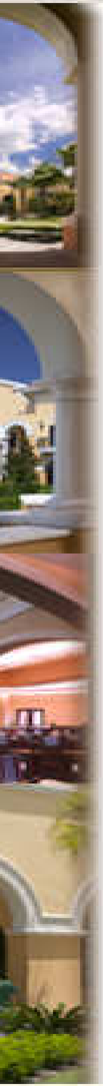


Data Mining in Tourism Data Analysis: Inbound Visitors to Japan

Ms. Valeriya Shapoval, University of Central Florida
Dr. Morgan C. Wang, University of Central Florida
Dr. Tadayuki Hara, University of Central Florida
Mr. Hideo Shioya, JTB Foundation

Introduction

- Japan has strong potential to have a strong and competitive presence in the world tourism market
- According to the JNTO, total arrivals to Japan in 2000 were 4,757,146 people, in 2012 total tourist arrivals to Japan were 8,358,105 and in 2013 total arrivals were 10,363,922 which increased by 24 % from previous year
- Potential
  - Little research is done about Japanese market
  - None known to us research has being done using big data

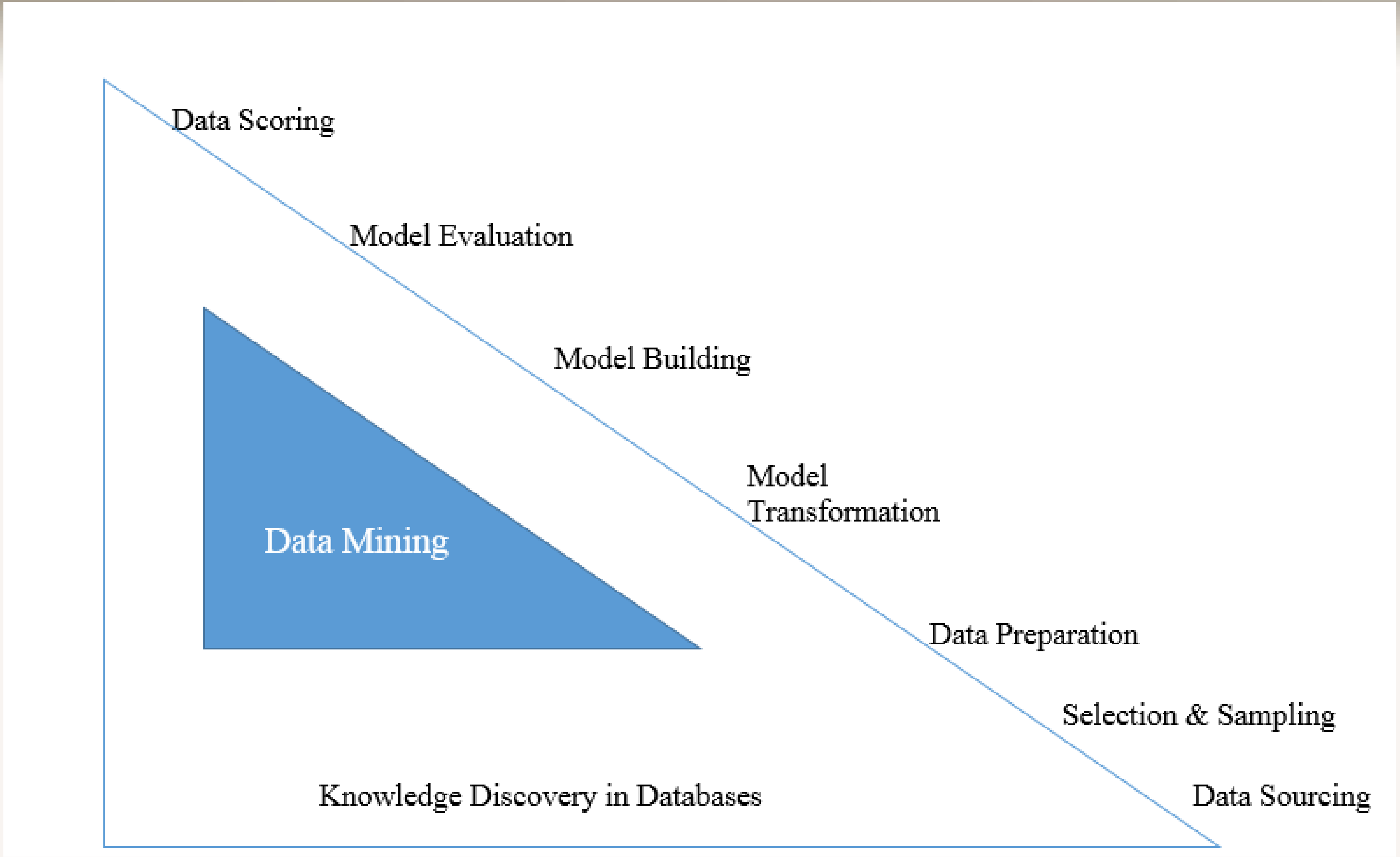
What is Data Mining?

