

Learning the student's happiness model

Third Year Project Report

By: Qifan Zhu

Supervisor: Dr. Xiaojun Zeng

May 2019

The University of Manchester, School of Computer Science
BSc Computer Science

Contents

1. Project context.....	5
1.1 The background of the project.....	5
1.2 Approach of the project.....	5
1.3 Project objectives	6
1.4 Motivations	6
2. Background research	7
2.1 Factors of happiness.....	7
2.2 Machine learning algorithms used	8
2.2.1 Linear regression.....	8
2.2.2 SMO Regression	10
2.2.3 REP tree	10
2.3 Tools used in project.....	11
3 Methodology and Experiment	13
3.1 Design of the project	13
3.2 Factor selection	13
3.3 Questionnaire design.....	14
3.4 Data collection	14
3.5 Machine Learning Approaches.....	15
3.6 Modelling Experiments.....	16
4 Analysis and evaluation	22
4.1 Importance of factors	22
4.2 Comparison of models.....	23
4.3 Comparison of factors among different groups	24
4.4 Comparison of results with student's approximation	25
5 Reflection and conclusion	26
5.1 Timing and Milestones.....	26
5.2 Reflection on the project	26
5.3 Conclusion.....	27
Reference	28

Abstract

The aim of this project is to quantify the importance of factors of happiness in university students and build a mathematical model which presents the relationship between factors and an overall happiness of students using machine learning method. The data for learning was collected from 202 responses of a questionnaire. The report also describes how the data was used to build different models and calculate the accuracy of models. It describes the differences and similarities of models for different group of people and choose the most accurate model. Finally, it compares the quantified importance of factors collected directly from questionnaire and the importance shown by the model.

Acknowledgement

I would like to thank my supervisor, Xiaojun Zeng, for his guidance on the project this year.

I would like to thank every student who help me complete the questionnaire.

I would like to thank my friends for their support and advice.

1. Project context

1.1 The background of the project

Happiness is one of the ultimate pursuits for many people in the world. Learning the happiness itself is called 'eudemonics', one of the main purposes of the subject is to find out what makes people feel happy. In modern philosophy, Utilitarianism is a classic model of pursuit of happiness. Since the 1960s, many science subjects carry out the research about happiness, including social psychology, happiness economy, medical research... The author of '**THE WISDOM OF LIFE ON HUMAN NATURE**', philosopher Arthur Schopenhauer said, the wisdom of life is the art of spending a lifetime as happily and cheerfully as possible. This project investigates happiness for students in university, by using machine learning techniques.

1.2 Approach of the project

This project is split into 3 parts to find out the factor of student's happiness.

1. Data collection

The aim of this part is to collect over 200 responses of questionnaire from university students for training the model. The data structure for each response is:

1. the information the student, e.g. gender, age.
2. A vector of indexes of factors of happiness which indicates to what extent does each factor affects their happiness
3. A index of happiness, which the interviewee considered himself to have.
4. A vector of indexes of importance of factors of happiness, which is their approximation of the importance of each factor.

2. Build the model

the aim of this part is to use different machine learning algorithm to build mathematical models of student's happiness from dataset. The structure of a completed model is:

Inputs:

The vector of factor of student's happiness.

Output:

A predicted happiness index.

3. Analysis

The aim of this part is to compare the accuracy of each model and analysis the difference of model of different group of students. Then compare the importance of factors from machine learning method and student's approximation.

It is an even-handed way to use machine learning method to investigate the importance of each factor of happiness. Because it is all determined by training set which collected from students, there is no subjective factor that may cause bias from the author.

1.3 Project objectives

The project has four main objectives:

1. Build mathematical models which shows the relationship between factors of happiness and the happiness index.
2. Compare the accuracy of each model built by different machine learning algorithms and determine the most accurate one.
3. Analysis the difference and similarities of importance of factor for different groups of people.
4. Analysis the difference of 'importance of factors' collected from student's approximation and calculated from the machine learning techniques.

1.4 Motivations

The project is being chosen for several reasons:

1. There are many studies on factors of happiness for people, but fewer specifically for students.
2. The importance of factors of happiness calculated by machine learning techniques has no subjective bias.
3. Student is a relatively confused group of people compare to older groups. It is more important for them to understand themselves better and pursue happiness.

2. Background research

For this project, three stages of the approach need research.

Firstly, the design of questionnaire needs a lot of background research. To design a non-biased questionnaire, it is necessary to investigate the template which minimize the subjective bias from the author. Also, the selection of factors requires research on psychology and philosophy, to choose the most appropriate factors to represent the student's happiness.

Secondly, the machine learning part of the project needs technical research on several machine learning algorithms, functions and tools.

Thirdly, the analysis of difference between different group of people need research of psychology, to justify the reason of the difference.

2.1 Factors of happiness

Happiness economics is a new research field investigate the relationship between happiness and economics. It indicates the 7 main factors of happiness are: individual values, individual freedom, health, friends and communities, work, family relationship, financial condition. Maslow's hierarchy of needs is a theory in psychology proposed by Abraham Maslow in his 1943 paper "A Theory of Human Motivation" in Psychological Review.[1] It indicates people's needs have five hierarchies: physiological, safety, love/belonging, esteem and self-actualization

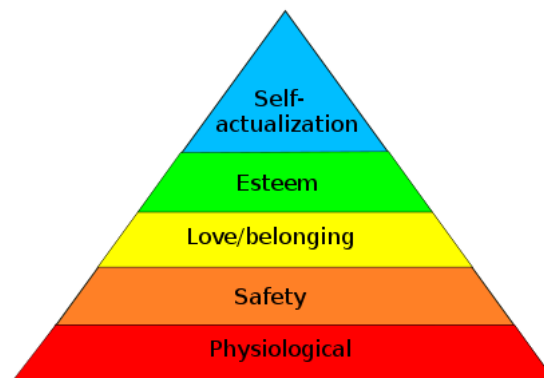


Figure1: Maslow's hierarchy of needs

At different stages, people have different needs. For example, for a person who have no safe place to live, living standard has more influence on his happiness than self-esteem. Because the research shows when current hierarchy is satisfied, all needs in this hierarchy is not important motivation factors anymore. People will always pursue next hierarchy of needs. In other words, healthy people usually would not feel much happy about their current healthy body, but sick people care more about their health.

In this project, since we choose a specific group of people, student, the factors of happiness can be more specific. According to research about student's happiness, the main factors of happiness for university students are: relationships formed, satisfaction about university environment, incomes, satisfaction about school work, extracurricular activities and health. In sum, students who had sufficient social relations were the happiest, which might not come as a big surprise. A far

noteworthy finding is that the level of satisfaction among students who can balance work and activities well and can meet deadlines or goals, or succeed in their studies, does not differ significantly from that of students who cannot perform these functions. [2]

Variable	Coefficient (std.err)	Wald (significance)
Age	−0.052 (0.118)	0.193
Gender	−1.012 (0.839)	1.154
Incomes (all)	0.000 (0.001)	0.200
Extracurricular activities	−1.579 (0.845)	3.496 *
Satisfaction with school work (achieve standard, AAS)	1.440 (0.859)	2.809 *
Satisfaction with school work (happy with marks, HWM)	−2.405 (1.010)	5.576 **
Satisfaction with school work (enjoy studying, ES)	0.485 (0.876)	0.306
Satisfaction with school work (interesting work, DWTI)	−0.642 (0.518)	1.538
Satisfaction with school work (cope with university work, COPE)	0.175 (0.608)	0.083
Satisfaction with resources and school environment	0.348 (0.389)	0.800
Relationships formed	1.249 (0.441)	8.025 ***
Time management (work balance, BWUAW)	0.802 (0.491)	2.665
Time management (meet deadlines, MD)	0.372 (0.470)	0.629
Time management (recreational time, SRT)	−0.051 (0.332)	0.024
Health	0.909 (0.673)	1.827
University reputations	0.265 (0.648)	0.167
Pseudo R ² , Cox and Snell	0.623	
Pearson goodness-of-fit, $\chi^2(83)$	128.919	
Model fitting, $\chi^2(16)$	33.159 ***	

