PREDICTION OF CUSTOMER PREFERENCE BASED ON CUSTOMER'S ACTIVITY HISTORY

By

Lee Yip Fung

TARC

TUNKU ABDUL RAHMAN
UNIVERSITY COLLEGE

FACULTY OF APPLIED SCIENCES AND COMPUTING
TUNKU ABDUL RAHMAN UNIVERSITY COLLEGE
KUALA LUMPUR

ACADEMIC YEAR
2016/17

Table content

1. INTRODUCTION 5

1.1 OBJECTIVES 5
1.2 RESEARCH QUESTION AND HYPOTHESIS 5
1.3 BACKGROUND 6
1.4 ADVANTAGES AND CONTRIBUTIONS 7
1.5 PROJECT PLAN 8
1.5.1 Features 8
1.5.2 Development model 8
1.5.3 Data analyze 8
1.5.4 Milestones 9
1.6 PROJECT TEAM AND ORGANIZATION 10
1.7 THESIS OUTLINE: 10

2. RESEARCH BACKGROUND AND RELATED WORK 12

2.1 INTRODUCTION TO PREDICTION MODEL 12
2.2 COMPETITORS COMPANY BACKGROUND 13
2.3 PROJECT BACKGROUND 14
2.4 LITERATURE REVIEW 15
2.4.1. Process model for building prediction model 15
2.4.1. Supervised Machine learning algorithm 16
2.4.2. Support Vector Machine(SVM) 16
2.4.3. Artificial Neural Network (ANN) 17
2.4.4. K-Nearest Neighbour(KNN) 18
2.4.5. Naive bayes 18
2.4.6. Random forest 19
2.5. FEASIBILITY STUDY 20
2.6. CHAPTER SUMMARY AND EVALUATION 21

3. METHODOLOGY AND REQUIREMENTS ANALYSIS 25

3.1. METHODOLOGY (EXTREME PROGRAMMING) 25
3.1.1 Fact Gathering 26
3.1.2 Fact Recording 27
3.1.3 Fact Analysis 29
3.2. REQUIREMENTS ANALYSIS 29
3.2.1. Project Scope 29
3.2.2. Development Environment 31
3.2.3. Operation Environment 31
3.2.4. External Interface Requirements 31
3.2.5. Non-functional Requirements 31
3.2.6. Functional Requirements 32
3.2.7. Discussions 32
3.3. CHAPTER SUMMARY AND EVALUATION 32

4. SYSTEM DESIGN 35

4.1 APPLICATION DEVELOPMENT PROJECT AND PACKAGE IMPLEMENTATION PROJECT 35
4.2.1. Pros and Cons of the system design 36
4.2.2. Alternative solution 36
4.2. RESEARCH PROJECT 36
4.2.3. Pros and cons of research framework 37
4.3. CHAPTER SUMMARY AND EVALUATION 37

5. APPENDIX 38

[Appendix - part A] 38
[Appendix- part B] 39

[Appendix - part C] *41*

6. REFERENCES **42**

Chapter 1

Introduction

1. Introduction

Nowadays people can access internet to get a lot of information and data such as we can simply search which food and restaurant is nearby just done by using online search engine. Sometimes the searching results might not meaningful to user is just simply because the searching is not "smart" enough to provide useful search results which match user preference.

In this project we were planned to do a food hunting mobile application which allow user to search reasonable food or restaurant and give suggestion food or restaurant. In our mobile application we will implement prediction model algorithm to analyze user activity history such as search history, place visited, restaurant facility to understand user preference.

In the prediction model we will roughly understand user preference such as: most likely will visit which area of restaurant or prefer which kind of food. Example: if a user is leaving in Setapak area and most of the food searching history is related to spicy, the searching result,suggestion search result and push notification will show 2 or more spicy food restaurant to user which are in setapak area.

1.1 Objectives

- Understand user preference by study user data.
  - we can study user data such as the searching history and the view history.

- provide more relevant suggestion to user which match with user preference.
  - example: 99% of view history are related to curry, so the suggestion will show curry food to user such as: curry fish head.

- Archive highest prediction accuracy if possible.
  - the count of user click the suggestion notification will determine the accuracy of prediction model. the higher count of user click suggestion notification, the higher accuracy will be awarded.

- Explore most efficient way to process user data.
  - minimize the computation cost on processing huge amount of data in terms of computation time and computation power require. Use the least compute resources to process huge amount of data if possible.

- Explore most suitable algorithm for prediction model.
  - find the lowest outlier rate algorithm which fit to this project.
  - to increase the productivity of prediction model.

1.2 Research Question and Hypothesis

research question is discuss about the current problem or uncertainty of this current project which we need to solve and ensure our project product have quality.

Research question

1.) what kind of algorithm should implement in our prediction model?
2.) is it possible have other way/ soution to improve the accuracy of prediction model?
3.) how to improve the data process efficiency of prediction model?

in this project I have form 2 hypothesis which I need to prove it either true or false in this project.

Hypothesis

1.)the more user activity history, the easier to understand user preference.
2.)the more user activity history, the higher accuracy of information push to user.

1.3 Background

In literature review of prediction model, most of the prediction model was frequently used in hospitality industry such as predict the patient disease state. Example: prediction model for breast, endometrial, and ovarian cancer.[1]

Nowadays food searching mobile application such as "HoChak!" and "OpenRice", all the similar application obviously does not implement prediction model and suggest relevant restaurant to user base on what kind of food user served the most. Only the OpenRice mobile application will push some notification which show to user the latest food hunting article but sometimes the location and food is not relevant to user preference and behaviour.

Most of the food searching mobile application are like a library application is because most of the application was like a food library and just search and show the results to user. Most of the time the searching result is not match user expected result. Example: while a spicy lover user search "nearby food" the application will show a list of restaurant but didn't show spicy food restaurant more frequent compare to other restaurant.

1.4 Advantages and Contributions

Below were few advantages and contributions of our research project.

- Predict user preference
    - Study user activity history and predict what user want and prefer to. Example: user may like to eat spicy food base on most recent 20records were eating spicy food.
    - Suggest relevant predict results to user such as a kind of food or restaurant user may like to visit.
- High efficient computation power prediction model
    - Study and adjust the sweet spot of data amount to predict user preferences.
    - Use the least of computation time to get highest productivity if possible.
    - Example: only process most recent 20 records and predict what kind of food user currently interested.
- Explore most suitable algorithm (in terms of prediction accuracy)
    - Find, study and modify to produce the high quality prediction model with high prediction accuracy.

- Research field
    - In manufacturing industry can be use to make forecasting and easy to make strategy planning.[2]
    - In hospitality industry need prediction model to make more accurate diagnostic on patient.[3]

1.5 Project Plan

1.5.1 Features

Table 1 below are the application features and its usage that we planned to build.

Features	usage
High quality searching engine	Analyze what user search the most and show more relevant searching results
Push suggestion notification	Predict what kind of food and restaurant user prefer and show to user via notification.
Auto improve prediction accuracy	The more data provided by user, the more data will be analyzed and the higher prediction accuracy will be archived.

1.5.2 Development model

The development model will be use in this project is 1 of the agile approach methodology which is extreme programming. This methodology allow to make a lot of requirement changes even in last minute and is high productive but the consequences is need to communicate and deliver many times to our client.[4] This is the first project and i expect will make a lot of mistake and maybe need to change requirement in last minute when the initial idea is not work.

1.5.3 Data analyze

Normally clickthrough data will be higher accuracy that meet user expected result is because user rarely click and view search results randomly[5]. In data analyze, i will use some machine learning algorithm to optimize the search results and prediction model by analyze user activity history. Table 2 below will briefly show what kind of data will be analyzed and the usage.

Analyze data	purpose	usage
Search results history	Observe what kind of data that user searched in most of the time.	Improve the prediction accuracy and search results.
View history	Observe what restaurant and food user have clicked and viewed.	Improve the prediction accuracy and search results.
User location	Roughly knowing the area user visited the most.	Suggest relevant food and restaurant within that area.

Table 2

1.5.4 Milestones

Table 3 below is the rough schedule of the whole project and what should be done on time.

Title/item to do	Planned deadline	What should do
Chapter 1 [introduction]	9/6/2017 [week2]	Intro and research
Chapter 2 [Research background]	30/6/2017[week5]	Background research, literature review, theoretical understanding
Chapter 3 [methodology & requirement analysis]	14/7/2017[week7]	Research approach, development model (extreme programming).
Chapter 4 [design]	28/7/2017[week9]	System design and algorithm. (system flow)
Coding and testing all the way…….	29/7/2017[week10] to 18/8/2017	Coding whole system and perform unit test.
Enhancement [efficiency]	18/8/2017 to 16/10/2017	Enhance the performance of the algorithm in terms of efficiency and bug less if possible.
Chapter 5 [system review]	16/10/2017	Test case and test plan
Final test with supervisor	30/10/2017	Final test of the system

Table 3

1.6 Project Team and Organization

Table 4 below is the distribution of job scope and module of the team members.

