GROUP ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT119-3-2-DMPM

DATA MINING AND PREDICTIVE MODELLING

Hand Out Date: 30 Nov 2022

Submission Date: 17 Feb 2023

Intake Code: APD2F2206CS(DA)

Lecturer: DR. PREETHI SUBRAMANIAN

Title: To identify the characteristics of people with high work-life balance score

Group Members:

1. Chee Wei Keng TP063861
2. Tan Yong Hong TP063960
3. Wong Kit Yee TP061080

Problem Statement **4**

Objective and Scope **4**

Methodology Selection **5**

Enterprise Miner - Process Flow Diagram **5**

Exploratory Data Analysis (EDA) **6**

Data Pre-processing **13**
a) Data Selection 13
b) Removing missing value 14
c) Dropping low variable worth column 15
d) Dropping low variance column 15
e) Data Partition 16

Model Construction & Optimization **17**
Wong Kit Yee TP061080 17
Model 1 - Regression 17
Model 2 - Neural Network 20
Chee Wei Keng TP063861 23
Model 1 - Clustering 23
Model 2 - Neural Network (5L 5N HPNN) 24
Tan Yong Hong TP063960 28
Model 1 & 2- Regressions & HP Regressions 28

Model Validation **33**
Wong Kit Yee TP061080 33
Model 1 - Regression 33
Model 2 - Neural Network 34
Chee Wei Keng TP063861 37
Model 1 – Clustering 37
Model 2 – Neural Network 38
Tan Yong Hong TP063960 41
Model 1 - 2nd Degree Polynomial Stepwise Regression 42
Model 2 - Stepwise Regression 43

Critical Interpretation of Outcomes **44**
Wong Kit Yee TP061080 44
Model 1 - Regression 44
Model 2 - Neural Network 48
Chee Wei Keng TP063861 50

Model 1 - Clustering 50
Model 2 - Neural Network (5L 5N HPNN) 54
Tan Yong Hong TP063960 56
Model 1 - 2nd Degree Polynomial Stepwise Regression 56
Model 2 - Stepwise Regression 60

Discussion & Conclusion **64**
Model Comparison 64
Ensemble Model 64

References **66**

1.0 Problem Statement

The phrase "work-life balance" refers to the method in which people allocate their time between the workplace and other facets of their daily life, such as relaxation, self-care, and interpersonal relationships (Moulder, 2022). According to the study, 72% of job seekers identified work-life balance as a key factor to consider (Statista, 2022). In 2021, 89% of Americans reported experiencing burnout due to their jobs. Numerous signs and symptoms of this burnout include: -- 44% physical exhaustion, 36% mental fatigue, 32% emotional drained, 26% lack of enthusiasm as well as 19% working without effort (*20+ Alarming Burnout Statistics [2022]: Stress and Lack of Motivation in the Workplace – Zippia*, 2022). There are certain things that individuals must do in order to lead a healthy life. By analysing the selected dataset, the group intends to figure out what are the key characteristics that score high in terms of work-life balance.

2.0 Objective and Scope

Objective:

- To explore the dataset and provide potentially useful insights to help people with maintaining work-life balance
- To create a predictive model that figures out the key characteristics to score high in terms of work-life balance
- To create a descriptive model that identifies the key metrics that can improve well-being

Scope:

- The dataset is obtained from Kaggle
    - https://www.kaggle.com/datasets/ydalat/lifestyle-and-wellbeing-data
- The dataset time period is between the years 2015 - 2021
- The dataset contains 24 columns and 15973 rows

3.0 Methodology Selection

Amongst the methodologies available, CRoss Industry Standard Process for Data Mining (CRISP-DM) is the method chosen to be applied to tackle the problem to be analyzed. According to Hotz (2018), CRISP-DM consists of 6 steps, which are business understanding, data understanding, data preparation, modelling, evaluation and deployment.

CRISP-DM is considered due to the benefits provided. It's generalize-able as it can provide clear guidance for data science activities, feasible as the steps in CRISP-DM are essentially the common sense required before diving into the data prepared, adaptable as this method doesn't require training or controversy to be implemented, beneficial when it comes to providing the proper initiation to tackle a problem as starting with business understanding helps to understand the context of the industry involved, flexible as users can cycle through the steps in CRISP-DM to gain more understanding as each cycle goes by which are crucial to tackling the problem and enables a strong finish with closing the usage of the methodology with the deployment phase. (Hotz, 2018)

4.0 Enterprise Miner - Process Flow Diagram

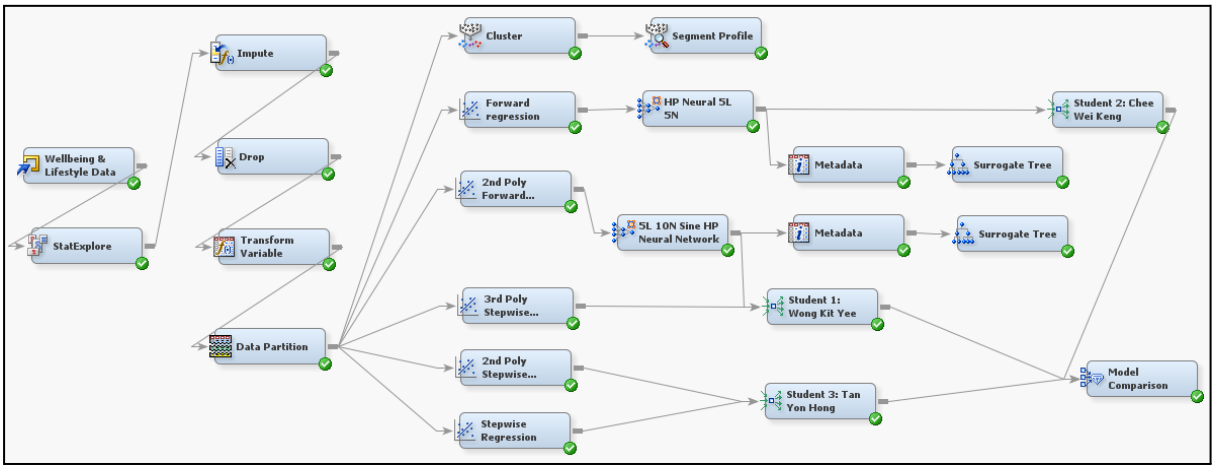


Figure 1: Process Flow Diagram in SAS Enterprise Miner

5.0 Exploratory Data Analysis (EDA)

The Work-Life Balance survey dataset holds information on 15,977 individuals, using 24 columns of data to evaluate their habits and lifestyles to maximize life satisfaction. The five dimensions evaluated include Healthy Body, Healthy Mind, Expertise, Connection, and Meaning. The data covers aspects like fitness, emotions, social network, and compassion. The goal is to analyze the data to optimize individual well-being and support the UN Sustainable Development Goals.

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	AGE	INPUT	4	0	21 to 35	38.24	36 to 50	29.14
TRAIN	GENDER	INPUT	2	0	Female	61.72	Male	38.28