Factors affecting student's happiness

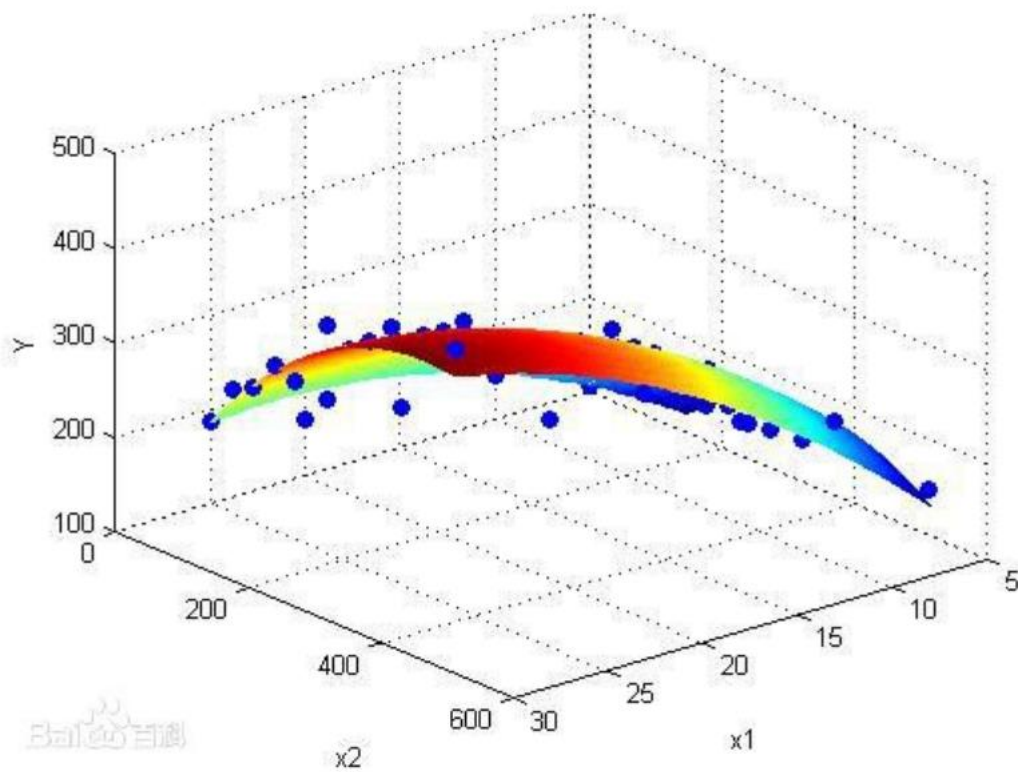
2.2 Machine learning algorithms used

This section is to introduce the machine learning algorithms used in this project.

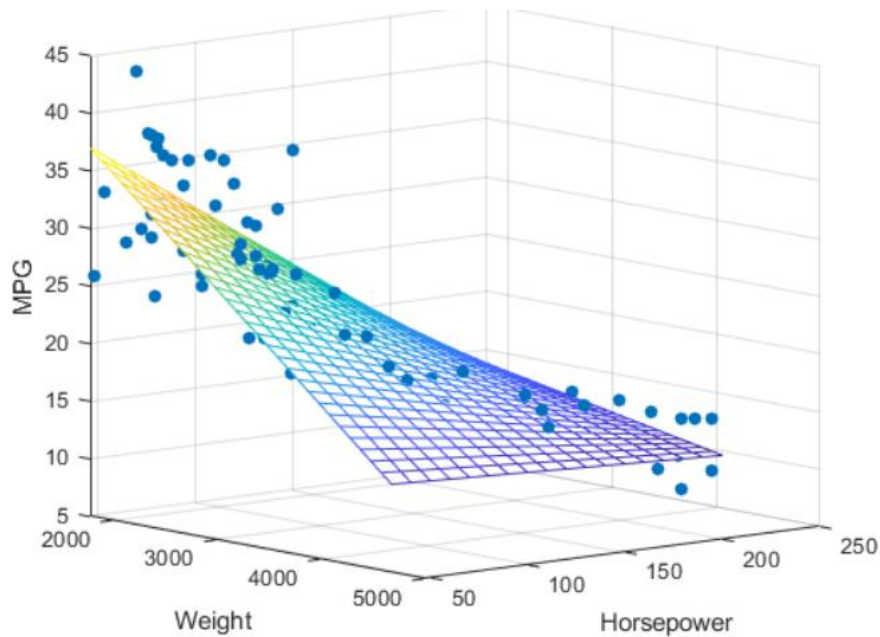
2.2.1 Linear regression

Linear regression is a statistical analysis method to investigate the relationship between two or more variables using regression analysis in mathematical statistics. The formula is: $y = w'x + e$, which e represents the error. [3] y is dependent variable and x is the independent variable. The aim of linear regression is to find out the best fitting line to represent the relationship between dependent variable and independent variable.

In this project, there are several independent variables (which are the factors of student's happiness) and one dependent variable (which is the overall happiness index).



Multi-variable linear regression model1



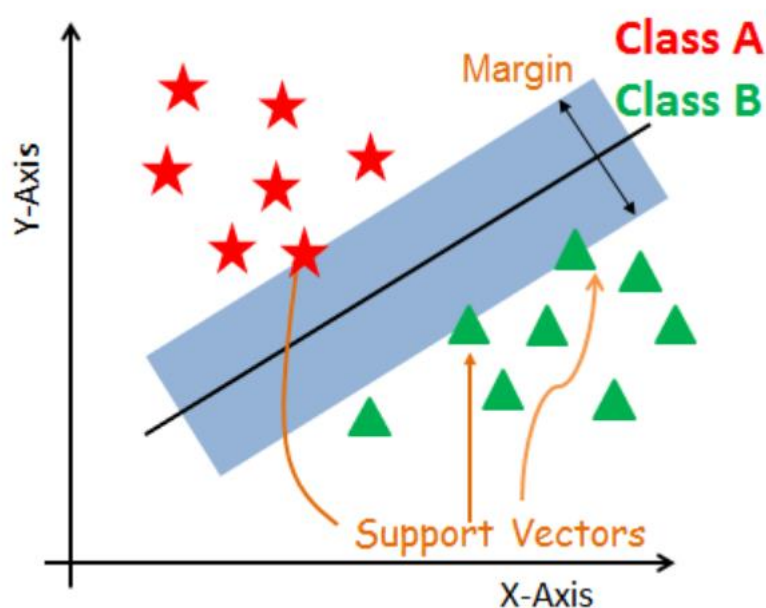
Multi-variable linear regression model [5]

In this project, multi-variable linear regression is a suitable model to represent the relationship between factors and the overall happiness of students. In this case, the coefficient of factors is the result. For each factor, coefficient of the factor reflects the importance of factor. The formula of multi-variable linear regression is:

happiness index = $w_1 \cdot \text{health} + w_2 \cdot \text{relationship} + w_3 \cdot \text{friend} + w_4 \cdot \text{finance} + w_5 \cdot \text{activity} + w_6 \cdot \text{grade} + w_7 \cdot \text{learning resource} + w_8 \cdot \text{interest on course} + w_9 \cdot \text{future job} + w_{10} \cdot \text{time management} + w_{11} \cdot \text{living condition} + w_{12} \cdot \text{self-esteem}$

2.2.2 SMO Regression

Sequential minimal optimization (SMO) is an optimization algorithm used during the training of SVM (support vector machine). The SMO algorithm is closely related to a family of optimization algorithms called Bregman methods or row-action methods. These methods solve convex programming problems with linear constraints. They are iterative methods where each step projects the current primal point onto each constraint. [4] Given a set of training instances, each training instance being marked as belonging to one or the other of the two categories, the SVM training algorithm creates a model that assigns a new instance to one of the two categories, making it non-probabilistic Meta linear classifier. The SVM model represents an instance as a point in space, such that the mapping separates the instances of the individual categories by as wide a clear interval as possible. Then, map the new instances to the same space and predict which category they belong based on which side of the interval they fall on.