System and sub-system	Pang Wai Kian	Lee Yip Fung
Activity history analyze engine		x
search result history		x
view history		x
location history		x
Customer Behaviour Estimation	x	
Text Classification	x	
UI design		
-Searching UI -FoodHuntArticle UI		x
-EPayment UI -QR Code Scanning UI - Payment Information UI	x	
Testing and Quality assurance	x	x
Activity history analyze engine		x

Table 4

1.7 Thesis Outline:

Thesis I. The more activity done by user, the more data will be capture.

a. Data
   i. Search results history
   ii. View history
   iii. User location

Thesis II. The more data analyzed, the higher the prediction accuracy.

a. Algorithm research
   i. Accuracy of algorithm result.
   ii. Efficiency of algorithm in terms of productivity and time to process

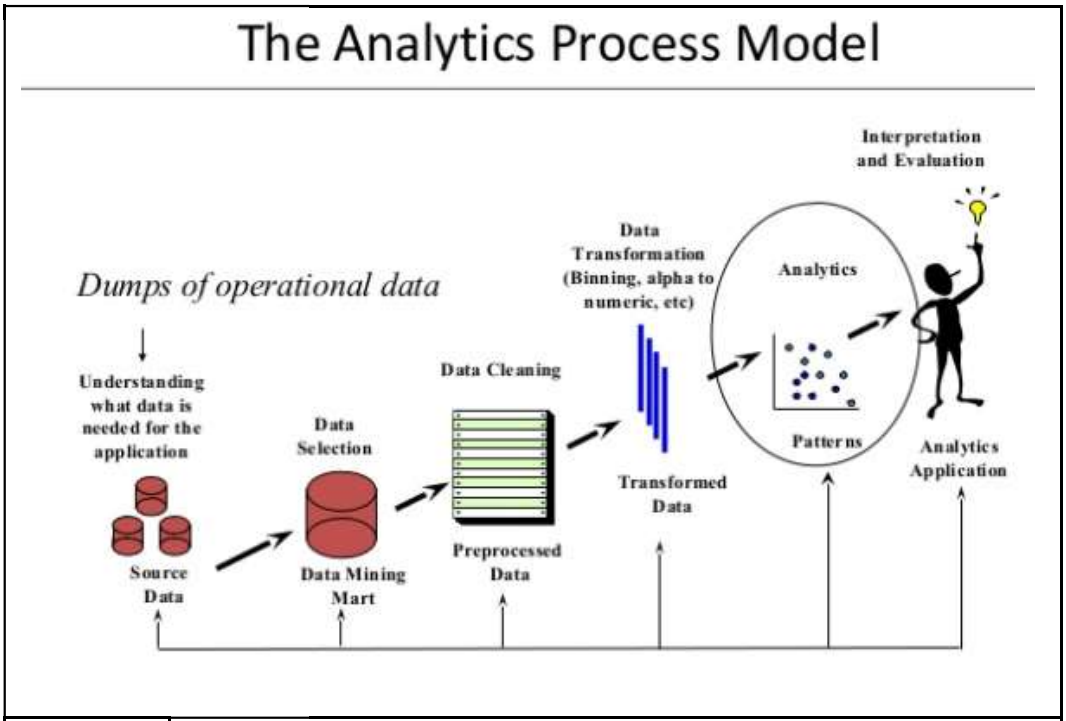
Chapter 2

Research Background and Related Work

2. Research Background and Related Work

2.1 Introduction to prediction model

Prediction model is a part of predictive analytic in data science. in easier understand manner, it is a model which study historical data to forecasting and make prediction for future. The prediction is using statistics number to show the likelihood of that particular prediction will happen in future. In the table 2.1 below with diagram and description will describe the whole process from collecting data stage until make prediction stage.



source data	data from various form such as: SQL database, microsoft excel
data mining	discover the relationship of data and discover the data pattern.
preprocess data	data filtering process to ensure data have quality to minimize errors of data. Example: decide what to do with missing data, whether to discard outlier or other solution.
transform data	because different data will store in various form and format, we need to transform it into reasonable or process-able manner to process those data by using machine learning.
analytic	visualize the data information to easier understandable manner. example: histogram,barchart,piechart
analytic application (knowledge)	extract knowledge from those visualized information. understand what is the logic behind those information. example: we observed majority of customer buy fried chicken and coke together in the analytic stage, so we understand (knowledge) fried chicken and coke is the perfect match and will increase the sales revenue when we put it together.

Table 2.1[6]

Basically in the prediction model, we need to include predictive analytic which is the most important component of prediction model.[7] In predictive analytic, we will using machine learning algorithm and data mining techniques to train and study huge amount of data sets (structured and unstructured data) in-order to discover the data pattern.[8] After discover data pattern, we can extract information and knowledge from data and make prediction with showing the likelihood numbers.

2.2Competitors company background

In Google Play store and apple store have a lot of similar mobile food finding application such as HoChak and OpenRice. The main competitor company to our mobile application is OpenRice. OpenRice was a famous dining experience sharing platform in Hong Kong and extended their platform to many country which included: Malaysia, China, Taiwan, Macau, Japan, Indonesia, Philippines, Singapore and Thailand.[9][10] Obviously, Malaysia does not have the culture like Hong Kong people love to using mobile application to find food and restaurant by observing the number of comments and feedback of particular restaurant on OpenRice website and mobile application.

In my personal user experience based on using OpenRice mobile application, this application was provided many restaurant to refer compare to other similar application such as included small hawker stall, "kopitiam", "mamak" restaurant, cafe and snacks delights shops. Furthermore, OpenRice provided more complete searching function compare to other similar application such as "HoChak" is because you can adjust your searching setting by change the location, distance, food type, environment and other more setting.

In my opinion, OpenRice most interesting function is allow user to write and post food article. User can write their personal thought and what they have experienced on having that particular food and restaurant environment. When you clicked the restaurant link inside the article, it will brings you to the restaurant page,show you the restaurant details, their famous dishes menu and you can view more reviews from others users.

Other than that, user will also can make reservation on certain restaurant and user can enjoy certain promotion by referring terms and condition of OpenRice. On the other hand, restaurant owner are allow to promote their food and restaurant in this platform, which means this platform not only will benefits to consumer and also benefits to restaurant owner.

Table 2.1 below was the sample screen capture in OpenRice Hong Kong version and Malaysia version.

Screen capture of Malaysia restaurant (4reviews)

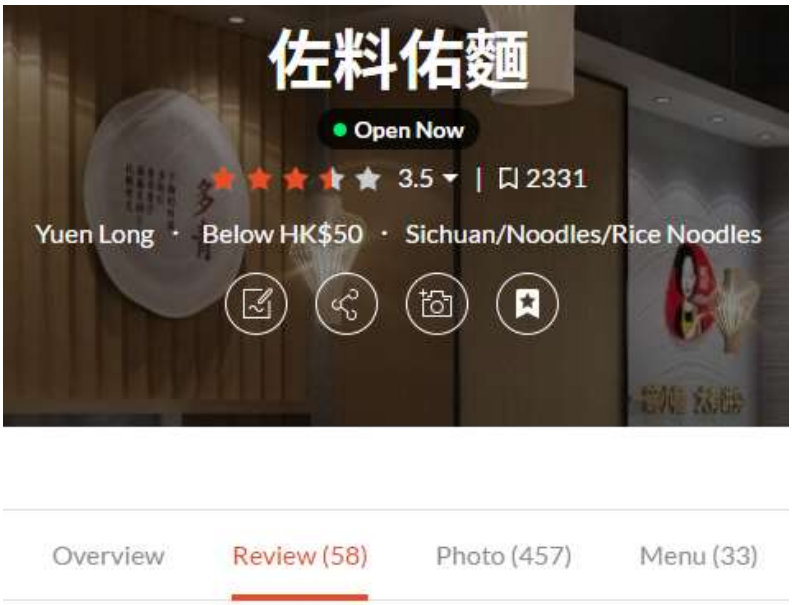
Screen capture of Hong Kong restaurant (58 reviews)


Table 2.1

2.3 Project Background

Prediction model was not a new term in now modern days. They were a lot of system has been implemented prediction model especially in online goods selling website such as: "Lazada", "Amazon"," 淘 宝 ". The main reason for all goods selling website were implemented prediction model is because we can briefly understand customer preference which means we can guess and understand what customer needs. Example: if a customer buy a desktop processor, the suggestion list will suggest customer to purchase thermal paste too is because logically a processor need thermal paste to transfer the heat to heatsink. In order to understand this logic, prediction model using machine learning algorithm to study millions of customer records and found out majority of them were buy both items at the same time in