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
ACHIEVEMENT	INPUT	4.000751	2.755837	15972	0	0	3	10	0.628884	-0.35502
BMI_RANGE	INPUT	1.410656	0.491968	15972	0	1	1	2	0.363255	-1.86828
CORE_CIRCLE	INPUT	5.508077	2.840334	15972	0	0	5	10	0.202431	-0.9668
DAILY_SHOUTING	INPUT	2.930879	2.676301	15972	0	0	2	10	1.118024	0.535427
DAILY_STEPS	INPUT	5.703606	2.891013	15972	0	1	5	10	0.0393	-1.16792
DAILY_STRESS	INPUT	2.791685	1.367801	15971	1	0	3	5	-0.08053	-0.79838
DONATION	INPUT	2.715314	1.851586	15972	0	0	3	5	-0.04791	-1.43948
FLOW	INPUT	3.194778	2.357518	15972	0	0	3	10	0.870127	0.266194
FRUITS_VEGGIES	INPUT	2.922677	1.442694	15972	0	0	3	5	-0.05098	-1.00704
LIVE_VISION	INPUT	3.752129	3.230987	15972	0	0	3	10	0.740936	-0.56921
LOST_VACATION	INPUT	2.898886	3.69218	15972	0	0	0	10	0.920513	-0.66794
PERSONAL_AWARDS	INPUT	5.711558	3.08963	15972	0	0	5	10	0.061245	-1.17737
PLACES_VISITED	INPUT	5.23297	3.311912	15972	0	0	5	10	0.170494	-1.25206
SLEEP_HOURS	INPUT	7.042888	1.199044	15972	0	1	7	10	-0.3558	1.172863
SOCIAL_NETWORK	INPUT	6.474267	3.086672	15972	0	0	6	10	-0.17518	-1.3704
SUFFICIENT_INCOME	INPUT	1.728963	0.444509	15972	0	1	2	2	-1.03031	-0.93857
SUPPORTING_OTHERS	INPUT	5.616454	3.242021	15972	0	0	5	10	0.055665	-1.27058
TIME_FOR_PASSION	INPUT	3.326572	2.729293	15972	0	0	3	10	0.841172	-0.19397
TODO_COMPLETED	INPUT	5.745993	2.624097	15972	0	0	6	10	-0.35608	-0.70454
WEEKLY_MEDITATION	INPUT	6.233346	3.016571	15972	0	0	7	10	-0.19252	-1.15012
WORK_LIFE_BALANCE_SCORE	TARGET	666.7515	45.01987	15972	0	480	667.7	820.2	-0.10467	-0.10842

Figure 2: Output from StatExplore: Summary Statistics

Based on the available data, the mean and standard deviation of the majority of the variables indicate that most of the data points are located close to the mean. However, sleep hours and work-life balance scores have a wider contrast between the mean and standard deviation, indicating that these variables have more widely spread data points. The only variable with a mean less than its standard deviation is lost vacation, which indicates data points are widely spread and the variability is high.

The data shows a mix of positive and negative skewness values, which indicate the asymmetry of the distribution. Positive skewness values suggest a longer tail to the right, and

some with negative skewness values suggest a longer tail to the left. The position of the tail in the graphs of these variables indicates the position of the outliers.

In terms of kurtosis, most of the variables have a negative value under 2, indicating that they are platykurtic and have a flatter distribution than a normal distribution, indicating that they are more evenly spread than a variable with normal distribution. However, sleep hours, work-life balance score, weekly meditation, todo completed, social network, fruit veggies, donation, and daily stress have a positive kurtosis value under 2, indicating that they are leptokurtic and have a more peaked distribution than a normal distribution.

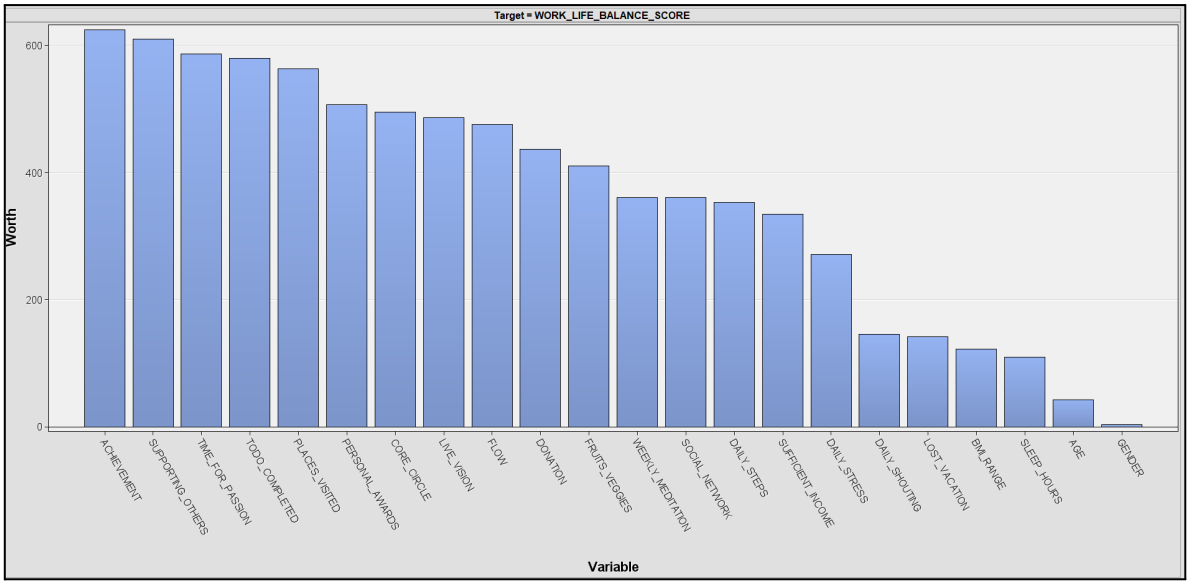


Figure 3: Variable Worth Graph

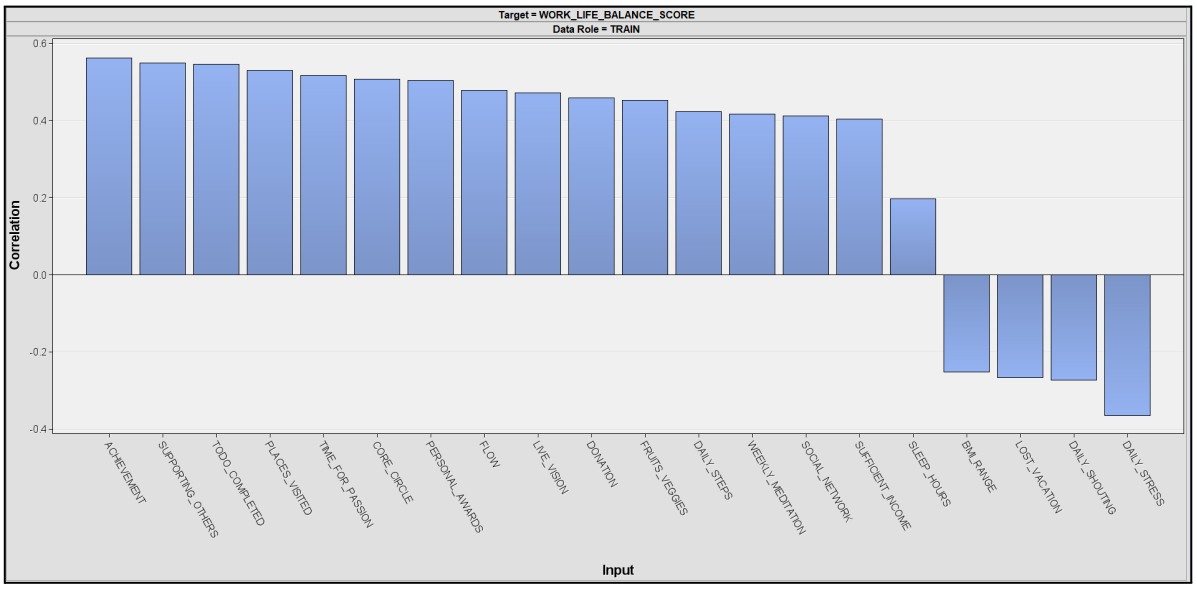


Figure 4:

The variable "Achievement" has the highest worth of nearly 600 and is estimated to have a correlation with the target variable of approximately 0.58. Meanwhile, the variable "Gender" is estimated to have a near-zero worth, meaning it is not significantly correlated with the target variable.

The correlations of the variables with the target variable are estimated to be in the range of -0.4 to 0.6, with all of them having an absolute value smaller than 0.6. This suggests that while some variables may play a significant role in predicting the target variable (such as "Achievement", "Supporting_Others", "Time_For_Passion", "ToDo_Completed", "Placed_Visited"), others may not have much impact on the prediction.