Support vector machine example

2.2.3 REP tree

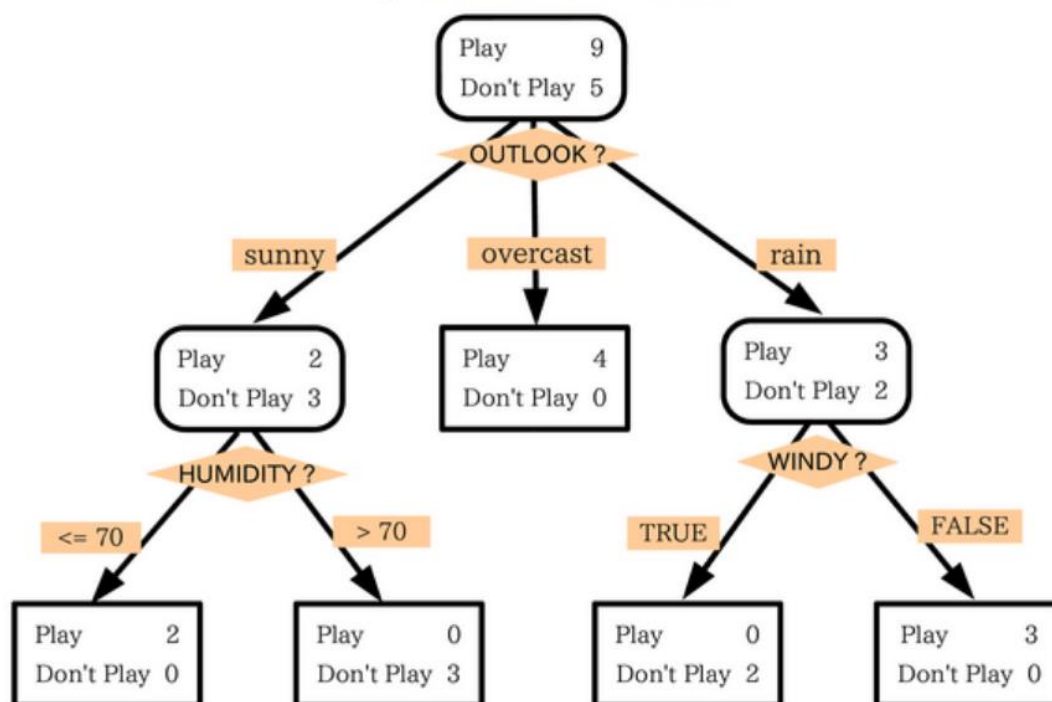
In machine learning, a decision tree is a predictive model; it represents a mapping between object properties and object values. Each node in the tree represents an object, and each forked path represents a possible attribute value, and each leaf node corresponds to the value of the object represented by the path from the root node to the leaf node. The decision tree has only a single output. If you want to have a complex output, you can create an independent decision tree to

handle the different outputs. Decision trees in data mining are a common technique that can be used to analyse data and can also be used for forecasting.

Decision tree learning is an ordinary method in data mining. Each decision tree represents a tree structure, which is branched by its branches to rely on attributes for that type of object. Each decision tree can rely on data testing of the partitioning of the source database. This process can recursively trim the tree. The recursion process is complete when no further segmentation or a separate class can be applied to a branch. In addition, the random forest classifier combines many decision trees to improve the accuracy of the classification.

In weka, it developed the REP tree algorithm which is a fast decision tree learner. Builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with backfitting). Only sorts values for numeric attributes once. Missing values are dealt with by splitting the corresponding instances into pieces (i.e. as in C4.5). [8]

Dependent variable: PLAY



Example of decision tree

2.3 Tools used in project

Weka (Waikato Environment for Knowledge Analysis):

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to these functions. [9] The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modelling algorithms implemented in other programming languages, plus data

pre-processing utilities in C, and a Make file-based system for running machine learning experiments. This original version was primarily designed as a tool for analysing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, for educational purposes and research. [10]

Microsoft Excel

Microsoft Excel is a powerful spreadsheet software used in data analysis and machine learning. In the project, it is used to manipulate data collected from responses in SmartSurvey and record the result from WEKA. The datatype of files exported from SmartSurvey website is CSS format. It is Cascading Style Sheets which is able to be manipulate directly from WEKA. It also used to manipulate the result in table to highlight the difference and similarity of the model.

Java

Java is a general-purpose programming language that is class-based, object-oriented, and designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA), meaning that compiled Java code can run on all platforms that support Java without the need for recompilation. [11] In this project, because the machine learning tool we used is written in java. So, JAVA must to be used to realize some specific function for achieve the objectives of project.

SmartSurvey

Smartsurvey is an online questionnaire builder and publisher. In this project, the questionnaire is built on this website. The responses can be easily collected from students who click the shared link and answer all the questions. The data can be exported directly from the website in CSS format which is compatible for the machine learning tool. Also, the website provides the service of visualize the statistics, filtering data before export, convenient survey organizing user interface.

3 Methodology and Experiment

This part of report aims to describes the creation of the achievements, including the general plan and approach of each stage.

3.1 Design of the project

The implementation of the project is split into 4 stages: factor selection, questionnaire design, data collection and modelling.

- **Factor selection**
The factors will be selected based on background reading of psychology and philosophy, with some interview of students.
- **Questionnaire design**
Design of questionnaire requires background reading to avoid bias and guarantee the accuracy of the data collected.
- **Data collection**
The data is collected by responses from students who complete the questionnaires. This stage requires no technical skill, but time consuming.
- **Modelling and analysis**
Using different machine learning algorithm to manipulate the data collected to build different models of student's happiness.

3.2 Factor selection

The factor selection needs background research on happiness itself and factors of happiness specifically for students. Several students from different university was interviewed by the author and given their advices for factor selection.

Health	Relationship satisfaction	Friendship satisfaction	Finance condition	Activity satisfaction	Grade satisfaction
Resource available	Interest on courses	Future job availability	Time management	Living condition	Self-esteem

12 factors selected

As a result, 12 factors were chosen based on literature review and interview of students. Based on research about people's happiness, heal, relationship satisfaction, friendship satisfaction, finance condition, living condition, self-esteem was selected. Based on research about university students' happiness, activity satisfaction, grade satisfaction, learning resource available, interest on courses, future job availability, time management was selected. The 12 factors were also checked with Maslow's hierarchy of needs, to ensure the correctness and comprehensiveness for themselves to be a factor of happiness. For which health, living condition, finance condition belongs to physiological stage. Living condition, belongs to safety stage. Relationship satisfaction, friendship

satisfaction belongs to love/belonging stage. Self-esteem, grade satisfaction belongs to esteem stage. Resource availability, interest on courses, future job availability, time management belongs to self-actualization stage. The number of factors selected is 12 because this number of factors would not require a time-consuming questionnaire.