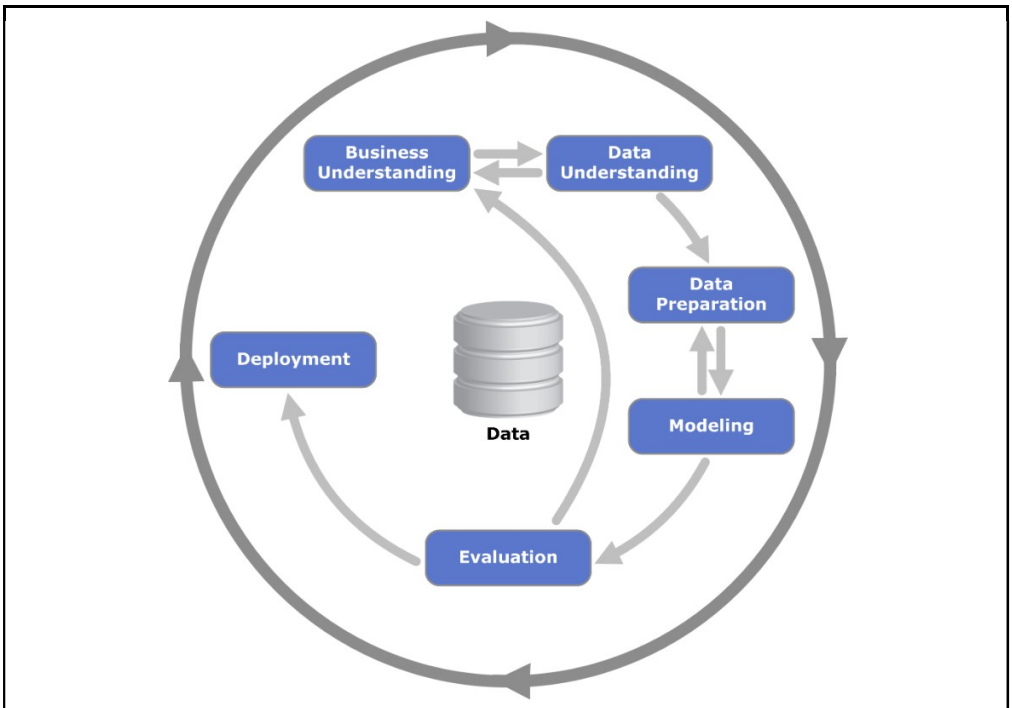
previous purchasing record in database, that's how prediction model understand what majority customer need by studying over millions even trillions of records.

In this project, we will design a "food hunt" mobile application which will implement prediction model to understand each user preference and suggest food or restaurant to user. In order to understand user preference, we will feeding all user data (user activity history) to train the prediction model such as user search result, clicked items and location visited. the prediction model will study and analyze all data and giving suggestion to user. Since this is the smart campus research project, so our main user are TARUC staff and students, secondly is public user.

2.4Literature Review

2.4.1. Process model for building prediction model

In develop and deploy a prediction model into our project, there was a process model we need to go through to produce a quality prediction model. Table 2.2 below was the process model to producing prediction model with description.



phases	description
business understanding	analyze and understand the business logic and the policy works in business perspective.
data understanding	understand all the available data and understand the relationship in between all available data in-order to perform data mining in next following stages.
data	perform data cleaning to ensure all data have quality and can produce

preparation	higher accuracy of prediction.
modeling	select the most suitable modeling techniques and apply on it. such as apply K-nearest neighbour , ANN and other techniques. observe the results of each techniques before making final decision.
evaluation	evaluate the built model in different perspective such as accuracy, performance and productivity. we should make sure the model is not over-fitting else will just perform well in testing phases but cannot deploy and use it in real life situation.
deployment	deploy the model into real life products.(food hunt application)

table 2.2[11]

2.4.1. Supervised Machine learning algorithm

Machine learning was the tools to study huge amount of data and visualize in understandable manner. For "food hunt" mobile application project, I have chosen 5 supervised machine learning algorithm to review which is suitable for prediction. below this paragraph was the brief literature review for each algorithm (SVM, ANN, KNN, Naive bayes, Random forest).

2.4.2. Support Vector Machine(SVM)

- This algorithm will plot all data as a point in multidimensional space and identify the hyper-plane to classify out the cluster of data.[12]
- Example with explanation:

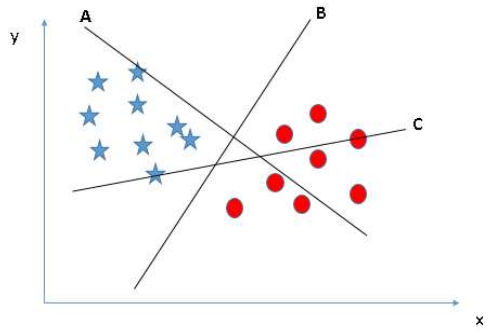


diagram 2.1

Explanation:
SVM form three hyper plane which is A,B and C in diagram 2.1. the rules of SVM is need to partition out 2 cluster in this situation. So for this situation, hyper-plane B is the best solution and will be chosen. When new data fall at the left side of plane B, which means will be predicted is a star else at the right side will be predicted is a circle.

- Pros
  - works well if data is clustered in well manner.

- High memory efficiency in training data compare with others method or algorithm.

- Cons
  - Does not work well if data is not clustered in well manner such as has outlier.
  - Does not include probability estimation.

2.4.3. Artificial Neural Network (ANN)

- This algorithm is work as a human brain neurons. This network algorithm is trained to associate outputs with input patterns which is suitable in pattern recognition.[13]
- Example with explanation:

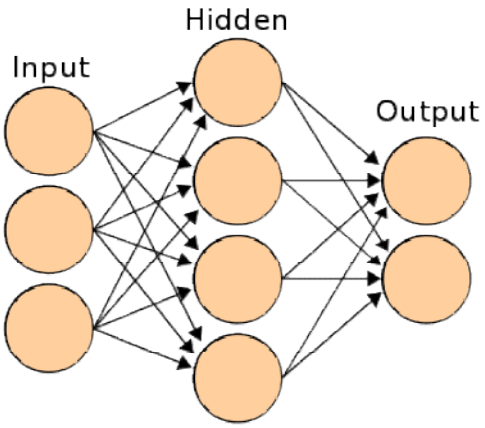


diagram 2.2[14]

Exaplaination:
The input will transfer to hidden layer, the hidden layer will process the information and transfer output layer. in this algorithm, input layer, hidden layer and output layer are affecting each other same as human brain learning some knowledge. [15]

Example in human perspective: in first time learning 1+1, human may answer equal to 3, but the actual answer is equals to 2, so we will correct our selve 1 increase by 1 is equal to 2.

Example apply to ANN: input is 1,1 but the hidden layer is perform subtraction so the answer equals to 0 which does not match with the expected output answer 2. so the hidden layer will change to addition in-order to match the output answer equals to 2.

- Pros
  - Working best in answering "what if" situation because it can generate multiple possible path.[13]
  - Hidden layer will be adjusted based on using what data to train ANN and the results will be more reasonable and rational
- Cons
  - Take longer computation time and higher computation resources consumption (use more RAM and CPU cores).
  - Overfitting the data. (may have lower accuracy)
  - Not suitable for predicting outcomes.

2.4.4. K-Nearest Neighbour(KNN)

- This algorithm is suitable for both regression and classification predictive problems. This algorithm is using the "K" value to determine the accuracy of the prediction.[16]
- Example and explanation:

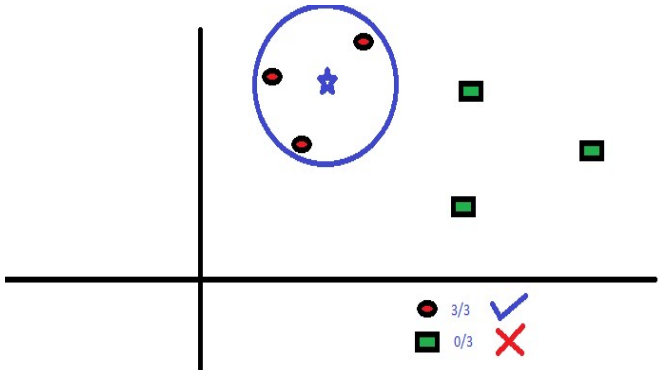


diagram 2.3[16]

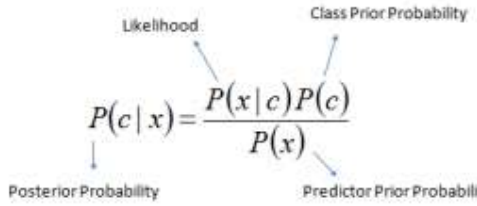
explanation:
In the diagram 2.3, we can clearly see there was 3 small circle and 3 small square. we assume that the small star is the new data which is unknown category. In the diagram we assume that the value "K" is equal to 3 so the radius from the big circle is value 3. Within the big circle we can see that 3small circle were inside the big circle (K value coverage) and non of small square inside the big circle. In conclusion, we can predict the small star is belongs to small circle category.

- Pros
  - Shorter training time to get the results with smaller data sets.
  - Using lesser memory with smaller data sets
  - Allow outlier data
- Cons
  - Take longer time to plot all data with large data sets
  - Take more memory with large data sets
  - Hard to determine the "K" value

2.4.5. Naive bayes

- This algorithm is using probability to predict the outcome.
- Example and explanation:

Explanation:
Data sets features will be converted into frequency table to calculate the likelihood table.[17] After we get the likelihood table, we will able to predict the outcome by calculation using formula in diagram 2.4 below. Example in diagram 2.5 below.