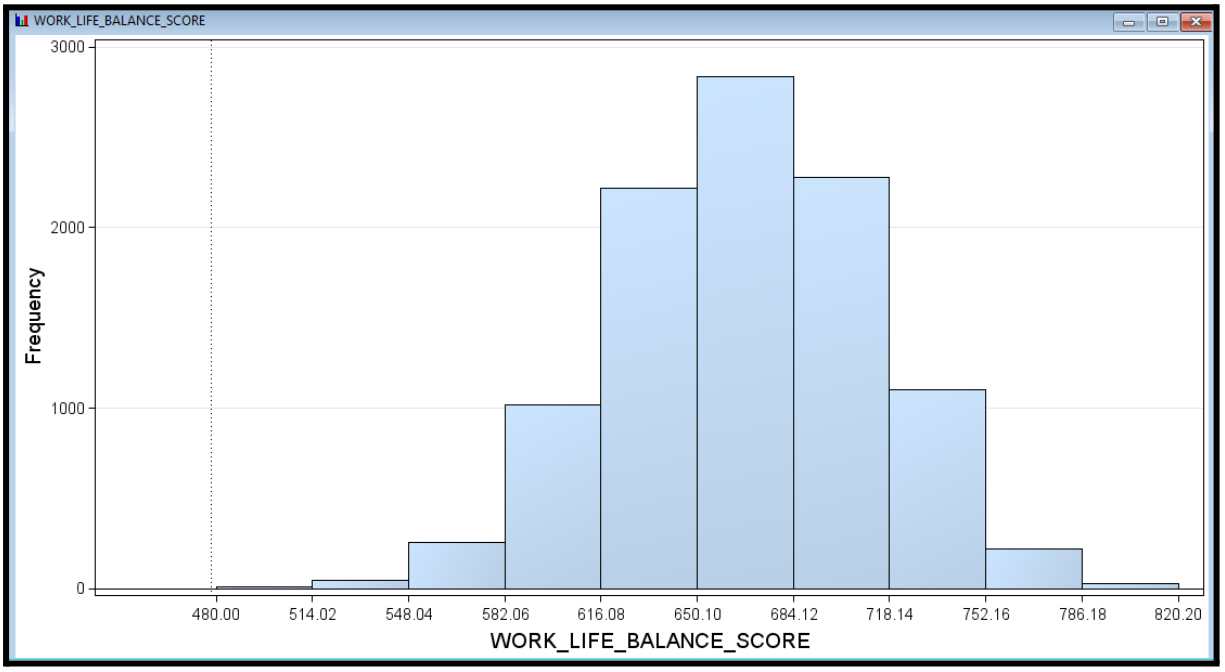


Figure 5: Histogram for target variable: Work-Life Balance Score

According to the target variable chart above, Work-Life Balance Score between 650 and 680 is the most attainable for an individual. Scores that are less than 548 or greater than 786 are outliers. However, these outliers are acceptable because each person's Work-Life Balance Score is determined by his or her personal lifestyle and habits. Therefore, these outliers will remain.

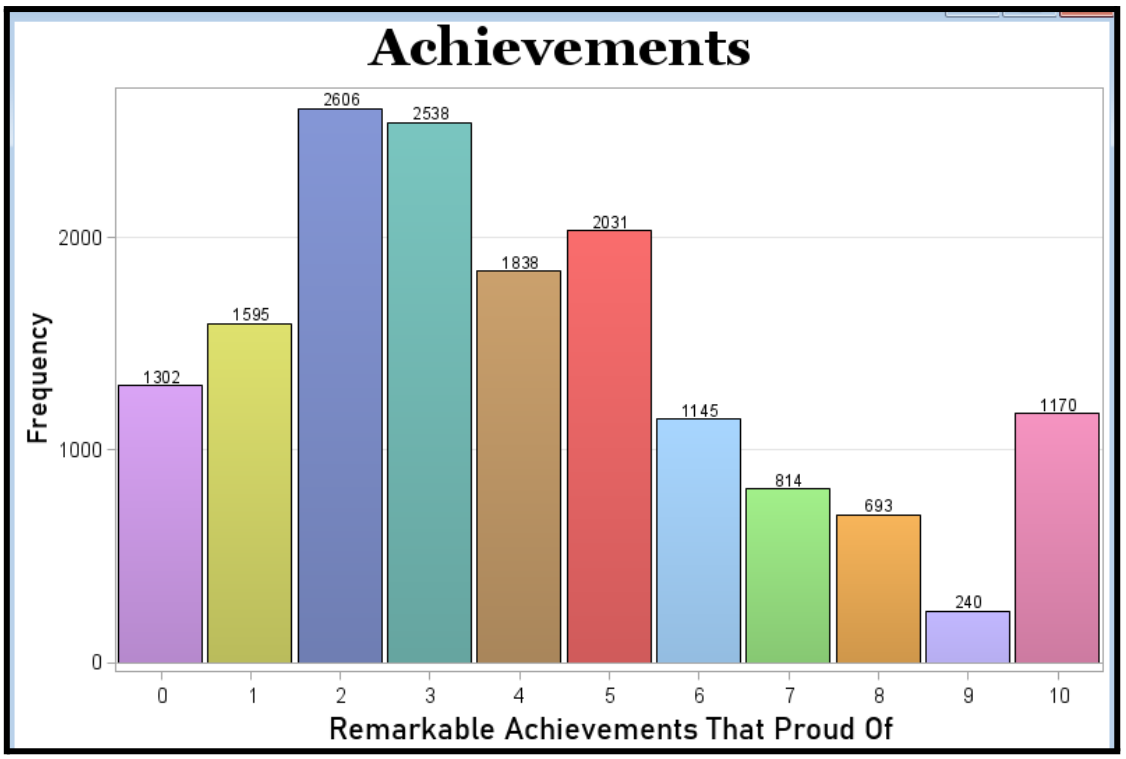


Figure 6: Bar Chart for Achievement

The majority of individuals are proud of two to three achievements in their lifetimes. There were 2606 and 2538 respective individuals in total. Moreover, we can observe a gradual drop in the number of individuals who feel proud of more than five achievements in their lifetimes.

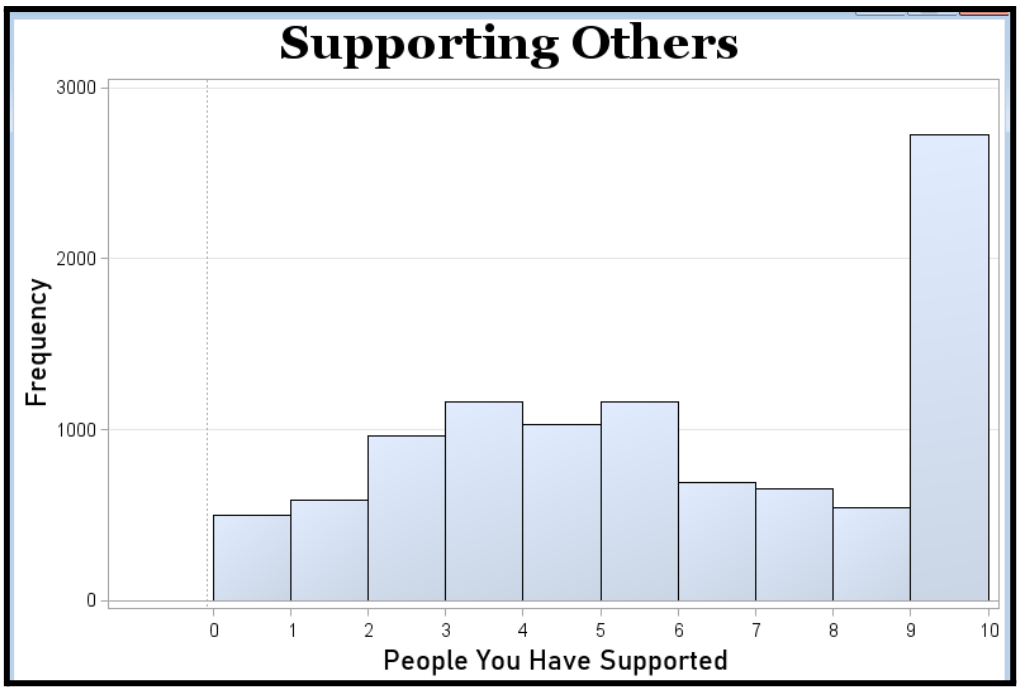


Figure 7: Histogram for Supporting Others

There are no missing values in the dataset. The survey question is defined as "How Many People Do You Help Achieve a Better Life?'. As shown in figure 7, people nowadays like supporting and helping others, however, there will still be a few people who claim they have not helped anyone.