3.3 Questionnaire design

One of the objectives of questionnaire design is to minimize the bias of answer caused by a subjective bias of students. To achieve the objective, background research shows ESS well-being module design template [12] is a comprehensive template to refer. Because the field of the template is well-being, which is similar to the research area of this project. The questionnaire was spitted into 2 parts, first part is the information of students. The students would answer the questions about their information, this part aims to determine their group. For example, the information part would ask some questions about their level of year, gender, country of birth... The second part is feature value part. The students were firstly asked a question about their overall happiness, to let themselves evaluate their happiness levels. Then they were asked a set of questions about their satisfaction on each factor of happiness, and a set of questions about their approximation of importance of each factor. The reason of this organization is to minimize the subjective bias when they think about their overall happiness, because setting the overall happiness to be the first question minimize the bias caused by answering the questions about factors of happiness. For example, if the student who recently be lovelorn, after he answered the question about the relationship satisfaction, he might immerse himself in a sad atmosphere and choose a lower overall happiness option than he really was.

Another objective of the questionnaire design is to minimize the time requires to complete it. The platform of the questionnaire is Smart Survey which is an online questionnaire builder. It can build and edit the questions online and publish the survey with a shared link. It minimizes the time of both interviewer and interviewee. Because the whole process of completing the questionnaire can hold online. The students can complete the questionnaire either on their mobile devices or personal computer, because the platform has compatibility of multiplatform.



3.4 Data collection








Data collection is the most time-consuming part of the implementing of the project. Because there is a large amount of response required to modelling and analysis. So, the objective of this stage of project is to get at least 200 responses. There are 2 sub-objectives in this part.




Firstly, the comprehensiveness of groups of people. In other word, the project requires not only large amount of responses, but also requires diversity of group of people. Because the modelling part of the project requires enough amount of data for each different group of people, such as male, female, year 1 students, year 2 students... The shared link of the questionnaire was published mainly on social networks. To guarantee there are enough data for each year of students for analysing, the link was sent to different social network groups in each year, and different school of university.

Secondly, the validity of data collected. In other word, the reliability of each responses. Some students may roughly complete the survey and not even read the questions. The purpose of

students to roughly complete the survey is the key to this problem. These students are not willing to complete the survey for their self-interest or kindness, they complete the questionnaire because of pressure. So, the questionnaire link did not share one-to-one but one-to-many, the link shared only in social network groups and the process of completing the questionnaire was anonymous. Also, completing the questionnaire would not get any reward. This method might result in low efficient of data collection, but the quality of response of data will be much better. Finally, the data cleaning proceeds on the platform of SmartSurvey which filter out any responses that takes faster than 1 minute to finish. As a result, the probability of the situation of 'Utilitaly complete' was minimized.

			Response Percent	Response Total
1	Male		59.31%	121
2	Female		40.69%	83
			answered	204
			skipped	0

			Response Percent	Response Total
1	1st year undergraduate		29.90%	61
2	2nd year undergraduate		14.22%	29
3	3rd year undergraduate		34.80%	71
4	4th year undergraduate		13.73%	28
5	postgraduate		5.39%	11
6	PHD or higher		0.98%	2
7	other		0.98%	2
			answered	204
			skipped	0

			Response Percent	Response Total
1	Science and Engineering		60.70%	122
2	Humanities		35.32%	71
3	Biology,Medicine and Health		3.98%	8
			answered	201
			skipped	3

The result of data collection

After the period of data collection, 204 valid responses were exported, and these data would be used to train the model.

3.5 Machine Learning Approaches

This is the data analysis part of the project. After getting the data from responses, the CSS format file was exported. The machine learning tool used in this project is WEKA, three machine learning algorithms would manipulate the dataset and three different machine learning models get trained. After the model was built, the accuracy of the model would be calculated and be compared. Finally, there was a comparison of models of different group of students.

machine learning algorithms

The project selected three machine learning algorithms which is suitable for learning student's happiness model. The objective of the model is to presents the relationship between the factor of happiness and student's overall happiness. As a result, the model can be used to predict the student's happiness index by inputting the index of factors of happiness. Three algorithms were chosen to learn the student happiness model: Linear Regression, SMO regression, REP tree.

Linear Regression

Model of linear regression can directly show the mathematical relationship between factors of happiness and overall happiness by coefficient for each feature. That achieves the objective of building model of student's happiness. Also, multi-linear regression is a relatively clearer approach to show the relationship compare to other algorithms, so the accuracy of linear regression can be an important reference.

SMO Regression

SMO regression model is able to build non-linear model to solve the problem of optimisation. If the model of student's happiness is non-linear for some factors, this model would perform higher accuracy.

REP tree

REP tree is a specific algorithm based on decision tree developed by WEKA. This algorithm can highlight the factors which split the students in different happiness groups. So, the result of it also showed the importance of factors.

3.6 Modelling Experiments

As the result of data collection shows, the students can be split into 6 groups: female students, male students, year1-2 undergraduate students, year 3 students, students in school of science and engineering, students in school of Humanities. The reason for this organization is the lack of data for year 2 and year 4 undergraduate students. So, they were combined with responses from year 1 and year 3 students. For the other 4 groups: male, female, humanities, science and engineering, the responses are enough for data analysis.

The three algorithms were implemented on WEKA. The overall dataset for all students was cleaned by filtering out the rough-finished response and extracted from CSS format file. All the answers of the responses were displayed as numbers (0-10) and the details of respondent are deleted from the table. As WEKA only support numbers of the data it manipulated, the blank answer which was displayed as '-' was replaced by '' (null). CSS format file is able to be manipulated by Microsoft Excel as a format of table, so the group of people was split by 'filter' tool inside the Excel. There was a question asked about the gender of students, the answer format is 0 or 1, which 0 represents male and 1 represents female students. The data was then split into 6 groups, based on the answer of three questions about the group of students. In the example of gender of students, the overall dataset was split into 2 group, one for the answer of gender is '0', another for the answer of gender

is '1'.

MAPE (mean absolute percentage error)

To achieve another objective of the project: compare the accuracy of each model, the project uses the MAPE (mean absolute percentage error) as the index of accuracy. WEKA provides the calculation of MAE (mean absolute error) of each algorithm. But MAE itself can not be used as a standard of accuracy, because it is a value but not percentage. So, it is impossible to investigate the accuracy of one algorithm by using MAE only. MAPE provides the advantage to directly show the accuracy of the machine learning algorithm. The formula of MAPE is
$$M = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|,$$
 where A_t is the actual value and F_t is the forecast value. The difference between A_t and F_t is divided by the actual value A_t again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n . Multiplying by 100% makes it a percentage error. [13]