The non-trivial extraction of implicit, previously unknown, an otentially useful information from data (Frawley et al., 1991)

ata mining uses machine learning algorithms to find pattern f relationship between data elements in large, noisy, and messy data sets, which can lead to actions to increase benef s some form (diagnosis, profit, detection, ect.) knowledge liscovery in data (Nisbet, Edler and Miner, 2009 p. 17).

Knowledge discovery in databases is the non-trivial process dentifying valid, novel, potential useful, and ultimately nderstandable patterns in data (Fayyad et al., 1996)

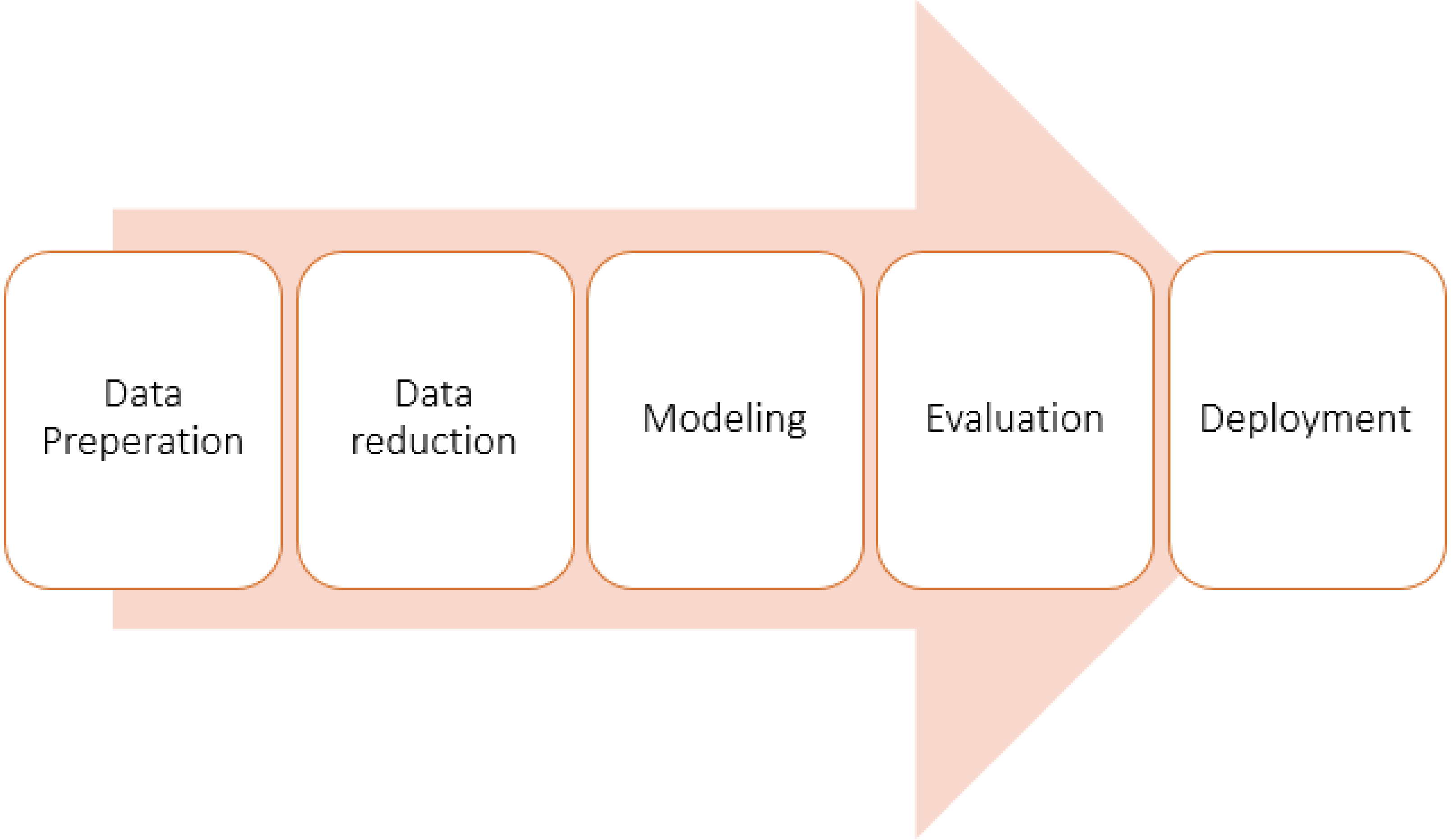
Processes in Data Mining



Data Mining versus classic Statistics

- Classical statistics has large subjective component, predictive model is known and main goal is to estimate parameters and/or confirm/reject hypothesis
- Statistical learning (Data mining) is much more manageable when there are no restrictions placed on the model for a given data, in other words where analysis are data driven and complexity of given machine learning are dependent on underlying distribution according of which we desire to learn (Hosking, Pednault & Sudan, 1997).