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

diagram 2.4[12]

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

diagram 2.5[17]

Example:
Scenario: What is the probability will play during sunny weather?
formula**: P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)**

P (Sunny |Yes) = 3/9 = 0.33
P(Sunny) = 5/14 = 0.36,
P(Yes)= 9/14 = 0.64

Conclusion: P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60

2.4.6. Random forest

- Random forest is the algorithm which create multiple CART model (decision tree) by using different sample and different random initial values from data sets. This method is to reduce the noise and variance of model.[18]
- Example and explanation:

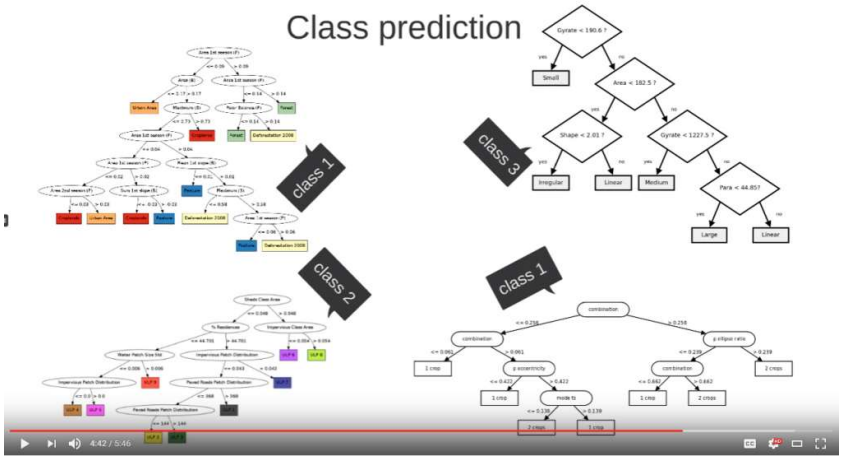


diagram 2.6[19]

Explanation:
Let's assume that we split the data set randomly into 4 subset and random initial values will be chosen in each subset respectively to produce 4 decision tree same as the diagram 2.6 above. When the new unknown data put inside this 4 model will show different prediction results same as diagram 2.6above. The higher occurrence same results will be chosen (class 1, 2times occurrence).

- Pros
  - Higher prediction accuracy
  - Lower overfitting data
- Cons
  - Consume higher computation power and resources
  - Take longer time to training data
  - Not suitable for small data sets

2.5. Feasibility Study

First of all, taste of food is good or bad actually is a very subjective to every single person. Example: Keeanu is my indian friend and unfortunately he was not a curry lover, as long as got curry in any dishes, he will feel disgusting when he smell curry. Normally, majority of indian was love curry because their food culture mainly is about curry. So in this situation, we cannot assume that our data sets will not have outlier and our algorithm must be tolerate will outlier in this situation.

Secondly, in current Malaysia market of food finding mobile application, majority does not implement prediction model. So we might be consider was the first food finding application with prediction model in Malaysia. Thirdly, data set size was the main concern in this application. We cannot assume we can get a very large data set which is usable. In table 2.3 below was the summary of the 5 machine learning algorithm which is suitable to be implement or not.

Lastly, real life data from Malaysia currently is not available, we will planning taking overseas data from kaggle which is "Burritos in San Diego".[20] This data is about Mexican food enthusiasts rate 10 dimensions of hundreds of burritos in San Diego and we use this for testing our prediction model in developing and testing stages for simulation purpose. The second solution is generate data artificially in lower data noise manner.

Algorithm	Comments	Suggestion to implement

SVM	• Fast and efficient • not tolerate with outlier	not suggested
ANN	• overfitting data • good approach in problem solving • not suitable in prediction	not suggested
KNN	• fast and efficient • suitable for small and large data sets • tolerate with outlier • hard to define "K" value	suitable for almost real time prediction (the food you may want to eat in your holiday), faster response
Naive Bayes	• fast and efficient • using probability • tolerate with outlier • calculate probability by using occurrence frequency	suitable for real time prediction (today food suggestion), faster response
Random forest	• higher prediction accuracy • works well with huge data sets • consume more time and computation power	suitable for forecasting report (example: food trend in 2025),take longer time to respond.

table 2.3

2.6. Chapter Summary and Evaluation

In this chapter, actually is focusing and research about prediction model. table 2.4 below was the few section I have already reviewed and concerned.

area focus	summary	examples
project background		

introduction of prediction model	• source data
	  o where's the data came from
	• data mining
	  o study relationship between data features or variable
	• preprocess data
	  o data filtering, make sure data have quality
	• transform data
	  o transform data into preferred format
	• analytic
	  o visualize the data to understandable manner
	• analytic application (knowledge)
	  o extract knowledge from visualized information.
		refer to "analytic process model" in section 2.1

main competitor	• features and showcase
	  o article/ blog
	  o food reservation
	  o promotion and advertisement
	  o bigger food and restaurant database
	  o largest food experience sharing platform in Asia.
		OpenRice mobile and web application

real life prediction model application	real life application which implement prediction model to improve the products and services.	amazon, taobao, lazada and etc.
Literature review		

process model to build prediction model	• business understanding
	  o understand the business culture and the business policy
	• data understanding
	  o understand the data in different perspective to perform data mining easily
	• data preparation
	  o pre-process data to ensure data have quality and minimize error of data
	• modeling
	  o review and select suitable algorithm to build a high accuracy model. (example: use KNN)
	• evaluation
	  o review the model in different perspective such as: accuracy, performance, productivity.
	• deployment
	  o implement and put into production
		refer to "Process model for building prediction model" in section 2.4.1

machine learning	algorithm suitable for prediction model • KNN • naive bayes • random forest • ANN • SVM	predict user preference and forecasting future trend
feasibly study		

data characteristics	• will have outlier
	  o different people different taste preference
	• will have incomplete data
	  o not all rolls and column have data
	• real datasets from Malaysia is not available
	  o data from kaggle
		-

	○ generate data artificially	
algorithm suggestion	algorithm suitable for current project • KNN • naive bayes • random forest	-

Since there was a lot of information in online article and online video links can refer to, I think this project is achievable, it just depends it can be achieve in proper and good manner or achieve with a lot of uncertainty manner. In this project, the main concern for me is about the performance of the prediction model. We need to make sure the prediction model is able to produce high accuracy of prediction in shorter time and minimize the consumption of computation power as much as possible.

Chapter 3

Methodology and Requirements Analysis

3. Methodology and Requirements Analysis

3.1. Methodology (Extreme programming)

Basically extreme programming methodology is a methodology which keep iterating the development cycle and deliver a lot of increments with end user. Example: developer meet with end user and understand the needs of end user then code a prototype first to show end user. After end user have reviewed the prototype and given feedback to developer. Developer will modify the current prototype based on the end user feedback and repeat meet with end user again and again until the whole end products is done. Diagram 3.1 below was the basic process model of extreme programming.

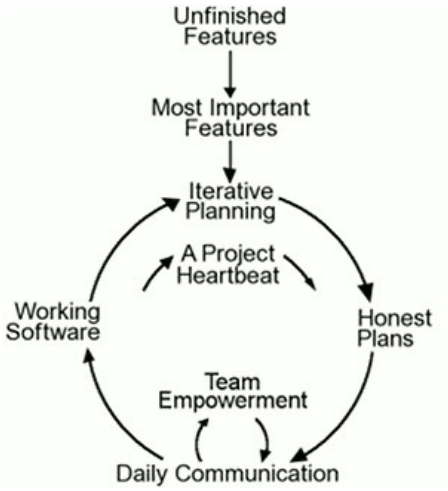


diagram 3.1[21]

This methodology was involved face-to-face meeting of developer with end user which means our end user is able to know how was the development progress and can know whether each requirement of customer is done in proper way.[22] Sometimes some requirement is hard to document it and very hard to understand by developer just by referring those few sentences in documentation without referring to the core person (example: end user or client). Table 3.1 below was the advantages of using extreme programming.

Advantages	example of situation	solution of extreme programming
Understand end user requirement clearer	Sometimes is hard to figure out whether developer have understand end user requirement correctly until developer deliver to end user only know which part have done it wrongly.	Extreme programming will deliver a lot of increment and demo to end user whether have match with end user requirement.
Allow to change requirement at anytime	Sometimes end user may missing out some requirement and may need to change the existing requirement. Sometimes some	This model is focusing more in coding, increment deliver and end user satisfaction rather than focus on documentation. Since everything is not

	end user may not even know what they really want.	fixed like waterfall model, developer can modify the whole code and add in new code for new requirement even in last minute without referring other department of developers or documentation.
High productivity	Example in SDLC approach only have few deliver to end user, if have any problem have found, the cost will be increase exponentially.	Every small increment is keep in touch with end user, so developer are able to make changes before it is too late to make changes and is lower cost compare with SDLC approach.
High customer satisfaction	Sometimes because of time and cost constraints, end user only can accept some bad design instead of asking developer to make changes. Example: the "change" module shouldn't appear at home page but only left 3 days and no extra money to make changes, so just accept and deploy it.	Since every increment is keep in touch with end user, so any dissatisfied part will be found and fix before it is too late.

table 3.1[21]

This is the first prediction model project for me, I think I will make a lot of mistake since everything is new to me. This approach allow me to make modification in every single deliver to end user and can make any requirement changes at anytime before it is too late.

3.1.1 Fact Gathering

Before this chapter, I have already briefly explained what is prediction model and how does it work in our project (for more information can refer to "appendix- part A" or refer to previous chapter under "literature review" section on page 15). Instead of let user understand our application, why don't we understand user? So the main objectives of our project application is make user feel satisfy and feel our system is understand them so well with high engagement. The following sections within this fact gathering section will describe how the system engage with user and how to enhance the user experience.

How the real system algorithm works

The main things we focusing is how long the time user spent on viewing something within our application such as: food viewing, restaurant viewing, article viewing and etc. Base on the time period threshold, we will assign the likelihood number and store in database. So we can roughly understand whether the user are interested on specific viewing item and how likely was it.