Regularly, the calculation of MAPE needs to output the model and forecast happiness index for every element.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	MAPE	error/actual	prediction	HAPPYNISS	health	relation	friend	finance	activity	grade	resource	interest	job	time	life	esteem
2	0.166742	0.02114	5.1057	5	7	8	8	1	10	1	5	2	7	1	3	5
3		0.06585	10.6585	10	10	6	10	1	10	10	4	10	1	10	10	10
4		0.19048	4.0476	5	5	10	5	10	6	3	10	6	8	6	7	4
5		0.0756125	8.6049	8	9	10	10	8	10	8	8	7	7	7	9	9
6		0.05792857	6.5945	7	8	10	8	10	7	7	6	3	10	6	7	7
7		0.05893333	6.3536	6	6	7	6	8	6	5	7	6	8	8	7	8
8		0.30716	6.5358	5	8	5	7	4	7	5	8	3	3	6	8	5
9		0.33456	6.6728	5	8	10	8	3	4	5	10	7	7	10	10	6
10		0.05172857	7.3621	7	8	9	8	2	8	8	6	6	9	6	8	8
11		0.0352125	7.7183	8	7	9	9	8	9	9	6	6	7	7	8	8
12		0.00922857	6.9354	7	7	6	7	5	6	6	7	7	7	7	7	8
13		0.00901429	6.9369	7	7	7	6	9	6	7	9	8	8	10	7	9
14		0.61215714	2.7149	7	4	4	3	2	7	4	6	5	10	5	4	3

Example of calculate the MAPE

In this project, since WEKA is an open source machine learning tool. It is able to modify the code to realize some functions. So, the WEKA is modified to calculate the MAPE in evaluation.java file. As a result, every time WEKA build the model, the MAPE will be evaluated and output.

```

=== Summary ===

Correlation coefficient          0.7417
Mean absolute error             0.8231
Mean absolute percentage error  0.1594
Root mean squared error        1.1324
Relative absolute error         61.4147 %
Root relative squared error     67.0048 %
Total Number of Instances      202

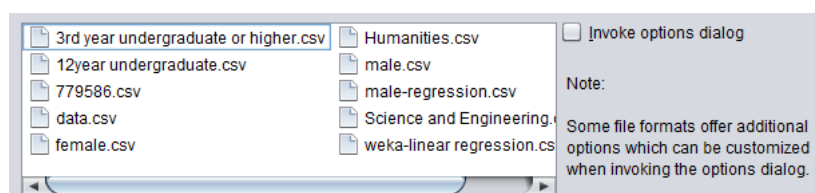
```

Example of output of evaluation

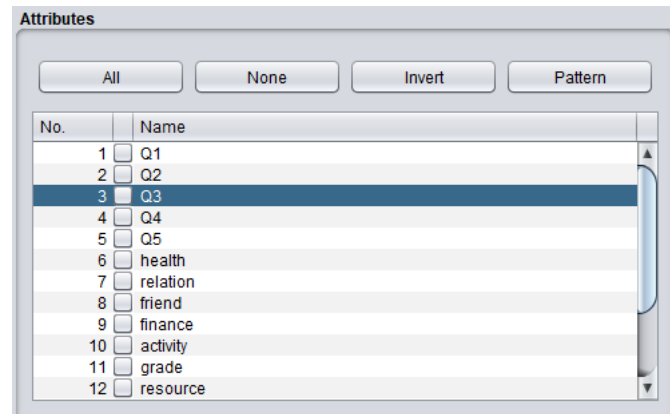
In this figure, 0.1594 means the mean absolute percentage error of the model is 15.94%.

Linear Regression

the first step is to select the data file.

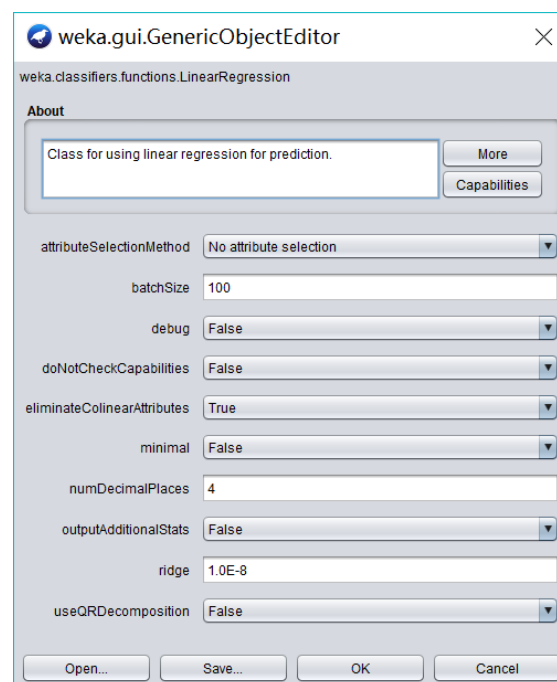
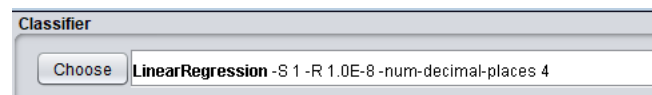


The next step is to select the attributes



In this project, the attributes Q1, Q2, Q3, Q4 are used to determine the group of people, and Q5 is the overall happiness index, the other attributes are all factors of happiness. So Q1, Q2, Q3, Q4 can be deleted before learning the model.

After the selection of attributes, choose the machine learning algorithm.



The linear regression algorithms in WEKA API has the function of attributes selection. It will select the attributes automatically based on the subset evaluation and greedy backwards search to select more important factor and ignore the relatively irrelevant factors.

After loading the dataset, the linear regression model is the output, together with mean absolute error.

```

Linear Regression Model

Q5 =

    0.0574 * health +
    0.2073 * relation +
    0.1304 * friend +
    0.0029 * finance +
    0.0653 * activity +
    0.0723 * grade +
    -0.0729 * resource +
    0.0442 * interest +
    -0.1506 * job +
    -0.0473 * time +
    0.1246 * life +
    0.3752 * esteem +
    0.9627
=== Cross-validation ===
=== Summary ===

Correlation coefficient          0.7417
Mean absolute error             0.8231
Mean absolute percentage error  0.1594

```

Example of result of linear regression model

datasets	number of students	testing condition	MAPE	Accuracy
total	202	10-fold cross validation	15.65%	84.3500%
female	81	10-fold cross validation	14.51%	85.4900%
male	121	10-fold cross validation	17.57%	82.4300%
Science and Engineering	122	10-fold cross validation	16.00%	84.0000%
Humanities	81	10-fold cross validation	17.16%	82.8400%
1st&2nd year undergraduate	89	10-fold cross validation	14.47%	85.5300%
3rd year undergraduate or higher	113	10-fold cross validation	18.42%	81.5800%
average MAPE			0.162542857	83.7457%

Table of experiments using Linear regression

As the table showed, the model was built for each group of students, and the MAPE and accuracy is calculated out. The highest accuracy of group of students are highlighted.

SMO Regression

```

weights (not support vectors):
-    0.0156 * (normalized) health
+    0.1998 * (normalized) relation
+    0.2256 * (normalized) friend
-    0.0088 * (normalized) finance
-    0.0059 * (normalized) activity
+    0.1086 * (normalized) grade
-    0.0782 * (normalized) resource
+    0.0547 * (normalized) interest
-    0.104 * (normalized) job
+    0.0229 * (normalized) time
+    0.1103 * (normalized) life
+    0.2976 * (normalized) esteem
+    0.1169

Number of kernel evaluations: 20503 (94.974% cached)

Time taken to build model: 0.03 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient          0.7531
Mean absolute error             0.8147
Mean absolute percentage error  0.1558
Root mean squared error        1.1103
Relative absolute error         60.7854 %
Root relative squared error     65.6973 %
Total Number of Instances      202