Procedural Steps in Data Mining



Neural Networks

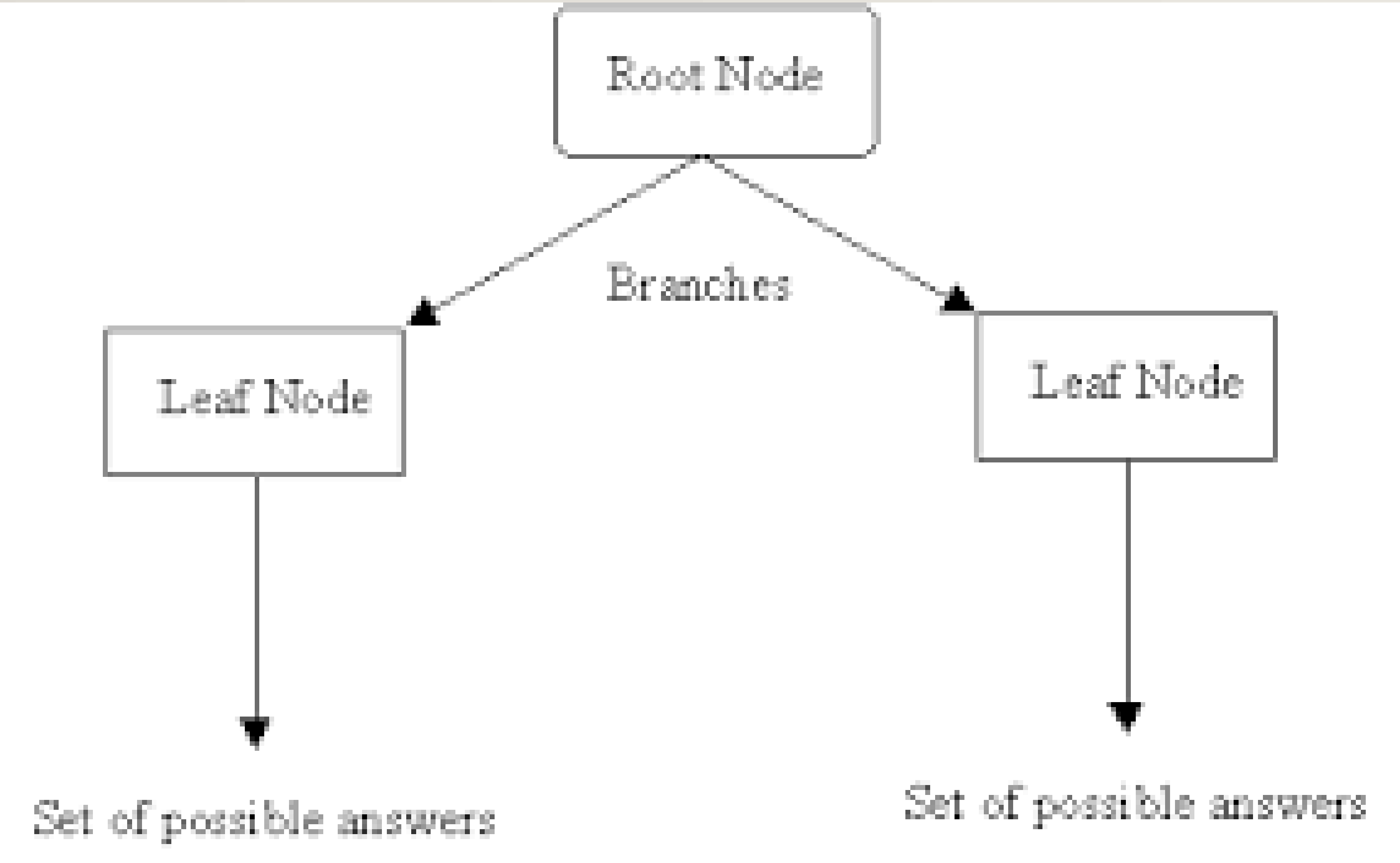
- Neural networks (NN) are capable to generalize and learn from data mimics, which can be in the way related to a one learning from one's own experience.
- Draw back of the technique is results of training NN are weight that are distributed through network and do not provide valid insight as to why given solution is valid.
- NN is a good tool for prediction and estimation problems.

Decision Trees

Decision Trees (DT) are form of multiple variable analyses.”… it is a structure that can be used to divide up a large collection of records into successfully smaller sets of records by applying a sequence of simple decision rules (Berry and Linoff 2004 p. 6).”

Nisbet, Edler and Miner, (2009) “DT is a hierarchical groups of relationships organized into tree-like structure, starting with one variable (like trunk or an oak tree) called a root node (p. 241)