Explanation example in sequence order:

1.)User A spent 2minutes reading "NZ curry house" restaurant and spent 5 seconds viewing "oldtown kopitiam" cafe, then the likelihood probability of "NZ curry house" is 0.7 and "oldtown" is 0.2.

2.)We assume that NZ curry house signature dish is curry fish head and the main element of majority dishes is curry.

3.)We assume that user A viewing others restaurant with curry element and with high likelihood numbers. Example viewed history in table 3.2 below

restaurant/food	time viewing	likelihood numbers
rampai court curry fish head	5minutes	0.78
curry mee business park	2minutes	0.65
ah leong curry fish head	10minutes	0.84
abu kari ikan	3minutes	0.71

table 3.2

4.)Because of the likelihood numbers and the history of user A, machine learning algorithm can understand user A is love to eat curry and fish head.

5.) When next time user searching some food, the first few search results will contain either have curry element or fish head element or both element. Promotion and other notification will also work in the same way.

Project specification

Based on literature review on few competitor mobile application, the few features we are planning to implement into our mobile application are listed in table 3.3 below with description.

features	description
prediction based on likelihood numbers	predict how likely was the specific user like specific food and restaurant.
push suggestion notification	suggested promotion, restaurant and food
basic and advanced searching	search nearby, by ranking, by area,by distance,by price
forum base design	allow user to post review and opinion
add restaurant/ food	add new restaurant or add new food on specific restaurant
bookmark (hunt list)	for end user future refer purpose and increase likelihood number

table 3.3

In this project, data is one of the important component in the process model and real free access data in Malaysia is currently not available for our project. The way to collect data that i choose is using questionnaire method. I will using google form to design questionnaire and distribute the form by sharing via various social media such as: facebook, whatsapp and wechat. The questionnaire is basically need to observe what kind of food is famous and which category (race,age group, gender, area staying and etc) of people will having those food.

3.1.2 Fact Recording

Diagram 3.1, 3.2 and 3.3 below was database table design to store restaurant data,user view history data and user profile data respectively. For more information about sample data can be refer to Appendix part-B.

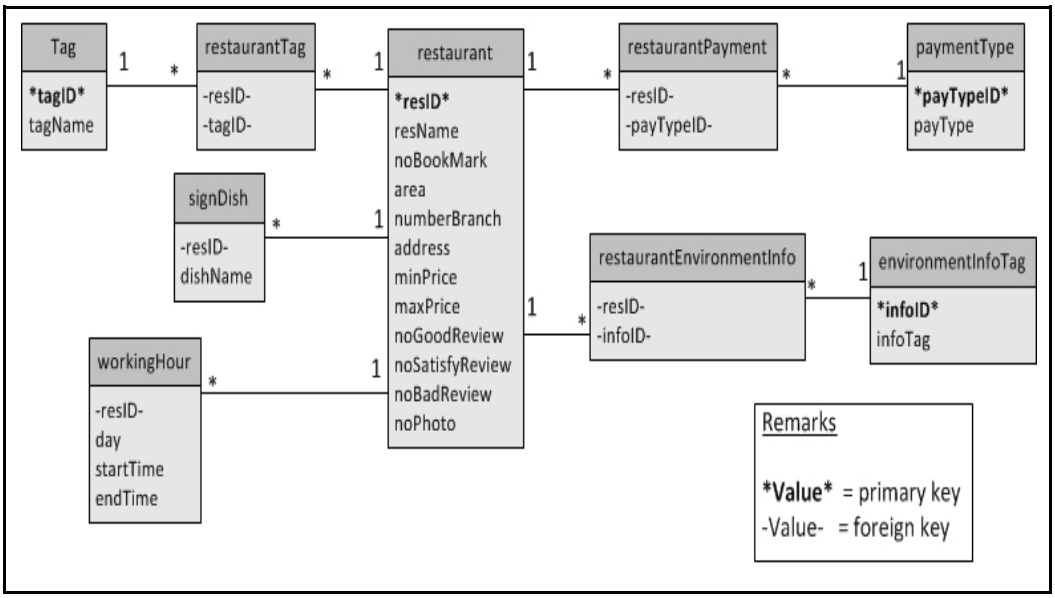


diagram 3.1

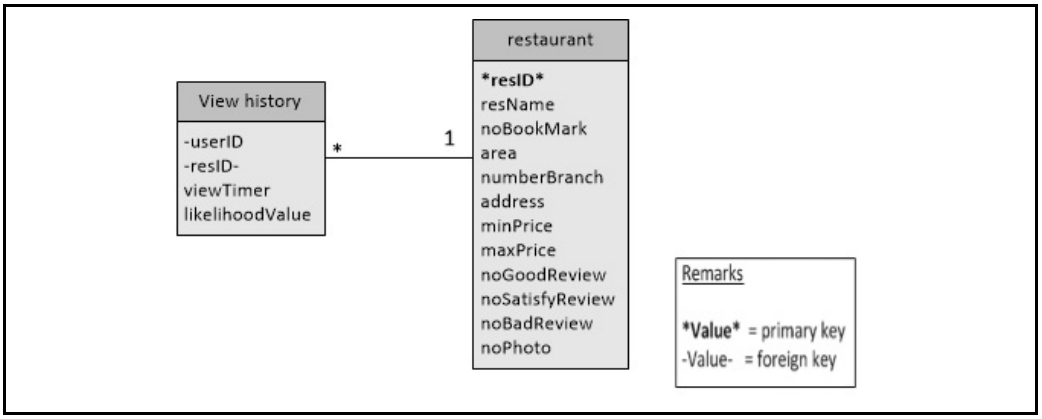


diagram 3.2

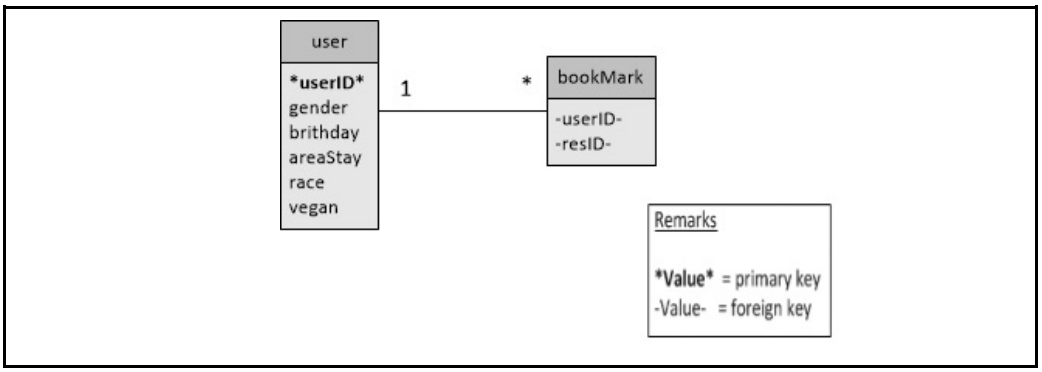


diagram 3.3

diagram 3.4 below was the data flow diagram of the whole prediction model works and how the data flow within our system.

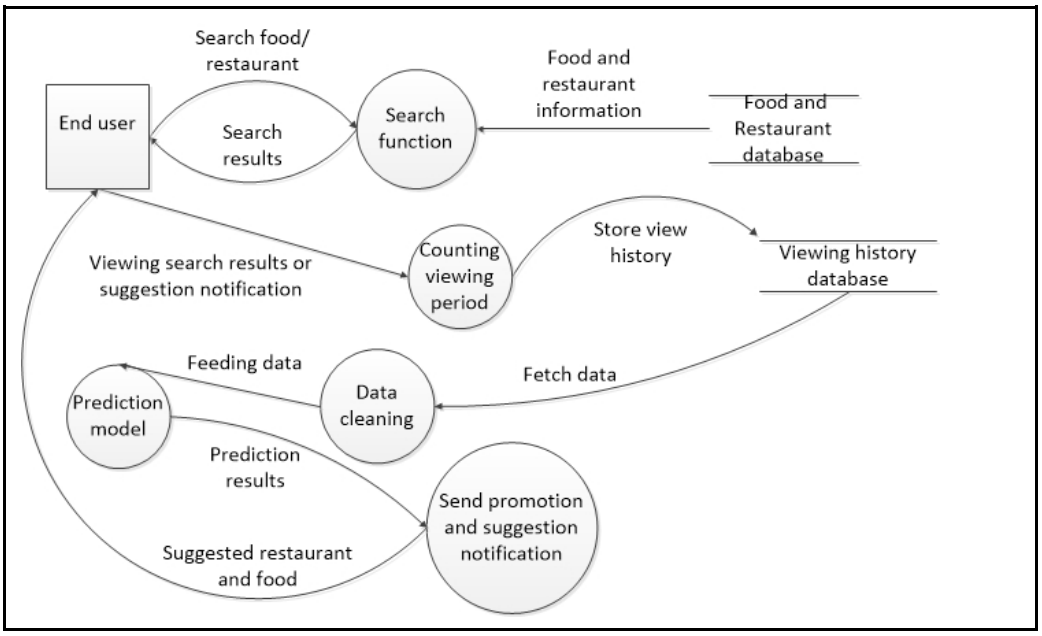


diagram 3.4

3.1.3 Fact Analysis

After we collected data, we will perform data cleaning and data mining to reduce the noise of the data before feeding it into prediction model. Data cleaning will filter the data in different perspective such as: filter missing value data, outlier data and others data. In data cleaning process, we will observe the data and decide whether to modify the filtered data or remove it because it will influence the final results. Data mining is to study the relationship between data variables. Example: "age group" and "food type" variable can describe which age group of people most likely prefer to what kind of food. For more information about the data cleaning and data mining process explanation, please refer to Appendix part C analytic process model or refer to previous chapter (Chapter 2) under section of "2.1 Introduction to prediction model" on page 12 .