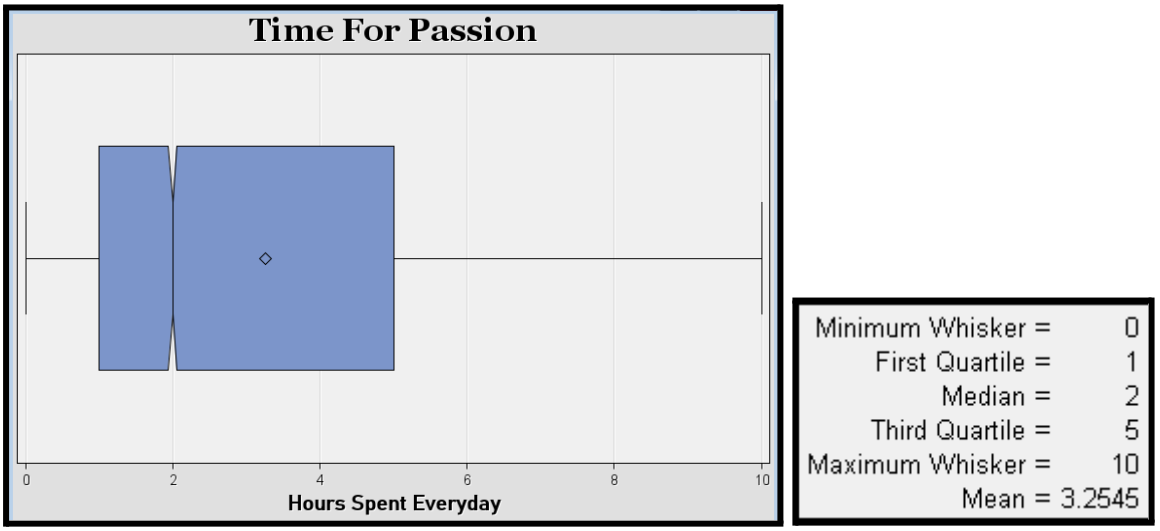


Figure 8: Boxplot for Time For Passion

The survey question is defined as "How Many Hours Do You Spend Everyday Doing What You Are Passionate About?" It is shown that there are no outliers in the dataset. They spend an average of three hours every day doing what they are passionate about.

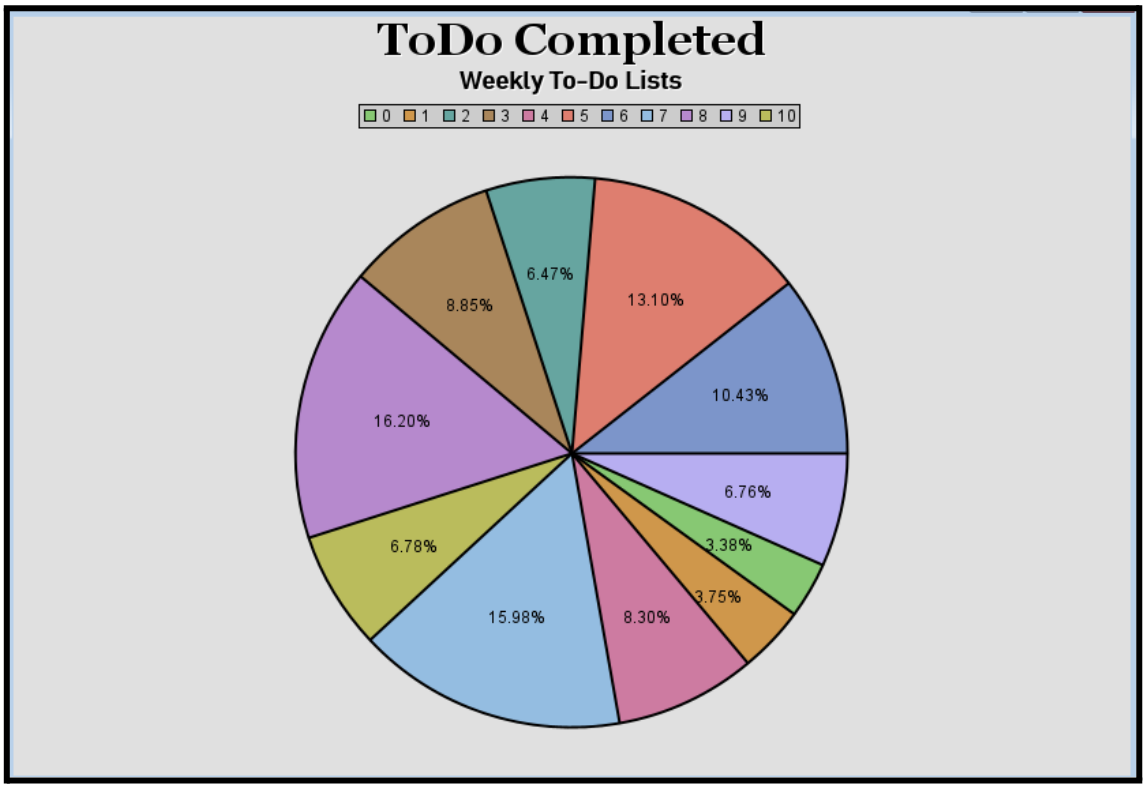


Figure 9: Pie Chart for ToDo Complete

The survey question is defined as "How Well Do You Complete Your Weekly To-Do Lists?" To-Do List is a list of tasks that needs to be completed by the individual. Figure 9 indicates that around 16% of people claim to have 80% confidence in their ability to accomplish their weekly To-Do list.