```

Result of SMO Regression

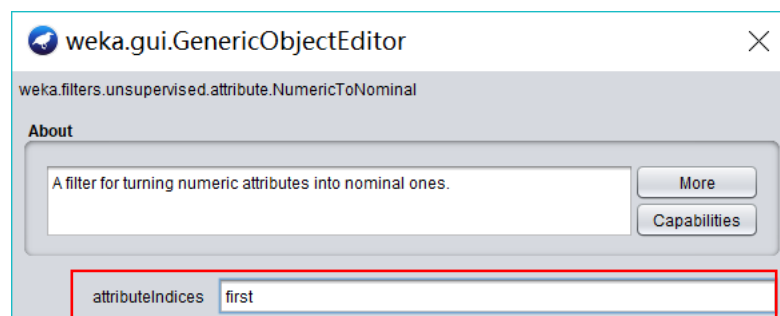
The approach of SMO Regression is similar to linear regression because they are all regression. And the Mean absolute percentage error also calculated out.

datasets	number of students	testing condition	MAPE	Accuracy
total	202	10-fold cross validation	0.1558	84.4200%
female	81	10-fold cross validation	0.1457	85.4300%
male	121	10-fold cross validation	0.1681	83.1900%
Science and Engineering	122	10-fold cross validation	0.1737	82.6300%
Humanities	81	10-fold cross validation	0.1776	82.2400%
1st & 2nd year undergraduate	89	10-fold cross validation	0.1528	84.7200%
3rd year undergraduate or higher	113	10-fold cross validation	0.1799	82.0100%
average MAPE			0.1648	83.5200%

Table of experiments using SMO regression

REP tree

The approach of decision tree is a classification. Before building the model, the cluster of overall happiness needs to be defined. In this project, the happiness index is in range 1-10. 1-4 are in unhappy cluster, 5-7 are in neutral cluster, 8-10 are in happy cluster. To do this, the numeric data of happiness index should be converted to nominal form first.



Select the attribute (happiness index) needs to convert

Selected attribute	
Name: Q5	Type: Numeric
Missing: 0 (0%)	Unique: 1 (0%)
Distinct: 10	
Statistic	Value
Minimum	1
Maximum	10
Mean	6.545
StdDev	1.687



Selected attribute

Name: Q5
Missing: 0 (0%)

Distinct: 10

Type: Nominal
Unique: 1 (0%)

No.	Label	Count	Weight
1	1	1	1.0
2	2	2	2.0
3	3	8	8.0
4	4	11	11.0
5	5	27	27.0
6	6	41	41.0
7	7	55	55.0
8	8	37	37.0
9	9	13	13.0
10	10	7	7.0

The next step is to merge the values of happiness into three clusters.

weka.filters.unsupervised.attribute.MergeManyValues

About

Merges many values of a nominal attribute into one value. More Capabilities

attributeIndex: first

debug: False

doNotCheckCapabilities: False

ignoreClass: False

label: unhappy

mergeValueRange: 1,2,3

Open... Save... OK Cancel

Selected attribute

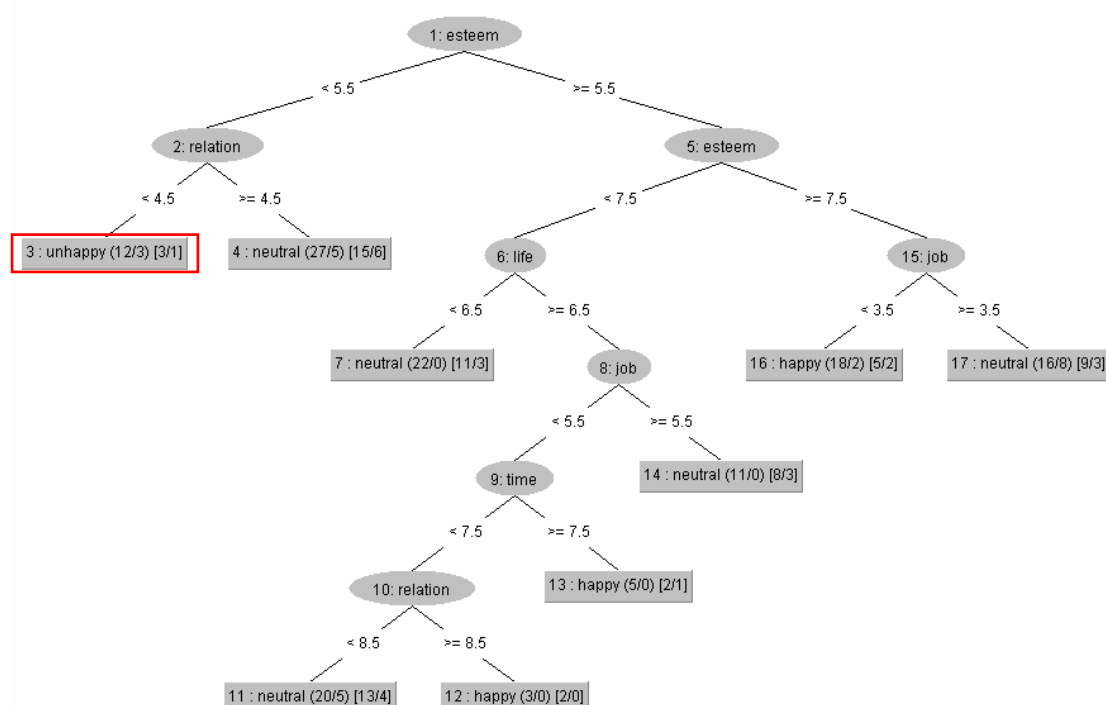
Name: Q5
Missing: 0 (0%)

Distinct: 3

Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	unhappy	22	22.0
2	neutral	123	123.0
3	happy	57	57.0

Happiness attribute after clustering



REP Tree for all students

Each node of the REPTree is a factor of happiness, students were split into two subtrees based on the rule of the node. For example, for the leave 3 which is highlighted, the total students were split

into 2 group after the root node: students with self-esteem over 5.5 and students with self-esteem under 5.5. For students has esteem under 5.5, the students were split into 2 group based on relationship satisfaction: students with relationship satisfaction under 4.5 and students with relationship satisfaction over 4.5. For group of students with self-esteem under 5.5 and relationship satisfaction under 4.5, the amount of this group of students is 3. The tree graph shows the split of student's happiness are mostly based on their self-esteem and relationship satisfaction of them.

datasets	number of students	testing condition	MAPE	Accuracy
total	202	10-fold cross validation	0.1746	82.5400%
female	81	10-fold cross validation	0.1303	86.9700%
male	121	10-fold cross validation	0.174	82.6000%
Science and Engineering	122	10-fold cross validation	0.1633	83.6700%
Humanities	81	10-fold cross validation	0.2014	79.8600%
1st & 2nd year undergraduate	89	10-fold cross validation	0.1486	85.1400%
3rd year undergraduate or higher	113	10-fold cross validation	0.1946	80.5400%
average MAPE			0.169542857	83.0457%

Table of experiments using REPTree

4 Analysis and evaluation

This section of the report is the analysis of the model. The objectives of the section are:

- Analysis of the importance of factors of student's happiness.
- Comparison of accuracy of each model for different groups of people.
- Find out the similarities and differences of importance of factor between different groups of students.
- Find out the similarities and differences of importance of factor from student's approximation and result of machine learning.