3.2. Requirements Analysis

3.2.1. Project Scope

In my project scope, basically my main focus job is to tracking and analyze user activity history to predict user preference and send push suggestion notification to user. Figure 3.1 is the Hierarchical Chart showing Project Sub-system and table 3.4 is the description of each component within Hierarchical Chart.

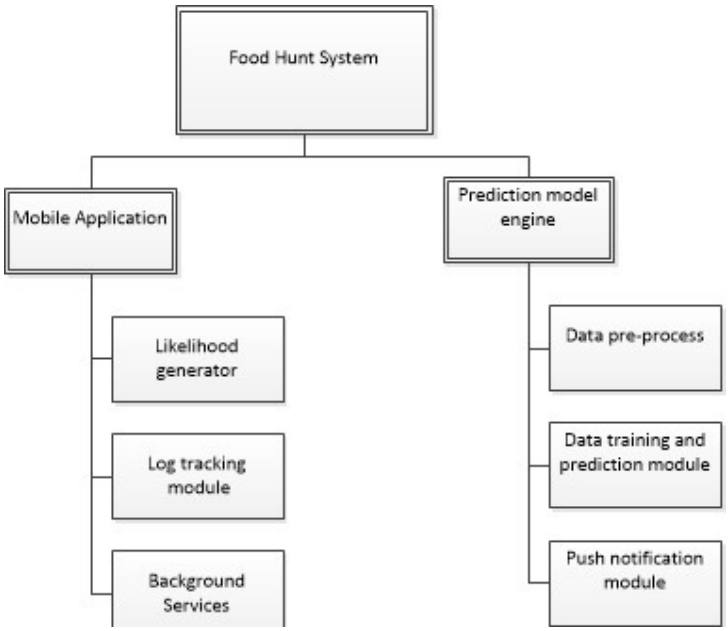


Figure 3.1

Module	Sub-Module	Description	Example
Mobile Application	Likelihood generator	calculate the likelihood number base on the length of time user viewing each item (restaurant/food)	1minute=0.5 3minute=0.67 20seconds=0.45
	Log tracking module	collect and tracking user activity within our mobile application	1.)search "curry" 2.)view "curry fish" for 2 minutes 3.)view "curry laksa" 4minutes 4.)view "curry laksa" reviews
	background services	send and receive data from prediction model engine	suggestion notification, user log tracking data.
Prediction model engine	data pre-process	receive data and perform data cleaning	filter low likelihood data
	data training and prediction module	feed data into prediction model and make prediction	predict user A love "curry"
	push notification module	search food or restaurant with "curry" tagging or comment.	suggest "Ah Keong curry fish head" to user A

table 3.4

3.2.2. Development Environment

Table 3.5 below was the software and hardware tools will be using in this project

IDE for development	Netbean IDE and android studio
Programming language	Java, SQL language
Database	MySQL
Programming Library tools	JavaML (java machine learning)
Hardware for development and testing	HTC E9+ (android 5.0.2), Laptop (windows 10 ,core i5, 8GB ram)

table 3.5

3.2.3. Operation Environment

Since this is a research project and is not necessary using high computation power hardware to deploy. The prediction model engine will be deploy on my personal computer and the mobile application will be deploy on my smartphone (HTC E9+) and the connection is using smartphone hotspot to connect and communicate with my personal computer. Table 3.6 below was the specification of hardware for this project.

	personal computer (laptop)	smartphone (HTC E9+)
processor	Dual core core i5, 7400M 3.1GHZ	Mediatek Octa core 2.0GHz
ram	8GB DDR3	3GB
OS version	Windows 10	Android 5.0.2
data communication	Wifi, bluetooth	Mobile Hotspot, Wifi, 4G LTE, bluetooth

table 3.6

3.2.4. External Interface Requirements

The library interface that i planned to use are Java WiFi peer-to- peer API or Java Bluetooth API, I will just choose either 1 to implement in this project for communication between pc and smart devices. Table 3.7 below was the minimum requirement of both API.

Java API	Minimum Android API level
WiFi peer-to-peer	Android 4.0 (API level 14)
Bluetooth	no limit
Bluetooth Low Energy	Android 4.3 (API level 18)

table 3.7

3.2.5. Non-functional Requirements

- easy to use

- high engagement with user
- make user feel enjoy and satisfy

3.2.6. Functional Requirements

- suggest relevant content to user
- high accuracy of prediction
- minimize noise of data
- set reasonable timer threshold for likelihood number determination

3.2.7. Discussions

This is the first project that including a smartphone application which need a back-end supporting engine (prediction model engine) to support during runtime. The few main concern in coding are about : mobile application coding, prediction model coding, background communication services coding on both device.

The problem that I facing in this project are the communication between pc and smart devices. I have discover 2 method to communicate both device which is using Bluetooth and WiFi technology.[25][26] I need to study both method and choose which 1 is more easier to implement and high efficient. Luckily this is a research project, we are not stick to certain rules or regulation, we are able to choose our own way to implement the communication service.

3.3. Chapter Summary and Evaluation

Summary list below was stated the main section that I have discover within this chapter.

- Methodology
    - Extreme programming
        - keep in touch with customer by interact with end user in each deliver of development.
    - Fact gathering
        - tracking user activity history and determine the likelihood numbers based on the time period user viewing a restaurant/ food.
        - using Google Form to perform questionnaire
    - Fact recording
        - ERD diagram (table design)
        - DFD diagram (user and back-end engine)
    - Fact analysis
        - data preprocessing (data cleaning)
- Requirement analysis
    - Project Scope
        - Mobile application
            - likelihood generator
            - log tracking module
            - background services
              Prediction model engine
            - data pre-process
            - data training and prediction module
            - push notification module

          Development environment
              IDE : Netbean, Android studio
              language : JAVA, SQL command
              database: MySQL

          Operation environment

personal computer (Laptop, i5 dual core,8GB RAM, windows 10)

android smartphone (HTC E9+, octa core,3GB RAM,android 5.0.2)

External interface requirement

Java WiFi library interface ,android 4.0 (API level 14)

Java bluetooth (LE) library interface, android 4.3 (API level 18)

- Non functional requirement
  - easy to use
  - high engagement with user
  - make user feel enjoy and satisfy
- Functional requirement
  - suggest relevant content to user
  - high accuracy of prediction
  - minimize noise of data
  - set reasonable timer threshold for likelihood number determination

Chapter 4

System Design

4. System Design

4.1 Application Development Project and Package Implementation Project

My partner Pang Wai Kian will be design the mobile application UI and I will focusing in back end prediction model engine. Since this is a research project, we are planning to deploy the back end prediction model engine on our own laptop and communicate with mobile phone via wifi hotspot. User activity history data will be send from mobile smart device and store into laptop MySQL database. Prediction model engine will fetch data from MySQL database and process the user data. After processed user data, the suggestion notification will send to user smart device via mobile wifi hotspot. Diagram 4.1 below was the data flow and data connection between back end engine (prediction model engine) to our mobile phones application.