Decision Trees



Root Node

Branches

Leaf Node

Leaf Node

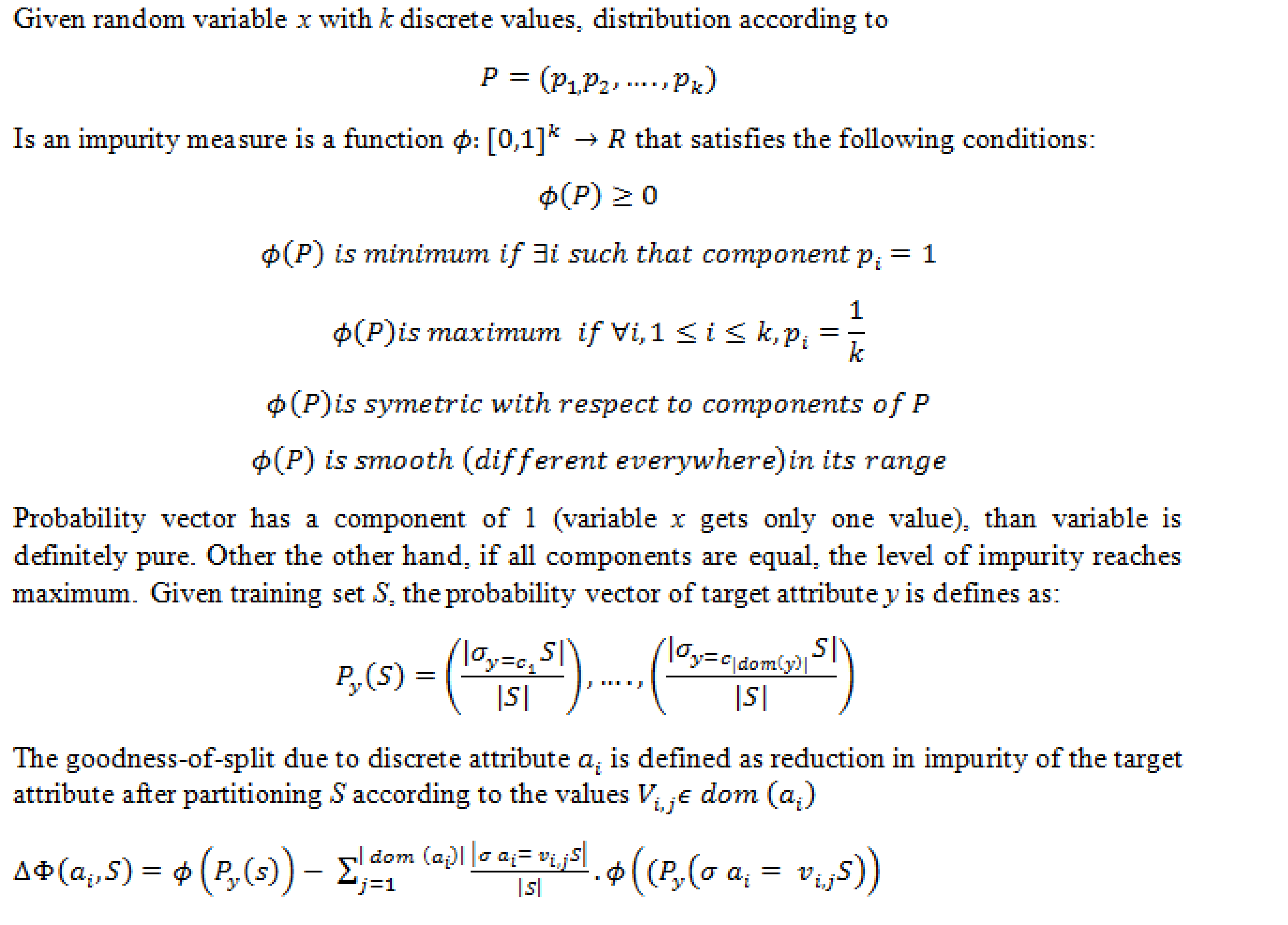
Set of possible answers

Set of possible answers

Impurity-based Criteria

- In many cases in Decision Tree split is done according to the value of single variable. Most common criteria for a split is an impurity based split.

Impurity-based Criteria



Given random variable x with k discrete values, distribution according to

$$P = (p_{1,}p_2, \ldots., p_k)$$

Is an impurity measure is a function $\phi: [0,1]^k \rightarrow R$ that satisfies the following conditions:

$$\phi(P) \geq 0$$

$$\phi(P) \text{ } is \text{ } minimum \text{ } if \text{ } \exists i \text{ } such \text{ } that \text{ } component \text{ } p_i = 1$$

$$\phi(P) is \text{ } maximum \text{ } \text{ } if \text{ } \forall i, 1 \leq i \leq k, p_i = \frac{1}{k}$$

$$\phi(P) is \text{ } symetric \text{ } with \text{ } respect \text{ } to \text{ } components \text{ } of \text{ } P$$

$$\phi(P) \text{ } is \text{ } smooth \text{ } (different \text{ } everywhere) in \text{ } its \text{ } range$$

Probability vector has a component of 1 (variable x gets only one value), than variable is definitely pure. Other the other hand, if all components are equal, the level of impurity reaches maximum. Given training set S, the probability vector of target attribute y is defines as:

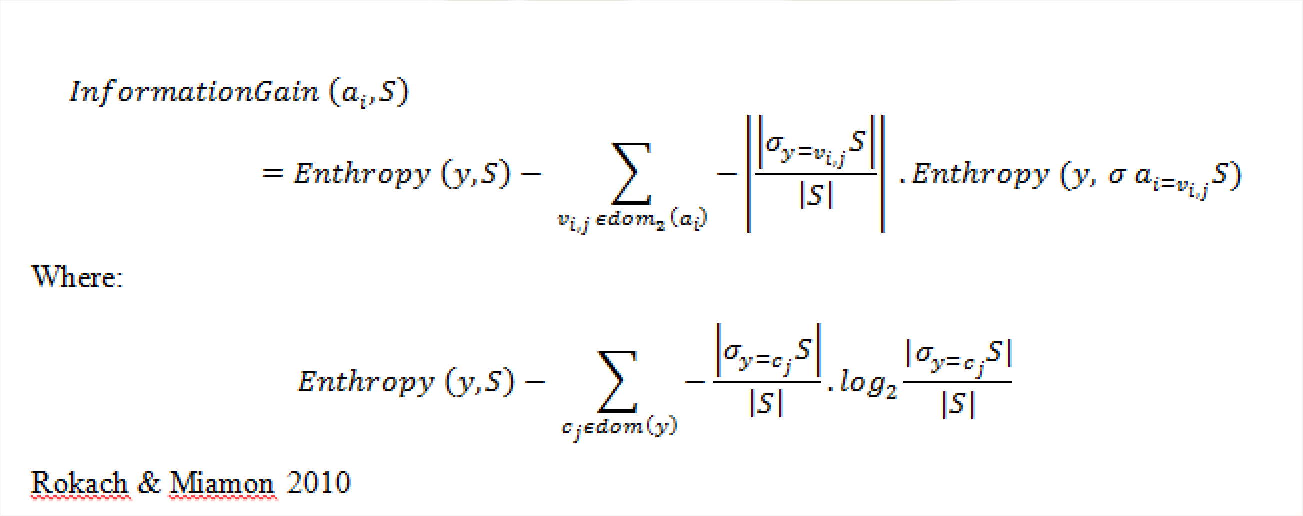
$$P_y(S) = \left(\frac{|\sigma_{y=c_1}S|}{|S|}\right), \ldots., \left(\frac{|\sigma_{y=c_{|dom(y)|}}S|}{|S|}\right)$$

The goodness-of-split due to discrete attribute a_i is defined as reduction in impurity of the target attribute after partitioning S according to the values $V_{i,j} \epsilon \text{ } dom \text{ } (a_i)$

$$\Delta\Phi(a_i, S) = \phi\left(P_y(s)\right) - \Sigma_{j=1}^{|\,dom\ (a_i)|} \frac{|\sigma\ a_i = v_{i,j}S|}{|S|} . \phi\left((P_y(\sigma\ a_i = \ v_{i,j}S)\right)$$

Information Gain

- Entropy information gain was used. Information gain is impurity based criterion that uses the entropy measure as an impurity measure.



$$InformationGain\ (a_i,S)$$

$$= Enthropy\ (y,S) - \sum_{v_{i,j}\,\epsilon dom_2(a_i)} - \left| \frac{\left|\sigma_{y=v_{i,j}} S\right|}{|S|} \right| .\, Enthropy\ (y,\ \sigma\ a_{i=v_{i,j}} S)$$

Where:

$$Enthropy\ (y,S) - \sum_{c_j \epsilon dom(y)} - \frac{\left|\sigma_{y=c_j} S\right|}{|S|} .\, log_2 \frac{|\sigma_{y=c_j} S|}{|S|}$$

Rokach & Miamon 2010

Theoretical Background

- Tourism is one of the world’s major industries that contributes significantly to the global economy and became one of the major sources of wealth for some developing and developed counties.
- Due to the increasing competition among tourist destinations in the last several decades, destination marketing managers and industry practitioners have become concerned about their destinations’ images in the minds of tourists (Wang & Pizam, 2011).

Theoretical Background

ccording to UNWTO Japan had a 23% of positive growth in nternational tourism receipts, this creates a need in understanding a atterns of consumer expenditures in Japan.

estination marketing organizations need to know how their destinatio s perceived by potential visitors, so they can better target their marke nd develop more appropriate tourism products and increase destinat ttractiveness (Phillips and Back, 2011).

arketers should take consumer behavior into consideration, where ultural differences, extend of planning time before vacation and numb f people in the group influences expenditure of tourist (Leasser and olnicar, 2012).

Data and Methods

ata were collected by JTB-Foundation on behalf of Japan Tourism gency during year 2010 at the airport and seaport. Inbound tourists t apan were approached at random by representatives of JTB foundati articipants were asked to participate in the survey. Data were collect n the likert, binary scale and sample size of 4,000 usable observatio

his study employed casual research design. The survey questionnai onsisted of following major sections:

tourist attributes of satisfaction, overall satisfaction, intention to retu

and questions that consists of tourists' demographical questions su as country, party size, gender age, and number of children.

