visiting_Relatives_ &_ Friends: Frequent_Visitors (6.0)
|  Nationality = North-East Asia: Frequent_Visitors (55.0/2.0)
|  Nationality = Northern Europe: Frequent_Visitors (54.0/2.0)
|  Nationality = Norway: Frequent_Visitors (60.0)
```

| Nationality = Other Americas: Frequent_Visitors (63.0/1.0)
Age_Group = Group_1
|| No_of_Visitors = Low_No_Visitors: First_Time_Visitor (315.0)
| | Nationality = Other Europe
| | | Travel_Purpose = Transit: First_Time_Visitor (6.0)
| | | Travel_Purpose = Visiting_Relatives_ &_ Friends: Frequent_Visitors (6.0/1.0)
| | Nationality = Greece
| | | Travel_Purpose = Visiting_Relatives_ &_ Friends: Frequent_Visitors (13.0/1.0)
| | Nationality = India
| | | Duration_of_Stay = LoS_Very_Ex
| | | | Travel_Purpose = Leisure_and_Holiday: First_Time_Visitor (3.0)
| | | | Travel_Purpose = Visiting_Relatives_ &_ Friends: Frequent_Visitors (2.0)

Rule generated By Apriori for travel frequency

1. Duration_of_Stay=LoS_Med No_of_Visitors=Very_Low_No_Visitors Travel_Frequancy=Frequent_Visitors 2448 ==> Age_Group=Group_3 2448 <conf:(1)> lift:(2.08) lev:(0.12) [1270] conv:(1270.72)
2. No_of_Visitors=Very_Low_No_Visitors Age_Group=Group_2 2017 ==> Travel_Frequancy=First_Time_Visitor 2017 <conf:(1)> lift:(1.87) lev:(0.09)
3. No_of_Visitors=Very_Low_No_Visitors Source _of_Motivation=History Age_Group=Group_3 1827 ==> Duration_of_Stay=LoS_Med 1827 <conf:(1)> lift:(1.79) lev:(0.08) [803] conv:(803.5)
4. Duration_of_Stay=LoS_Med Source _of_Motivation=History Age_Group=Group_3 1829 ==> No_of_Visitors=Very_Low_No_Visitors 1827 <conf:(1)> lift:(1.42) lev:(0.05) [541] conv:(181.1)
5. Source _of_Motivation=Cultural_Life Age_Group=Group_1 1895 ==> No_of_Visitors=Very_Low_No_Visitors 1872 <conf:(0.99)> lift:(1.41) lev:(0.05) [539] conv:(23.45)
6. Source _of_Motivation=History Age_Group=Group_3 1854 ==> Duration_of_Stay=LoS_Med 1829 <conf:(0.99)> lift:(1.76) lev:(0.08) [790] conv:(31.36)
7. Source _of_Motivation=History 3181 ==> No_of_Visitors=Very_Low_No_Visitors 3137 <conf:(0.99)> lift:(1.4) lev:(0.09) [900] conv:(21)
8. Duration_of_Stay=LoS_Med Age_Group=Group_2 1968 ==> No_of_Visitors=Very_Low_No_Visitors 1937 <conf:(0.98)> lift:(1.4) lev:(0.05) [553] conv:(18.27)
9. No_of_Visitors=Very_Low_No_Visitors Source _of_Motivation=History 3137 ==> Duration_of_Stay=LoS_Med 3082 <conf:(0.98)> lift:(1.75) lev:(0.13) [1324] conv:(24.64)
10. No_of_Visitors=Very_Low_No_Visitors Source _of_Motivation=Cultural_Life 1915 ==> Age_Group=Group_1 1872 <conf:(0.98)> lift:(4.34) lev:(0.14) [1441] conv:(33.73)
11. No_of_Visitors=Very_Low_No_Visitors Age_Group=Group_3 Travel_Frequancy=Frequent_Visitors 2541 ==> Duration_of_Stay=LoS_Med 2448 <conf:(0.96)> lift:(1.72) lev:(0.1) [1024] conv:(11.89)
12. Source _of_Motivation=History Travel_Frequancy=First_Time_Visitor 1743 ==> Duration_of_Stay=LoS_Med 1677 <conf:(0.96)> lift:(1.72) lev:(0.07) [700] conv:(11.44)
13. No_of_Visitors=Very_Low_No_Visitors Age_Group=Group_2 2017 ==> Duration_of_Stay=LoS_Med 1937 <conf:(0.96)> lift:(1.71) lev:(0.08) [807] conv:(10.95)
14. No_of_Visitors=Very_Low_No_Visitors Age_Group=Group_2 Travel_Frequancy=First_Time_Visitor 2017 ==> Duration_of_Stay=LoS_Med 1937 <conf:(0.96)> lift:(1.71) lev:(0.08) [807] conv:(10.95)
15. Source _of_Motivation=Natural_and_Wildlife Age_Group=Group_3 1953 ==> Duration_of_Stay=LoS_Med 1758 <conf:(0.9)> lift:(1.61) lev:(0.06) [663] conv:(4.38)
16. Age_Group=Group_2 3082 ==> Travel_Frequancy=First_Time_Visitor 2774 <conf:(0.9)> lift:(1.69) lev:(0.11) [1128] conv:(4.65)
17. Duration_of_Stay=LoS_Med Travel_Frequancy=First_Time_Visitor 3104 ==> No_of_Visitors=Very_Low_No_Visitors 2769 <conf:(0.89)> lift:(1.27) lev:(0.06) [587] conv:(2.74)