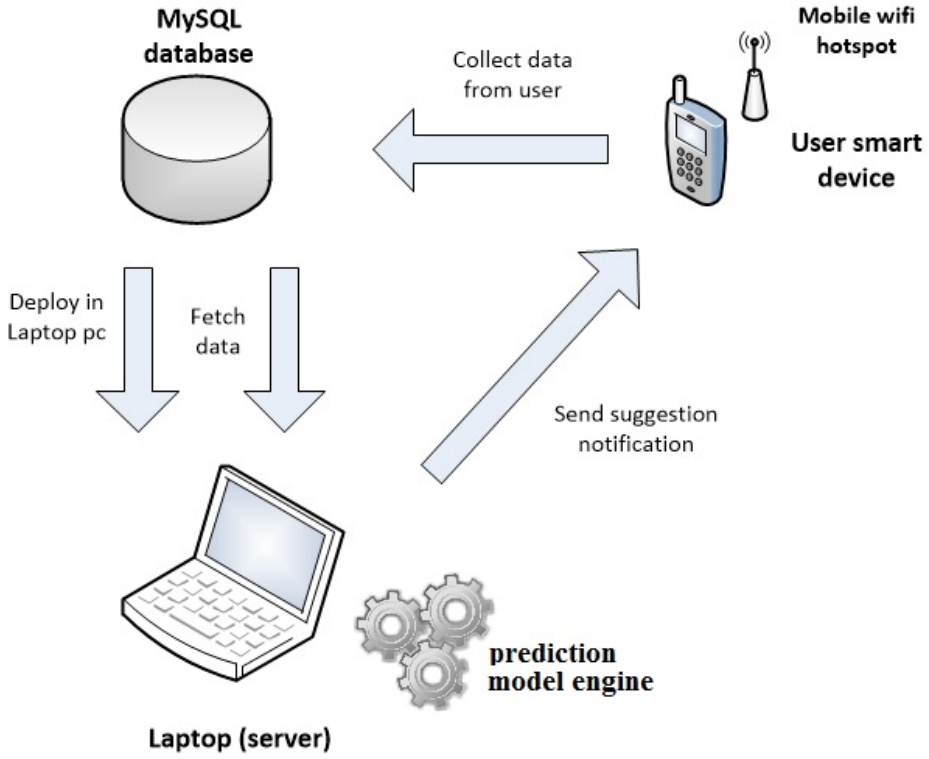
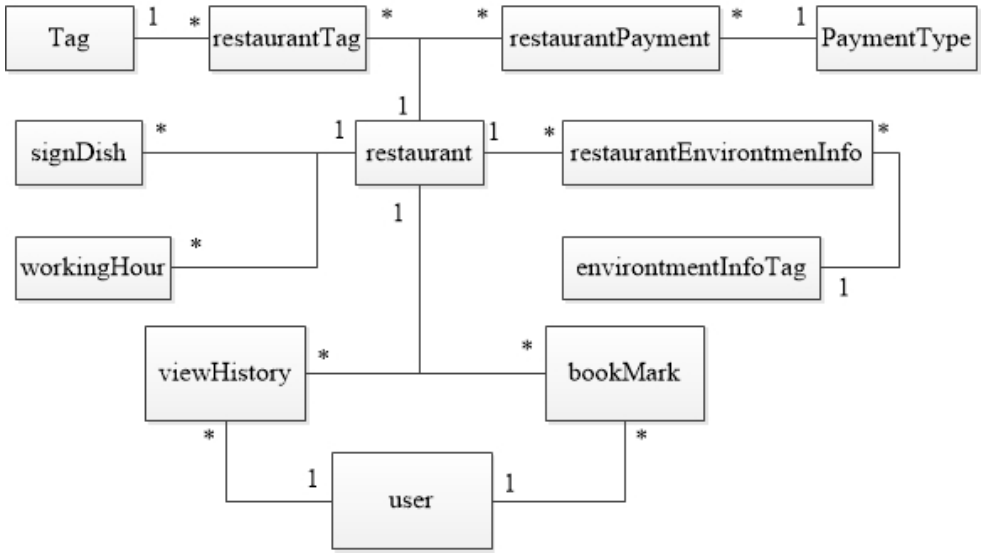


Diagram 4.1

ERD diagram 4.2 below was the full database design that i planning to implement into our food hunting system. For more detail table variable can be refer to previous chapter on Chapter 3.



ERD diagram 4.2

4.2.1. Pros and Cons of the system design

The prediction model will send suggestion notification to user smart device in daily basis may be 3 times(breakfast, lunch and dinner time) or more (tea time, supper) in each day. This solution is more user friendly is because this system will auto generate suggestion notification to user to suggest user to visit which restaurant but at the same time it will cause system overhead during peak period such as breakfast, lunch and dinner period because need to process thousands of user history to send suggestion to thousands of user. If our user is over million it may happen server down in this situation.

4.2.2. Alternative solution

We may try to reschedule the period to process user activity history. Example: during system is not fully utilize, the system will process user history and make prediction and store the prediction info into database first, after that in peak hour (meal time) only fetch the suggestion notification info from database and send to user directly.

4.2. Research Project

In this project we will just demonstrate how the prediction model works perfectly to predict what user want and make user feel satisfy which feels like the prediction model was understand user very well. Since this is a food hunting application which need malaysia user historical data is not available, we will try to prepare a questionnaire to collect similar data pattern to show how the prediction model. after collect all data from user, I will perform data cleaning process as usual to make sure all data is valid and accurate. After that will perform data mining process and discover which variable is correlated to each other. Next step will be data mining process, we will fit all the sample data into prediction model and train all sample data. Now we can feed new data into trained prediction model to make prediction to predict the user preference. Diagram 4.3 below was the whole process from raw data to make prediction.

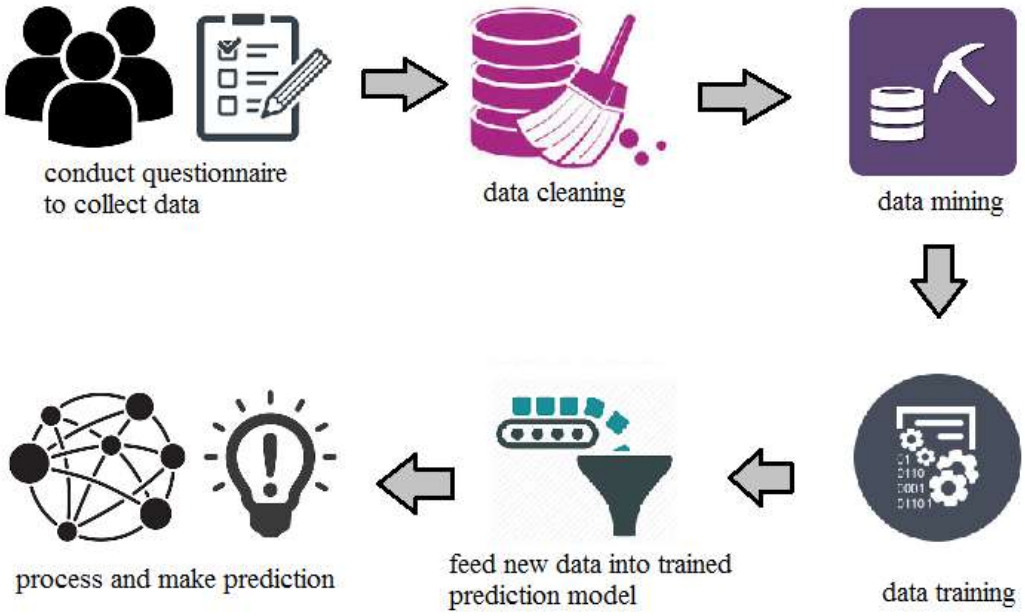


diagram 4.3

4.2.3. Pros and cons of research framework

I using questionnaire to collect data is simulating as the real application to understand the user how likely to visit what kind of restaurant or food. This simulation can prove our algorithm whether it can perform prediction well or not. on the other hand, the cons is the choice to let user to choose inside questionnaire is very limited. We cannot fit all kind of food and restaurant into the questionnaire to let user fill in for us is because the longer the questionnaire, the higher chance to done question incorrectly because user may skip the question by simply enter incorrect data.[27] We will try to adjust the sweet spot of the number of the questions and is enough for data analytics and making prediction.

4.3. Chapter Summary and Evaluation

The whole system design mainly follow to data science prediction model framework which is follow the sequence of: collect data> data cleaning> data mining > train data > feed new data >make prediction. We will repeat the process until we have discover the best sample data and choose the best prediction algorithm.

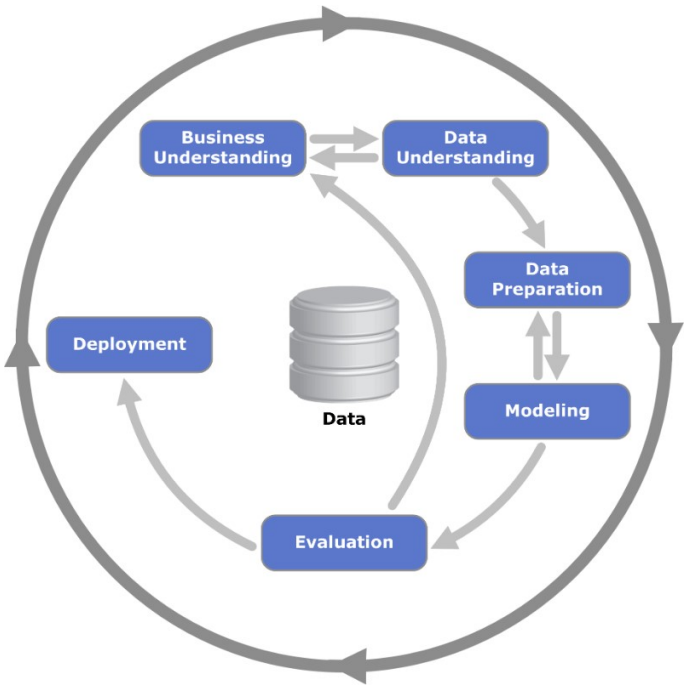
Since we have not yet start develop the mobile application and the back end prediction model engine, we cannot measure the performance of the whole system whether the prediction model can perform well or not in terms of computation resources and processing time. Data source for training the prediction model was the main problem, because we doesn't have the real data but we collect similar data by distribute questionnaire to public.

Although we have done a lot of researching paper works, we cannot ensure the whole future system will be implement as we planned which have already discussed from chapter 1 to chapter 3. We just can try our best to develop and implement our future system same as project planning as possible.