Results: Future intention to return

Variable	Description
5_1_01	Experienced Japanese Food
5_1_06	Shopping
3_02	Transportation
1_01	Lonely Planet as a major source of information about Japan prior to visit
1	Which airport did you land in Japan
2	How many time have you visited Japan including this visit
5_1_1Area	Main area (destination) in Japan visited
2_06	Internet as a main helpful source in obtaining information while in Japan
5_2_04	Desire to experience nature/scenery sightseeing next visit
_E	Flight cost
esident	Country of residency
5_2_05	Want to walk around downtown in the future
4_b_ck	Catering cost
3_e5	Cosmetics and pharmacy expenditure
ational	Nationality
2_07	Credit Cards as a method of payment in Japan
ge	Age
esidents of China	Residents of China

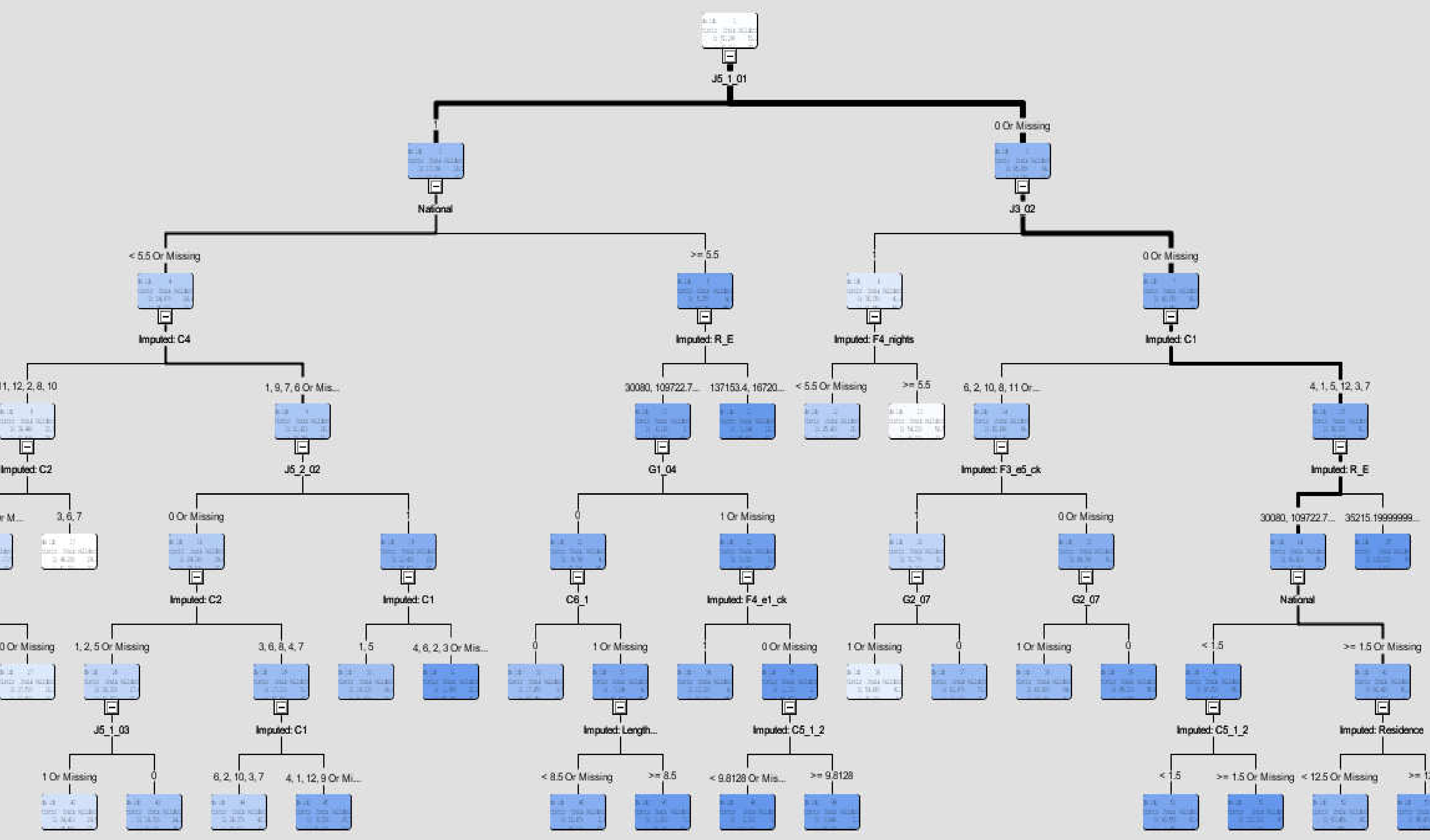
Results: Satisfaction

Variable	Description
J5_1_01	Japanese food
J5_1_06	Shopping
J3_02	Availability of Information on transportation
Residence	Country of residence
National	Nationality
C1	Airport
C5_1_1	Main area (destination) in Japan visited
C4	Main purpose of the visit
C5_1_2	Secondary destination visited in Japan
F4	Main place where tourist stayed in Japan
C2	Prior visit to Japan
J5_1	Business trip
F3_e5	Cosmetics and Pharmacy expenditure
G2_07	Credit Cards as a method of payment in Japan
Length of stay	Length of stay
J2_02	Would like to stay in Japanese style inn next time/appeal of Japanese hospitality
J1_03	Hot spring experience
J5_2_04	Desire to experience nature/scenery sightseeing next visit
C7	Organized tour