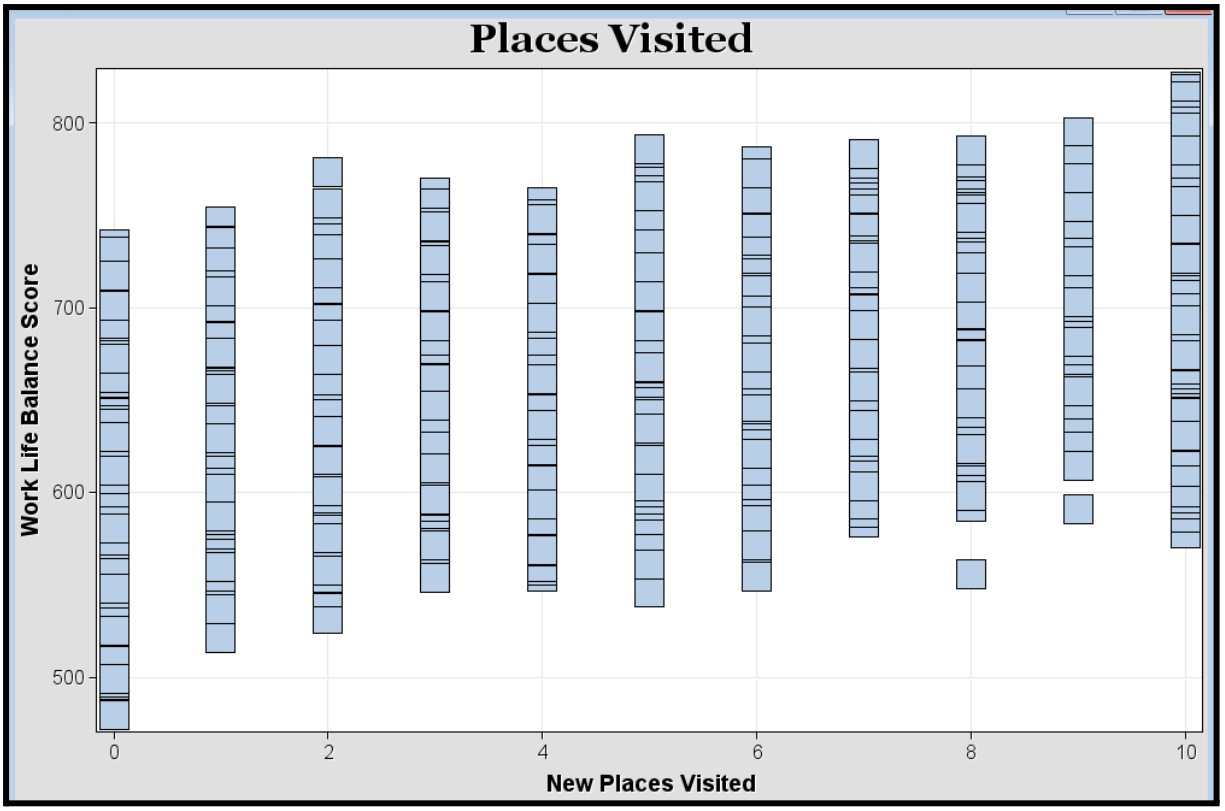


Figure 10: Scatter Plot for Places Visited

The survey question is defined as "How Many New Places Do You Visit?" It is shown that individuals with a higher Work-Life Balance Score are more likely to visit new places in their lifetimes. Individuals who have not visited any new places only existed among those with work-life balance scores less than or equal to 500.

6.0 Data Pre-processing

Data pre-processing involves the cleaning, transforming, and preparing of raw data to make it suitable for analysis and modelling. This step is crucial for guaranteeing that the data is in a format that is fit for the intended analysis and modelling methods. The WORK_LIFE_BALANCE_SCORE variable is set to the target variable in the dataset. The "DAILY_STRESS" column had a missing value when discovering the dataset. Not only did the “BMI_RANGE” and “SUFFICIENT_INCOME” variables have low standard deviation values, but also the “GENDER” variable has a low value of 3.78 on the variable worth graph. Also, the data has been divided into two parts with 80% of it being designated for training and the other 20% for validation.

The respective detailed steps for each process involved in this dataset are as follows: --

a) Data Selection

Name	Role	Level
ACHIEVEMENT	Input	Interval
AGE	Input	Nominal
BMI_RANGE	Input	Interval
CORE_CIRCLE	Input	Interval
DAILY_SHOUTING	Input	Interval
DAILY_STEPS	Input	Interval
DAILY_STRESS	Input	Interval
DONATION	Input	Interval
FLOW	Input	Interval
FRUITS_VEGGIES	Input	Interval
GENDER	Input	Nominal
LIVE_VISION	Input	Interval
LOST_VACATION	Input	Interval
PERSONAL_AWARDS	Input	Interval
PLACES_VISITED	Input	Interval
SLEEP_HOURS	Input	Interval
SOCIAL_NETWORK	Input	Interval
SUFFICIENT_INCOME	Input	Interval
SUPPORTING_OTHERS	Input	Interval
TIME_FOR_PASSION	Input	Interval
TODO_COMPLETED	Input	Interval
VAR1	Time ID	Interval
WEEKLY_MEDITATION	Input	Interval
WORK_LIFE_BALANCE_SCORE	Target	Interval

Figure 11: Target Variable Selection

b) Removing missing value

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	AGE	INPUT	4	0	21 to 35	38.24	36 to 50	29.14
TRAIN	GENDER	INPUT	2	0	Female	61.72	Male	38.28

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
ACHIEVEMENT	INPUT	4.000751	2.755837	15972	0	0	3	10	0.628884	-0.35502
BMI_RANGE	INPUT	1.410656	0.491968	15972	0	1	1	2	0.363255	-1.86828
CORE_CIRCLE	INPUT	5.508077	2.840334	15972	0	0	5	10	0.202431	-0.9668
DAILY_SHOUTING	INPUT	2.930879	2.676301	15972	0	0	2	10	1.118024	0.535427
DAILY_STEPS	INPUT	5.703606	2.891013	15972	0	1	5	10	0.0393	-1.16792
DAILY_STRESS	INPUT	2.791685	1.367801	15971	1	0	3	5	-0.08053	-0.79838
DONATION	INPUT	2.715314	1.851586	15972	0	0	3	5	-0.04791	-1.43948
FLOW	INPUT	3.194778	2.357518	15972	0	0	3	10	0.870127	0.266194
FRUITS_VEGGIES	INPUT	2.922677	1.442694	15972	0	0	3	5	-0.05098	-1.00704
LIVE_VISION	INPUT	3.752129	3.230987	15972	0	0	3	10	0.740936	-0.56921
LOST_VACATION	INPUT	2.898886	3.69218	15972	0	0	0	10	0.920513	-0.66794
PERSONAL_AWARDS	INPUT	5.711558	3.08963	15972	0	0	5	10	0.061245	-1.17737
PLACES_VISITED	INPUT	5.23297	3.311912	15972	0	0	5	10	0.170494	-1.25206
SLEEP_HOURS	INPUT	7.042888	1.199044	15972	0	1	7	10	-0.3558	1.172863
SOCIAL_NETWORK	INPUT	6.474267	3.086672	15972	0	0	6	10	-0.17518	-1.3704
SUFFICIENT_INCOME	INPUT	1.728963	0.444509	15972	0	1	2	2	-1.03031	-0.93857
SUPPORTING_OTHERS	INPUT	5.616454	3.242021	15972	0	0	5	10	0.055665	-1.27058
TIME_FOR_PASSION	INPUT	3.326572	2.729293	15972	0	0	3	10	0.841172	-0.19397
TODO_COMPLETED	INPUT	5.745993	2.624097	15972	0	0	6	10	-0.35608	-0.70454
WEEKLY_MEDITATION	INPUT	6.233346	3.016571	15972	0	0	7	10	-0.19252	-1.15012
WORK_LIFE_BALANCE_SCORE	TARGET	666.7515	45.01987	15972	0	480	667.7	820.2	-0.10467	-0.10842

Figure 12: Summary of Variable Statistics

The "DAILY_STRESS" column has the presence of 1 missing value when exploring the dataset.

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
DAILY_STRESS	MEAN	IMP_DAILY_STRESS	2.79168	INPUT	INTERVAL		1

Figure 13: Summary of Imputation

By using a mean imputation approach to fill up the value for the only missing value. As this method is chosen because the mean is a representative value of the central tendency of the data, which can help to preserve the overall distribution and relationship between variables in the dataset.

c) Dropping low variable worth column

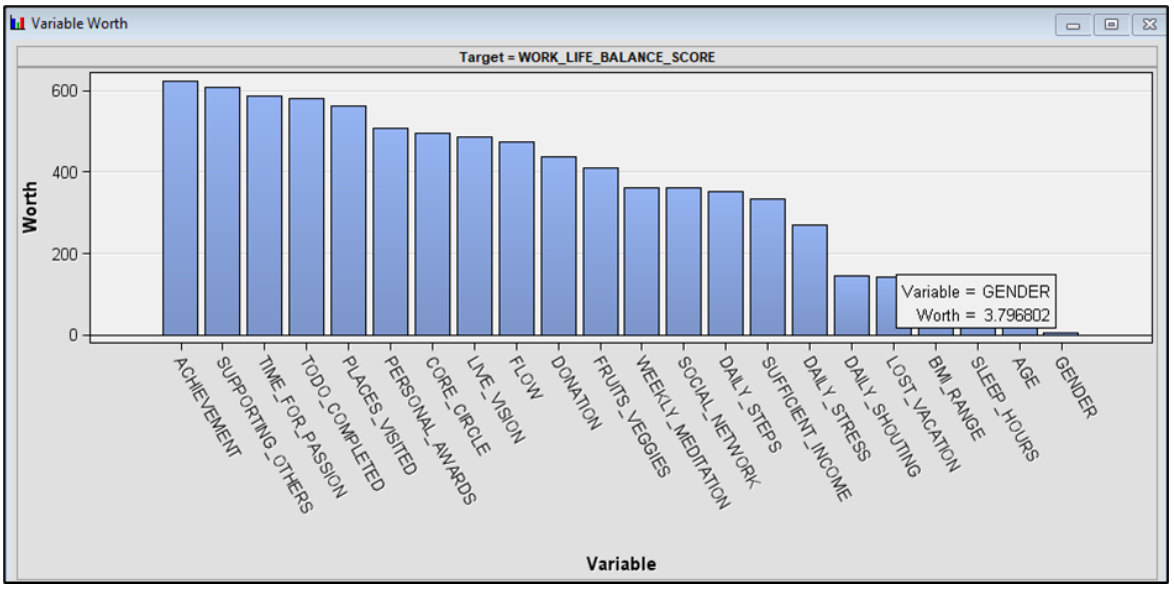


Figure 14: Variable Worth Graph

As the "GENDER" variable having the lowest worth value in the graph, the group decided to drop the variable as low variable worth variables may contain redundant information and therefore dropping it can reduce noise. Besides, the group is interested in discovering the characteristics of individuals who have a work-life balance lifestyle. Thus, dropping the "GENDER" variable will not have any impact on further analysis.