5. Appendix

[Appendix - part A]

Process model for building prediction model[7]



phases	description
business understanding	analyze and understand the business logic and the policy works in business perspective.
data understanding	understand all the available data and understand the relationship in between all available data in-order to perform data mining in next following stages.
data preparation	perform data cleaning to ensure all data have quality and can produce higher accuracy of prediction.
modeling	select the most suitable modeling techniques and apply on it. such as apply K-nearest neighbour , ANN and other techniques. observe the results of each techniques before making final decision.
evaluation	evaluate the built model in different perspective such as accuracy, performance and productivity. we should make sure the model is not over-fitting else will just perform well in testing phases but cannot deploy and use it in real life situation.
deployment	deploy the model into real life products.(food hunt application)

[Appendix- part B]

Example data for table design

Restaurant table data

features	value
restaurant name	Hanbing Korean Dessert Cafe
restaurant rating	3.5
no.bookmarked	239
area	bangsar
number branches	3
address	33, Jalan Telawi 3, Bangsar Baru, Bangsar, 59100
price range	21-40
food tag	korean, noodles
signature dish	Green Tea Snow Ice/ Mango Snow Ice /Cheese Topokki /Korean Honey Fried Chicken/ Seafood Ramyeon /Blueberry Smoothie
working hours	Mon-Thu 12:00-00:00 Fri-Sat 12:00-01:00 Sun 12:00-00:00
payment method	visa,master cash
no.good review	1
no.satisfy review	2
no.bad review	0
other info	air-conditioned,wifi,pork-free,open till late
no.photo	45

View history table data

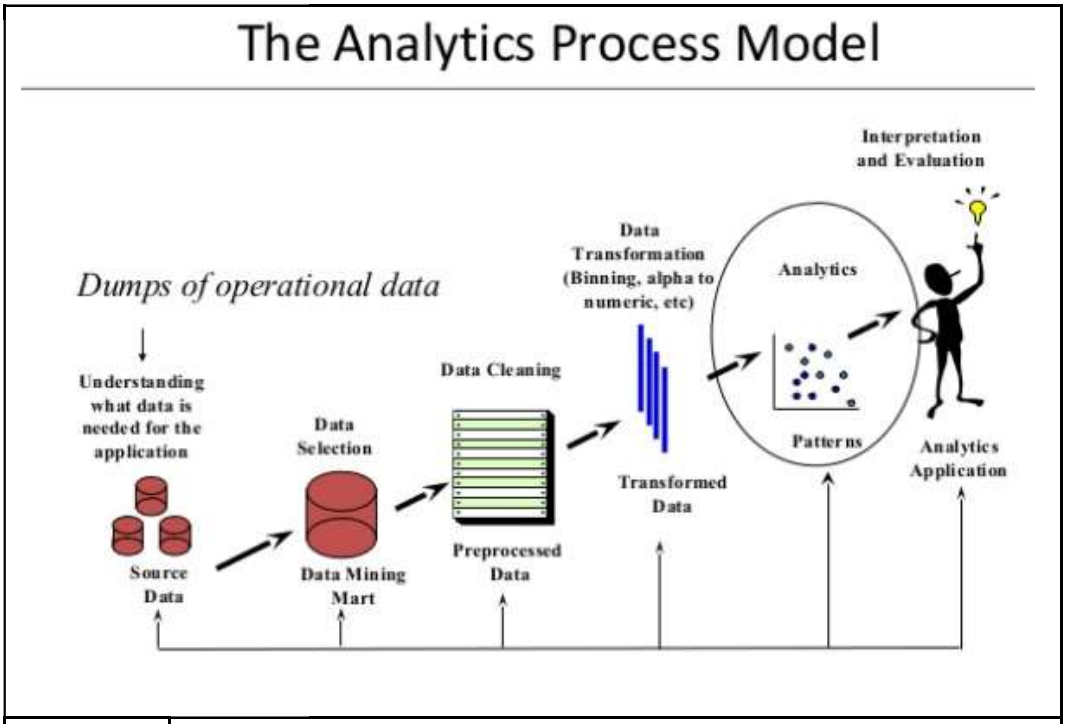
features	value
timer on view	30sec
viewed restaurant	res001 (restaurant id)
likelihood	0.8

User profile table data

features	
gender	male
birthday	1996-3-16
area stay	setapak
bookmark	old town,jonker street
race	chinese
vegan	no

[Appendix - part C]

The Analytics Process Model [8]



source data	data from various form such as: SQL database, microsoft excel
data mining	discover the relationship of data and discover the data pattern.
preprocess data	data filtering process to ensure data have quality to minimize errors of data. Example: decide what to do with missing data, whether to discard outlier or other solution.
transform data	because different data will store in various form and format, we need to transform it into reasonable or process-able manner to process those data by using machine learning.
analytic	visualize the data information to easier understandable manner. example: histogram,barchart,piechart
analytic application (knowledge)	extract knowledge from those visualized information. understand what is the logic behind those information. example: we observed majority of customer buy fried chicken and coke together in the analytic stage, so we understand (knowledge) fried chicken and coke is the perfect match and will increase the sales revenue when we put it together.

6. References

[1] Holmberg, L. and Vickers, A., 2013. Evaluation of prediction models for decision-making: beyond calibration and discrimination. *PLoS Med*, *10*(7), p.e1001491.

[2] Padhy, N. and Panigrahi, R., 2012. Data Mining: A prediction Technique for the workers in the PR Department of Orissa (Block and Panchayat). arXiv preprint arXiv:1211.5724.

[3] Furness, D. (2017, January 17). Artificial intelligence can now predict heart failure, and that may save lives. Retrieved from Digital Trends: http://www.digitaltrends.com/computing/ai-heart-disease/

[4] Wells, D. (2013, October 8). *Extreme Programming: A gentle introduction*. Retrieved 6 1, 2017, from Extreme Programming: http://www.extremeprogramming.org/

[5] Joachims, T., 2002, July. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 133-142). ACM.

[6]Leuven, A. (2013, March 28). *AFC Café: Bart baesens*. Retrieved June 19, 2017, from LinkedIn: https://www.slideshare.net/AFCLeuven/bart-baesens-3pdf

[7]Rouse, M. (2016, November). *predictive modeling*. Retrieved June 19, 2017, from TechTarget: http://searchdatamanagement.techtarget.com/definition/predictive-modeling

[8]Rouse, M. (2016, October). *predictive analytics*. Retrieved June 19, 2019, from TechTarget: http://searchbusinessanalytics.techtarget.com/definition/predictive-analytics

[9]Chew, N. (2015, March 15). *Top 5 Malaysian Food Apps*. Retrieved June 19, 2017, from mobile88: http://www.mobile88.com/news/feature/top-5-malaysian-food-apps/

[10]*About Us*. (n.d.). Retrieved June 20, 2017, from Openrice: http://www.openrice.com/info/corporate/eng/about-us.html

[11]Gray, K. (2015, October 7). *Demystifying Predictive Analytics*. Retrieved June 20, 2017, from LinkedIn: https://www.linkedin.com/pulse/demystifying-predictive-analytics-kevin-gray

[12]RAY, S. (2015, October 6). *Understanding Support Vector Machine algorithm from examples (along with code)*. Retrieved June 20, 2017, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/

[13]Siganos, C. S. (n.d.). *NEURAL NETWORKS* . Retrieved June 20, 2017, from https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

[14]SRIVASTAVA, T. (2014, October 20). *How does Artificial Neural Network (ANN) algorithm work? Simplified!* Retrieved June 20, 2017, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2014/10/ann-work-simplified/

[15]*Artificial Neural Networks Technology*. (n.d.). Retrieved June 23, 2017, from DoD DACS: http://www.psych.utoronto.ca/users/reingold/courses/ai/cache/neural2.html

[16]SRIVASTAVA, T. (2014, October 10). *Introduction to k-nearest neighbors : Simplified*. Retrieved June 23, 2017, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2014/10/introduction-k-neighbours-algorithm-clustering/

[17]RAY, S. (2015, September 13). *6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python)*. Retrieved June 23, 2017, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/

[18]SRIVASTAVA, T. (2014, June 10). *Introduction to Random forest – Simplified*. Retrieved June 23, 2017, from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/

[19]Körting, T. S. (2014, April 4). *How Random Forest algorithm works*. Retrieved June 24, 2017, from YouTube: https://www.youtube.com/watch?v=loNcrMjYh64

[20]Cole, S. (2016, October). *Burritos in San Diego*. Retrieved June 27, 2017, from kaggle: https://www.kaggle.com/srcole/burritos-in-san-diego

[21] Wells, D. (8 October, 2013). *Extreme Programming:*. Retrieved 10 7, 2017, from Extreme Programming: http://www.extremeprogramming.org/

[22]Rouse, M. (August, 2008). *DEFINITION Extreme Programming (XP)*. Retrieved 10 7, 2017, from TechTarget: http://searchsoftwarequality.techtarget.com/definition/Extreme-Programming

[23] Retrieved 13 7, 2017, from https://developer.android.com/guide/topics/connectivity/bluetooth.html#SettingUp

[24] Retrieved 13 7, 2017, from https://developer.android.com/guide/topics/connectivity/wifip2p.html

[25] Patel, S. (25 November, 2016). *COMMUNICATION OVER WIFI IN ANDROID*. Retrieved 13 7, 2017, from yudiz: http://www.yudiz.com/communication-over-wifi-in-android/

[26] Bevilacqua, F. (12 August, 2013). *Building a Peer-to-Peer Multiplayer Networked Game*. Retrieved 13 7, 2017, from Envatotuts: https://gamedevelopment.tutsplus.com/tutorials/building-a-peer-to-peer-multiplayer-networked-game--gamedev-10074

[27] Debois, S. (16 3, 2016). *9 Advantages and Disadvantages of Questionnaires*. Retrieved 20 7, 2017, from SurveyAnyPlace: https://surveyanyplace.com/questionnaire-pros-and-cons/