4.1 Importance of factors

datasets	number of students	health	relation	friend	finance	activity	grade	resource	interest	job	time	life	esteem
total	202	0.046	0.1822	0.1557	0.0083	0.0608	0.0679	-0.0938	0.0693	0.1454	-0.0388	0.1436	0.3685
female	81	0.0315	0.4575	-0.1072	0.0306	-0.0846	0.0801	-0.0341	0.1094	0.0681	-0.0347	0.0683	0.2272
male	121	0.0706	0.0133	0.2865	0.0041	0.1895	0.0923	-0.0545	0.0044	0.1503	-0.0068	0.1273	0.2594
Science and Engineering	122	0.0882	0.2206	0.243	0.0131	0.061	0.0375	-0.1142	0.0212	0.1926	0.0614	0.1666	0.265
Humanities	81	0.0468	0.1162	0.2253	0.0572	-0.0737	0.1823	-0.1035	0.0875	0.0785	-0.2381	0.2255	0.4417
1st & 2nd year undergraduate	89	-0.01	0.163	0.1741	0.0152	0.003	0.0005	-0.1117	0.2242	0.0343	-0.0119	0.1188	0.4273
3rd year undergraduate or higher	113	0.0546	0.2007	0.0513	0.0224	0.1182	0.1312	-0.0883	-0.0134	0.2328	-0.051	0.1839	0.2637

Table of the importance of factors

The table shown above is the result of linear regression model, which directly reflects the weights of each factor. It presents the importance of factors among different groups of students. In this table, there are 7 groups of people shows in rows. There are 12 factors show in columns: health-health, relation-relationship satisfaction, friend-friendship satisfaction, finance-finance condition, activity-activity satisfaction, resource-learning resource availability, interest-interest on courses, job-future job availability, time-time management satisfaction, life-living condition satisfaction, esteem-self-esteem.

Green blocks represent the most positive factor in the group of people. Most positive factor is the

factor which has maximum weight of the model. As the table shows, the most important positive factor of student's happiness is obviously self-esteem. For female students, the most important positive factor is relationship satisfaction, for male students, the most important positive factor is friendship satisfaction. For other cases of group of students, the most important positive factor is self-esteem. As a result, the students who has higher degree of self-esteem are happier. The question about self-esteem in questionnaire is: "what degree of self-esteem do you think you have?", the answer to this question is non-bias because the people who chose 10 are actually relatively confident, people know their self-esteem very well.

Yellow blocks represent the most neutral factor in the group of people, the result shows the most neutral factor for total students is finance. The most neutral factor is the factor which has the least influence on student's overall happiness. In other words, the value which is nearest to '0'.

Red blocks represent the most negative factor in the group of people. Most negative factor is the factor which has least value coefficient and negative influence on student's happiness. The result shows the most negative factor is learning resource availability. However, the extent of the negative factor is not large compares to weights of other factors. So, there is two reason of that, the students might not care much about the availability of learning resources, or the students who are satisfied with their availability of learning resources are more likely to be unhappy.

4.2 Comparison of models

datasets	Linear Regression	SMOReg	REPTree
total	0.1565	0.1558	0.1746
female	0.1451	0.1457	0.1303
male	0.1757	0.1681	0.174
Science and Engineering	0.16	0.1737	0.1633
Humanities	0.1716	0.1776	0.2014
1st & 2nd year undergraduate	0.1447	0.1528	0.1486
3rd year undergraduate or higher	0.1842	0.1799	0.1946
average MAPE	0.162542857	0.1648	0.169542857

Accuracy of each model among different groups of students

This table presents the accuracy of different model used on different groups of students. Every time the model is built, the MAPE is recorded on the table for 7 groups of students. The green blocks highlight the most accuracy model with smallest mean absolute percentage absolute error for the current group of people. The red blocks highlight the least accuracy model with highest mean absolute percentage absolute error for the current group of people.

In average, three of the algorithms has a similar performance on the dataset. Linear Regression is the most accuracy model with an average accuracy of **83.8%** with least mean absolute percentage error. SMO regression is on the second place with an average accuracy of **83.5%**, and REPTree algorithm has least accuracy of **83.0%**.

In total, the best performance appears when REPTree algorithm is used to train the model of female student's happiness, the accuracy reaches **87%** which is the highest point throughout the whole project.

4.3 Comparison of factors among different groups

datasets	number of students	health	relation	friend	finance	activity	grade	resource	interest	job	time	life	esteem	constant
total	202	0.046	0.1822	0.1557	0.0083	0.0608	0.0679	-0.0938	0.0693	0.1454	-0.0388	0.1436	0.3685	0.8662
female	81	0.0315	0.4575	-0.1072	0.0306	-0.0846	0.0801	-0.0341	0.1094	0.0681	-0.0347	0.0683	0.2272	2.5459
male	121	0.0706	0.0133	0.2865	0.0041	0.1895	0.0923	-0.0545	0.0044	0.1503	-0.0068	0.1273	0.2594	0.6693
Science and Engineering	122	0.0882	0.2206	0.243	0.0131	0.061	0.0375	-0.1142	0.0212	0.1926	0.0614	0.1666	0.265	0.2995
Humanities	81	0.0468	0.1162	0.2253	0.0572	-0.0737	0.1823	-0.1035	0.0875	0.0785	-0.2381	0.2255	0.4417	0.7397
1st & 2nd year undergraduate	89	-0.01	0.163	0.1741	0.0152	0.003	0.0005	-0.1117	0.2242	0.0343	-0.0119	0.1188	0.4273	0.3561
3rd year undergraduate or higher	113	0.0546	0.2007	0.0513	0.0224	0.1182	0.1312	-0.0883	-0.0134	0.2328	-0.051	0.1839	0.2637	1.9668

Table of comparison of factors among different groups

The table shown above is the result from linear regression model, which directly reflects the weights of each factor. It shows the total students were split into 6 groups, which is 3 pair of comparison: male-female, students in school of science and engineering – students in school of humanities, first and second year undergraduate students – third and fourth year undergraduate or higher students. Each pair of comparison, the project reflects some interesting results.

datasets	number of students	health	relation	friend	finance	activity	grade	resource	interest	job	time	life	esteem
female	81	0.0315	0.4575	-0.1072	0.0306	-0.0846	0.0801	-0.0341	0.1094	0.0681	-0.0347	0.0683	0.2272
male	121	0.0706	0.0133	0.2865	0.0041	0.1895	0.0923	-0.0545	0.0044	0.1503	-0.0068	0.1273	0.2594

Comparison of factors between female and male students

Firstly, for different genders. The female students have much higher weight of their relationship factor. That reflects the importance of relationship satisfaction for female students is higher than male students. The satisfaction on relationship has high influence on happiness of female students. But for male students, their friendship satisfaction plays an important role on their overall happiness. The result shows, female student's happiness get affected more by their boy/girlfriend, however, male students care more about their friends. In 'Intimate Relationship (Rowland S. Miller)', the researcher investigates the difference in sexual selection of male and female, based on the evolutionary psychology. The parental investment for male and female has large difference, it results in the difference of attitude of relationship.

datasets	number of students	health	relation	friend	finance	activity	grade	resource	interest	job	time	life	esteem
Science and Engineering	122	0.0882	0.2206	0.243	0.0131	0.061	0.0375	-0.1142	0.0212	0.1926	0.0614	0.1666	0.265
Humanities	81	0.0468	0.1162	0.2253	0.0572	-0.0737	0.1823	-0.1035	0.0875	0.0785	-0.2381	0.2255	0.4417

Comparison of factors between science and engineering and humanities students

The significant difference between these two group of students is the finance condition factor. The humanities students have much higher weight of their finance condition factor. That reflects the importance of finance condition for humanities students is higher than science students. The reason for this result is predictable. Because most of students in school of Humanities is in Business school or learning some courses about business and economies. These students might be more interested in business and things to do with money. As a result, the happiness of students in school of humanities students might be affected by their finance condition more than students in school of science and engineering.

datasets	number of students	health	relation	friend	finance	activity	grade	resource	interest	job	time	life	esteem
1st & 2nd year undergraduate	89	-0.01	0.163	0.1741	0.0152	0.003	0.0005	-0.1117	0.2242	0.0343	-0.0119	0.1188	0.4273
3rd year undergraduate or higher	113	0.0546	0.2007	0.0513	0.0224	0.1182	0.1312	-0.0883	-0.0134	0.2328	-0.051	0.1839	0.2637

Comparison of factors between 1-2year students and >3rd year students

The significant difference between these two group of students is the health factor. The students over 3rd year undergraduate have much higher weight of their health factor. That reflects the importance of health for students over 3rd year undergraduate is higher than first- or second-year undergraduate students. The reason of this result is interesting to explore, new students are more likely not to care their health, some of them attend parties and hangover every night. As they get older, their body might no longer support them to do these. Moreover, the academic or job pressure will get higher as they get older.