Demographical Factors

- Asia (62%) such as Korea (19.51%), Taiwan (18.10 %), Main Land China (14.16%). Second largest visitors are from USA (10.65%).
- From Main Land China two largest groups Beijing and Shanghai.
- man (56%) and woman (43%).
- Average age was 23 years with standard deviation of 13 years.
- Majority of the tourists arrived in Narita (53.88%), Kansai (17.63%), and New Chitose (Sapporo) (6.212%).
- 42% of respondents visited Japan for the first time, 15% visited for the second time and 10% for the third time.
- General distribution of group travelers are: alone (17%), family (21%), work colleague (19%), and friends (19%). 57.9% of respondents travel for tourism and leisure (57.9%), and business training, conference or trade fair (25 %).

Decision Tree on Satisfaction



Odds Ratio

- Odds ratios are used to compare the relative odds of the occurrence of the outcome of interest (e.g. disease or disorder), given exposure to the variable of interest (e.g. health characteristic, aspect of medical history). The odds ratio can also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome.
  - OR=1 Exposure does not affect odds of outcome
  - OR>1 Exposure associated with higher odds of outcome
  - OR<1 Exposure associated with lower odds of outcome

Decision trees node rules: Satisfaction

Rule 1:

g Odds Ratio of tourists being satisfied is higher by 1.39 if they are from non-Asian country, experienced Japanese food, came for business purposes or visit friend, and shopped at local department store.

Rule 2:

g Odds Ratio of tourists being satisfied is higher by 1.64 if they are from neighboring Asian country (Korea, China, Taiwan, Hong Kong and Thailand), stayed at Japanese style inn, experience Japanese food, came for tourism/leisure, Incentive travel, Study, or International conference, and came through two main airports (Narita/Haneda)

Rule 3:

g Odds Ratio of tourists being satisfied is higher by 1.64 if they are mainly non-Asian countries, experienced Japanese food, paid between $300 and $1,500 for air fare, and used accommodations other than western-style hotels

Decision trees node rules: Satisfaction

Rule 4:

g Odds Ratio of tourists being satisfied is higher by 2.32 if tourists are mainly from non-Asian countries, had experience with Japanese food, paid between $300 and $1,500 for air fare, purchased Japanese fruits, and shopped at supermarket.