d) Dropping low variance column

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
ACHIEVEMENT	INPUT	4.000751	2.755837	15972	0	0	3	10	0.628884	-0.35502
BMI_RANGE	INPUT	1.410656	0.491968	15972	0	1	1	2	0.363255	-1.86828
CORE_CIRCLE	INPUT	5.508077	2.840334	15972	0	0	5	10	0.202431	-0.9668
DAILY_SHOUTING	INPUT	2.930879	2.676301	15972	0	0	2	10	1.118024	0.535427
DAILY_STEPS	INPUT	5.703606	2.891013	15972	0	1	5	10	0.0393	-1.16792
DAILY_STRESS	INPUT	2.791685	1.367801	15971	1	0	3	5	-0.08053	-0.79838
DONATION	INPUT	2.715314	1.851586	15972	0	0	3	5	-0.04791	-1.43948
FLOW	INPUT	3.194778	2.357518	15972	0	0	3	10	0.870127	0.266194
FRUITS_VEGGIES	INPUT	2.922677	1.442694	15972	0	0	3	5	-0.05098	-1.00704
LIVE_VISION	INPUT	3.752129	3.230987	15972	0	0	3	10	0.740936	-0.56921
LOST_VACATION	INPUT	2.898886	3.69218	15972	0	0	0	10	0.920513	-0.66794
PERSONAL_AWARDS	INPUT	5.711558	3.08963	15972	0	0	5	10	0.061245	-1.17737
PLACES_VISITED	INPUT	5.23297	3.311912	15972	0	0	5	10	0.170494	-1.25206
SLEEP_HOURS	INPUT	7.042888	1.199044	15972	0	1	7	10	-0.3558	1.172863
SOCIAL_NETWORK	INPUT	6.474267	3.086672	15972	0	0	6	10	-0.17518	-1.3704
SUFFICIENT_INCOME	INPUT	1.728963	0.444509	15972	0	1	2	2	-1.03031	-0.93857
SUPPORTING_OTHERS	INPUT	5.616454	3.242021	15972	0	0	5	10	0.055665	-1.27058
TIME_FOR_PASSION	INPUT	3.326572	2.729293	15972	0	0	3	10	0.841172	-0.19397
TODO_COMPLETED	INPUT	5.745993	2.624097	15972	0	0	6	10	-0.35608	-0.70454
WEEKLY_MEDITATION	INPUT	6.233346	3.016571	15972	0	0	7	10	-0.19252	-1.15012
WORK_LIFE_BALANCE_SCORE	TARGET	666.7515	45.01987	15972	0	480	667.7	820.2	-0.10467	-0.10842

Figure 15: Summary of Variable Statistics

Name	Drop	Role	Level
ACHIEVEMENT	Default	Input	Interval
AGE	Default	Input	Nominal
BMI_RANGE	Default	Input	Interval
CORE_CIRCLE	Default	Input	Interval
DAILY_SHOUTING	Default	Input	Interval
DAILY_STEPS	Default	Input	Interval
DONATION	Default	Input	Interval
FLOW	Default	Input	Interval
FRUITS_VEGGIES	Default	Input	Interval
GENDER	Yes	Input	Nominal
IMP_DAILY_STRESS	Default	Input	Interval
LIVE_VISION	Default	Input	Interval
LOST_VACATION	Default	Input	Interval
PERSONAL_AWARDS	Default	Input	Interval
PLACES_VISITED	Default	Input	Interval
SLEEP_HOURS	Default	Input	Interval
SOCIAL_NETWORK	Default	Input	Interval
SUFFICIENT_INCOME	Yes	Input	Interval
SUPPORTING_OTHERS	Default	Input	Interval
TIME_FOR_PASSION	Default	Input	Interval
TODO_COMPLETED	Default	Input	Interval
VAR1	Default	Time ID	Interval
WEEKLY_MEDITATION	Default	Input	Interval
WORK_LIFE_BALANCE_SCORE	Default	Target	Interval

Figure 16: Dropping Variable

"BMI_RANGE" and "SUFFICIENT_INCOME" variable having low variance value, in case of high number of dimensions, the variables should be dropped compared to others because this variable indicates that the values in the variable are relatively similar, which may not provide much useful information for the analysis.

e) Data Partition

Data Set Allocations	
Training	80.0
Validation	20.0
Test	0.0

Figure 17: Partitioning Data

The Figure above refers to the process of dividing the selected dataset into two parts, where 80% of the data is used for training and 20% of the data is used for validation. The purpose of this partitioning is to have a large enough sample for training the model and a smaller sample for evaluating its performance. This helps to ensure that the model is not overfitted or under fitted, and that it can generalise well to unseen data.

7.0 Model Construction & Optimization

7.1 Wong Kit Yee TP061080

The two models that I have decided to use to predict an individual's work life balance score based on various key characteristics are Regression and Neural Network.

7.1.1 Model 1 - Regression

A regression model utilises a straight line to assess the connection between one dependent variable and one or more independent variables. It has two types of prediction formulas which are Linear Regression Prediction Formula to predict a continuous target variable and Logistic Regression Prediction Formula used by categorical target variable. In this project, Linear Regression will be used due to the data type of the target variable is interval.

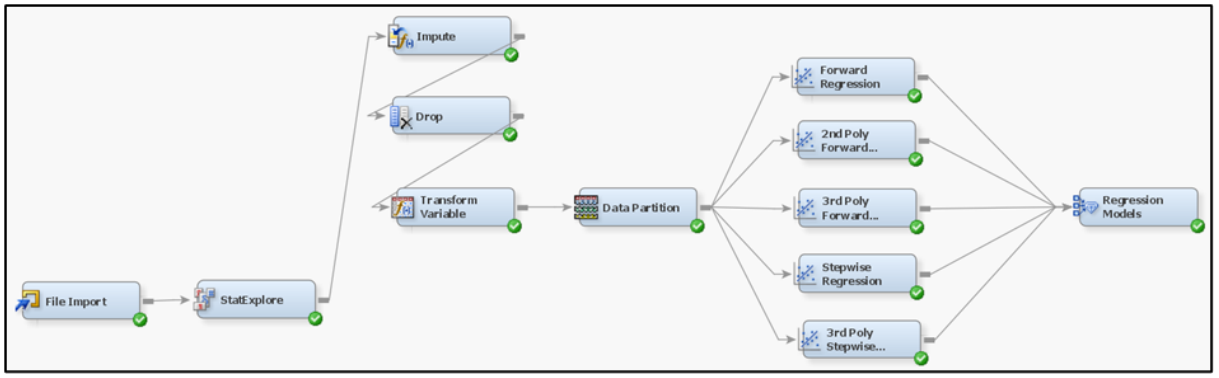


Figure 18: Regression Models Construction & Optimization

Initially, five different regression models were constructed in order to find out which model performed better with the lowest ASE values. Each regression model's properties is shown below:

Forward Regression

.. Property	Value
General	
Node ID	Reg3
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour

Figure 19: Configuration of Forward Regression

Polynomial Terms have been set to “No”. Linear Regression method has been selected, and the Link function being used is Logit. The model selection method is Forward, with the Validation Error criterion being used to evaluate model performance.