The result also shows the younger students cares more about the interest of courses, older students cares more about the job availability of their course. The younger students are more likely to care about the interest of the course because they just start to learn the courses and they must spend several years on that, the interest on the course is important. For older students, they are nearly graduated and nearly get to work, so the job availability is important subject for them to think about. If a student studies a course which is hard to find a job, he might feel the pressure everyday until he successfully finds the job.

4.4 Comparison of results with student's approximation

In the questionnaire, the project set one question after each question about satisfaction of factor, to record the student's approximation of each factor. The part of these questions is not mandatory, but most students answers.

datasets	health	relation	friend	finance	activity	grade	resource	interest	job	time	life	esteem
machine learning approach	4.13%	16.34%	13.96%	0.74%	5.45%	6.09%	-8.41%	6.21%	13.04%	-3.48%	12.88%	33.05%
student's approximation	9.37%	8.53%	8.94%	8.11%	7.62%	8.07%	7.45%	7.79%	8.46%	7.93%	8.81%	8.92%

Comparison of machine learning approach and student's approach

The result shows for both machine learning approach and student's approximation, the learning resource availability is the least important factor of student's happiness. But the student's approximation presents the most important factor is health, while the machine learning approach presents the most factor is self-esteem.

In the student's approximation, the importance they think for relationship satisfaction and friendship satisfaction is also relatively high. Based on research about self-esteem, self-esteem strongly related to the satisfaction in relationships and friendships. A person with high self-esteem will get happier in relationships and a person with low self-esteem is easily to be depressed in a relationship. So, the power of self-esteem in student's happiness is significant.

As the table shows, the more important factor on happiness highlights greener, less important factors highlights more red. The colour of the table reflects the trend of student's approximation is similar to machine learning approach even the value is very different. The standard deviation of student's approximation is much smaller but the rank of the importance is similar.

machine learning approach	esteem	relation	friend	job	life	interest	grade	activity	health	finance	time	resource
student's approximation	health	friend	esteem	life	relation	job	finance	grade	time	interest	activity	resource

Rank of importance of factors for both approach

5 Reflection and conclusion

This section is mainly about reflecting on the planning and management of this project.

5.1 Timing and Milestones

stages	begin date	end date
planning	26/9/2018	28/9/2018
background reading	28/9/2018	10/10/2018
questionnaire design	12/10/2018	26/10/2018
data collection	26/10/2018	13/1/2019
background reading	11/1/2019	20/1/2019
modelling	25/1/2019	22/2/2019
analysis	23/2/2019	5/3/2019

Timing and milestones of the project

The planning was recorded in a Microsoft Excel file, most of the stages was completed on time except the stage of data collection. As the introduction above, the questionnaire design makes the data collection process slower, but the quality of data will be higher. During the time of data collection, as the stage does not require much time if the students fill in the questionnaire online, the background reading about machine learning algorithm of modelling stage can concurrently proceed. In December of 2018, over 100 of data was collected, it did not achieve the objective of 200 responses in December, but the amount of data is enough to be the dataset of modelling. So, the implementation of machine learning algorithm was concurrently proceed using about 100 responses first to learn the skill of building models and analysis. As long as the amount of responses reaches 200, the process of machine learning stage is similar. That is the strategy to save time of the project.

5.2 Reflection on the project

Throughout progress of the project, the student learns the skill of planning and implement a complex project. The project contains both computer science techniques and social science techniques. In the early part of the project, the challenge is to design a questionnaire with selected factors of happiness and collect over 200 responses. The student learned how to investigate the required research to solve problems, not only researches about computer science techniques, but also in social science, philosophy, psychology fields. The student learned the machine learning

algorithms and their principles, and the knowledge of some machine learning tools such as WEKA.

5.3 Conclusion

Most objectives of the project have been achieved before May 2019. Over 200 high-quality responses were successfully collected from the university students by online questionnaire platform. Three machine learning algorithms (Linear Regression, SMO Regression, REP tree) were successfully used to build models of student's happiness and reflect the relationship between factors of student's happiness and their overall happiness. The index of accuracy, mean absolute percentage error, for the three machine learning algorithms was calculated to make a comparison between the models. A reasonable result of comparison of importance of factors among different group of people was made.

As a result, 'self-esteem', 'relationship satisfaction', 'friendship satisfaction', 'living condition satisfaction' were shown to be the most important factor on student's happiness. 'Linear Regression' had the best performance of modelling among the three algorithms. The amount of responses might not be enough to prove the correctness of the results, but it is still a successful year project and a very valuable experience of the student.

Reference

- [1] Maslow, A.H. (1943). "A theory of human motivation". *Psychological Review*. 50 (4): 370–96. CiteSeerX 10.1.1.334.7586. doi:10.1037/h0054346— via psychclassics.yorku.ca.
- [2] Hirvonen, T., & Mangeloja, E. (2005). What Makes University Students Happy? *International Review of Economics Education*, Volume 6, Issue 2, 2007, Pages 27-41
- [3] Cohen, J., Cohen P., West, S.G., & Aiken, L.S. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates. 2003.
- [4] Platt, John (1998), *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines* (PDF), CiteSeerX 10.1.1.43.4376
- [5] Mathworks.com. (2019). Multiple linear regression - MATLAB regress. [online] Available at: <https://www.mathworks.com/help/stats/regress.html> [Accessed 27 Apr. 2019]
- [6] Alisneaky, svg version by User:Zirguezi - CC BY-SA 4.0 , <https://commons.wikimedia.org/w/index.php?curid=47868867>
- [7] Ho, Tin Kam (1995). *Random Decision Forests*(PDF). *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.
- [8] Weka.sourceforge.net. (2019). REPTree. [online] Available at: <http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/REPTree.html?is-external=true> [Accessed 27 Apr. 2019].
- [9] Ian H. Witten; Eibe Frank; Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco. Retrieved 2011-01-19.
- [10] G. Holmes; A. Donkin; I.H. Witten (1994). "Weka: A machine learning workbench" (PDF). *Proc Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia. Retrieved 2007-06-25.
- [11] "1.2 Design Goals of the Java™ Programming Language". Oracle. January 1, 1999. Archived from the original on January 23, 2013. Retrieved January 14, 2013.
- [12] Europeansocialsurvey.org. (2019). [online] Available at: http://www.europeansocialsurvey.org/docs/round6/questionnaire/ESS6_final_personal_and_social_well_being_module_template.pdf [Accessed 28 Apr. 2019].
- [13] En.wikipedia.org. (2019). *Mean absolute percentage error*. [online] Available at: https://en.wikipedia.org/wiki/Mean_absolute_percentage_error [Accessed 30 Apr. 2019].