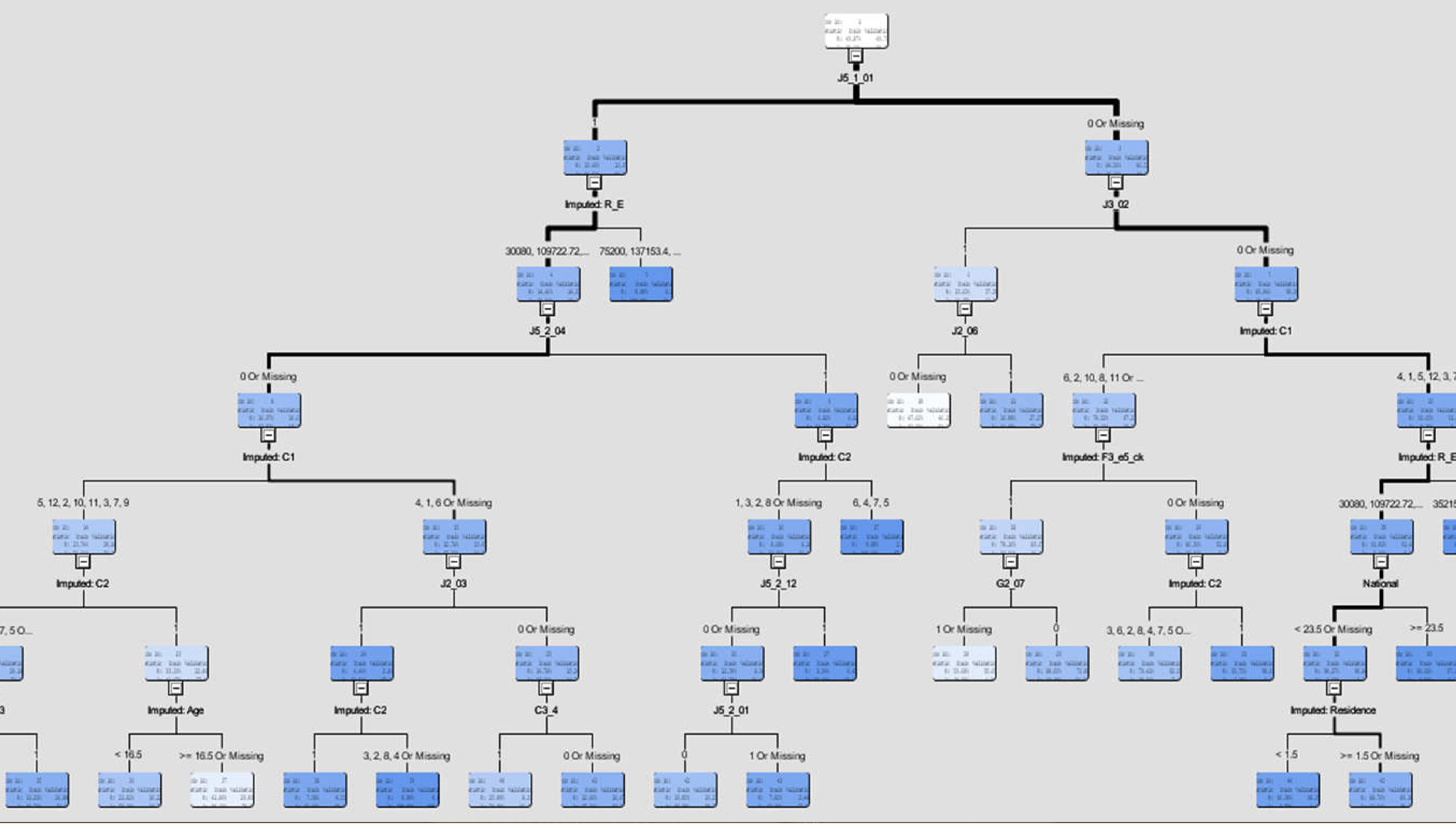
Rule 5

g Odds Ratio of tourists being satisfied is higher by 1.51 if tourists are from neighboring Asian country (Korea, China, Taiwan, Hong Kong and Thailand), experienced Japanese food, came for tourism or exhibition/conference/company meeting, and visited Japan more than once before.

Rule 6:

g Odds Ratio of tourists being satisfied is higher by 2.21 if tourists are mainly from non-Asian countries, paid between $300 and $1,500 for air fare, experienced Japanese food, stayed less than 8 days, stayed at western style hotel.

Decision Tree on Intention to return



Decision Tree Rules: Intention to Return

Rule 1:

g Odds Ratio of tourists having intention to return is higher by 3.9 if they experienced Japanese food, but have not experienced Japanese nature/scenery sightseeing, paid between $300 and $1,670 for air fare, visited Japan for the first time and came through airports such as Narita, New Chitose (Sapporo), or Fukuoka.

Rule 2:

g Odds Ratio of tourists having intention to return is higher by 3.9 if tourists experienced festival/event, Nature/scenery/sightseeing, Japanese food, paid between $300 and $1,670 for air fare, and visited Japan several times.

Rule 3:

g Odds Ratio of tourists having intention to return is higher by 1.30 if tourists experienced Japanese food, but not Nature/scenery/sightseeing, paid between $300 and $1,670 for air fare, first time visitors, and young age.

Decision Tree Rules: Intention to Return

Rule 4:

Odds Ratio of tourists having intention to return is 1.13 if tourists experienced Japanese ood, but not Nature/scenery/sightseeing, want to experience Japanese hot spring in the uture trip, paid between $300 and $1,670 for air fare, and came with family, spouse or riends.

Rule 5:

Odds Ratio of tourists having intention to return is 1.94 if tourists experienced Japanese ood, but not Nature/scenery/sightseeing, want to experience Japanese hot spring in the uture trip, and came with family, spouse or friends.

Rule 6:

Odds Ratio of tourists having intention to return is 1.49 if tourists want to experience in the uture nature/scenery/sightseeing, experienced Japanese food, and paid between $300 and $1,670 for air fare.

Conclusion on Satisfaction

ormation of Tourist satisfaction differs between two distinct roups which are Asian and non-Asian tourists with different references to achieve high level of satisfaction

lon-Asian tourists would include experience with Japanese ood, shopping at department store, stayed at western style otel, came on business or visit friend and air fare cost

or Asian tourists those factors would be experience with apanese food, stay at Japanese style inn, attending an even uch as conference, incentive travel or study, previous visit to apan and importantly they have no preferences for airfare co

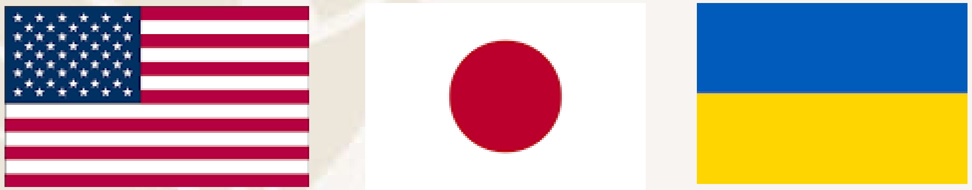
Conclusion: Intention to return

More family-oriented tourists or non-business tourists without previous visits are more likely to return.

Main drive for a future return is not experiences people had, but rather experiences people want to have in the future such as Japanese hot spring or nature.

Experience with a Japanese food remains universally attractive with all combinations.

Data analyses can indicate set of tourism policies and marketing strategies which would be less likely to fail (= higher likelihood of successes), because it is not based on emotion or subjective views.



Thank you! *From an international research team of Ukraine, USA and Japan*

(Corresponding author: valeriya.shapoval@ucf.edu)