2nd & 3rd Degree Polynomial Forward Regression

.. Property	Value
General	
Node ID	Reg4
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Yes
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour

.. Property	Value
General	
Node ID	Reg5
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Yes
Polynomial Degree	3
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour

Figure 20: Configuration of 2nd and 3rd Degree Polynomial Forward Regression

The modifications involve enabling polynomial terms by selecting "Yes” and adjusting the Polynomial Degree to the respective degree value.

Stepwise Regression

.. Property	Value
General	
Node ID	Reg6
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour

Figure 21: Configuration of Stepwise Regression

Major properties are the same with Forward Regression. The only changes made is the selection model will be Stepwise.

3rd Degree Polynomial Stepwise Regression

.. Property	Value
General	
Node ID	Reg7
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Yes
Polynomial Degree	3
User Terms	No
Term Editor	
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour

Figure 22: Configuration of 3rd Degree Polynomial Stepwise Regression

The modifications involve enabling polynomial terms by selecting "Yes" and adjusting the Polynomial Degree to 3rd degree.

7.1.2 Model 2 - Neural Network

The Neural Network model is a regression model extension. Since the target variable is continuous, it can show better accuracy when using neural networks.

A neural network's fundamental structure is made up of layers of linked nodes, or neurons, that are organised in a specific way to perform a specific task. A feedforward neural network, including an input layer, one or even more hidden layers, and an output layer. Hidden layers are used to process the input data. The output layer generates the neural network's final outcome, normally a forecast or categorization. The basic structure of a neural network can be further customised by choosing different activation functions, loss functions, and optimization algorithms.

However, it does not have a method to select inputs. There's no feature selection method, but it's mandatory. Therefore, a 2nd Poly Forward Regression Model has been selected as an external process to select useful inputs.

5 different Neural Networks have been created and compared.

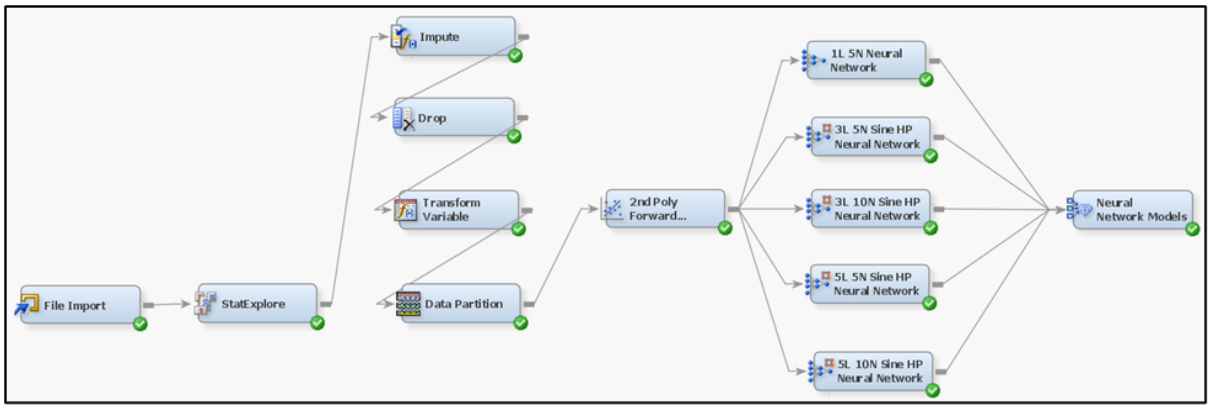


Figure 23: Neural Network Models Construction & Optimization

To optimise the neural networks, the preliminary training has been disabled since there is no further preprocessing required.

Single Layer Neural Network with 5 Hidden Neurons

Network

.. Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	5
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default
Target Layer Error Function	Default
Target Bias	Yes
Weight Decay	0.0

Optimization

.. Property	Value
Training Technique	Default
Maximum Iterations	1000
Maximum Time	5 Minutes
Nonlinear Options	
Use Defaults	Yes
Absolute	-1.34078E154
Absolute Function	0
Absolute Function Times	1
Absolute Gradient	1.0E-5
Absolute Gradient Times	1
Absolute Parameter	1.0E-8
Absolute Parameter Times	1
Relative Function	0.0
Relative Function Times	1
Relative Gradient	1.0E-6
Relative Gradient Times	1
Propagation Options	
Accelerate	1.2
Decelerate	0.5
Learn	0.1
Maximum Learning	50.0
Minimum Learning	1.0E-5
Momentum	0.0
Maximum Momentum	1.75
Tilt	0.0
Preliminary Training	
Enable	No
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

Figure 24: Configuration of Single Layer Neural Network with 5 Hidden Neurons

In the network property under the train section, the number of hidden units/neurons has been set to 5. Optimization of maximum iterations and time will be 1000 iterations and 5 minutes.

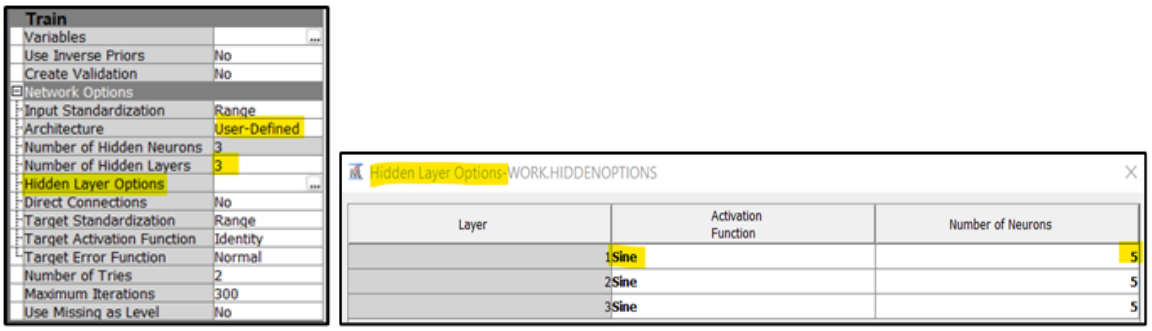
HP Neural Networks

For all the HP Neural Network Models, the network architecture is set to user defined. The number and options of hidden layers could be further customised based on each model's characteristics. The activation function chosen for the HP Neural Models is Sine function.

.. Property	Value
General	
Node ID	HPNNA2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	User-Defined
Number of Hidden Neurons	3
Number of Hidden Layers	3
Hidden Layer Options	
Direct Connections	No
Target Standardization	Range
Target Activation Function	Identity
Target Error Function	Normal
Number of Tries	2
Maximum Iterations	300
Use Missing as Level	No

Figure 25: Configuration of all HP Neural Networks

HP Neural Network with 3 Hidden Layers & 5 Hidden Neurons



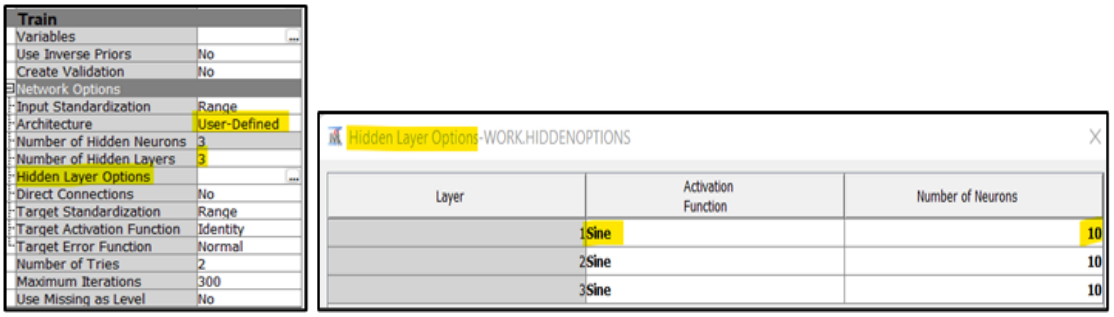
Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	User-Defined
Number of Hidden Neurons	3
Number of Hidden Layers	3
Hidden Layer Options	
Direct Connections	No
Target Standardization	Range
Target Activation Function	Identity
Target Error Function	Normal
Number of Tries	2
Maximum Iterations	300
Use Missing as Level	No

Hidden Layer Options-WORK.HIDDENOPTIONS

Layer	Activation Function	Number of Neurons
1	Sine	5
2	Sine	5
3	Sine	5

Figure 26: Configuration of 3 Hidden Layers HP Neural Network with 5 Hidden Neurons

HP Neural Network with 3 Hidden Layers & 10 Hidden Neurons



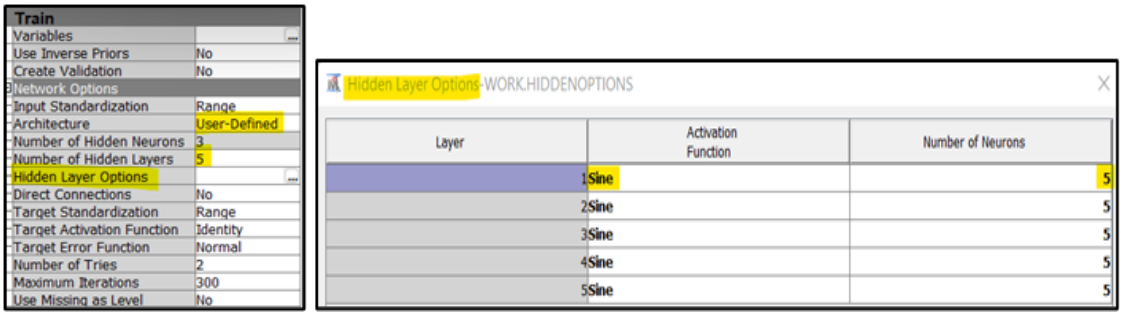
Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	User-Defined
Number of Hidden Neurons	3
Number of Hidden Layers	3
Hidden Layer Options	
Direct Connections	No
Target Standardization	Range
Target Activation Function	Identity
Target Error Function	Normal
Number of Tries	2
Maximum Iterations	300
Use Missing as Level	No

Hidden Layer Options-WORK.HIDDENOPTIONS

Layer	Activation Function	Number of Neurons
1	Sine	10
2	Sine	10
3	Sine	10

Figure 27: Configuration of 3 Hidden Layers HP Neural Network with 10 Hidden Neurons

HP Neural Network with 5 Hidden Layers & 5 Hidden Neurons



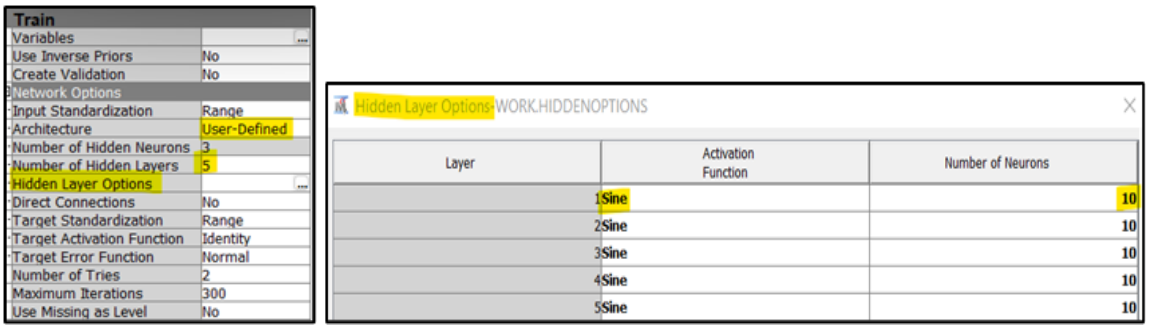
Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	User-Defined
Number of Hidden Neurons	3
Number of Hidden Layers	5
Hidden Layer Options	
Direct Connections	No
Target Standardization	Range
Target Activation Function	Identity
Target Error Function	Normal
Number of Tries	2
Maximum Iterations	300
Use Missing as Level	No

Hidden Layer Options-WORK.HIDDENOPTIONS

Layer	Activation Function	Number of Neurons
1	Sine	5
2	Sine	5
3	Sine	5
4	Sine	5
5	Sine	5

Figure 28: Configuration of 5 Hidden Layers HP Neural Network with 5 Hidden Neurons

HP Neural Network with 5 Hidden Layers & 10 Hidden Neurons



Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	User-Defined
Number of Hidden Neurons	3
Number of Hidden Layers	5
Hidden Layer Options	
Direct Connections	No
Target Standardization	Range
Target Activation Function	Identity
Target Error Function	Normal
Number of Tries	2
Maximum Iterations	300
Use Missing as Level	No

Hidden Layer Options-WORK.HIDDENOPTIONS

Layer	Activation Function	Number of Neurons
1	Sine	10
2	Sine	10
3	Sine	10
4	Sine	10
5	Sine	10

Figure 29: Configuration of 5 Hidden Layers HP Neural Network with 10 Hidden Neurons

7.2 Chee Wei Keng TP063861

The target variable in this assignment is “WORK_LIFE_BALANCE_SCORE”. To create the model to identify the traits of individuals with high work-life balance, a descriptive model -- clustering and a predictive model -- neural network was chosen.

7.2.1 Model 1 - Clustering

Clustering is a sort of unsupervised learning in which the data is grouped into clusters on the basis of the similarity of their characteristics. The aim is to partition the data into clusters (groups) such that data points inside a cluster are distinct from those in other clusters and comparable to one another. Clustering is an unsupervised learning technique, indicating it does not require labelled data and is used to uncover hidden patterns in the information (What Is Clustering? | Clustering in Machine Learning, 2019). Utilizing clustering is to group individuals with similar characteristics into clusters, and then analyse each cluster to determine what factors contribute to a high work-life balance score.

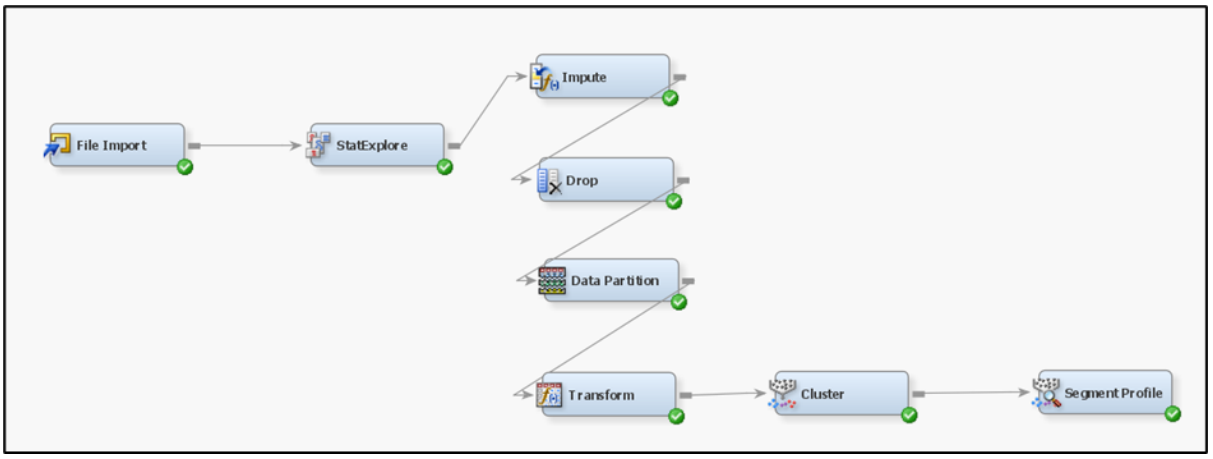


Figure 30: Clustering Model

Train	
Variables	
Internal Standardization	Standardization
Number of Clusters	
Specification Method	User Specify
Maximum Number of Clusters	6

Figure 31: Specifying the number of clusters

Referring to the figure above, after the team has completed all exploration and pre-processing of the dataset, cluster nodes are added to the process. A green check next to a cluster node

indicates that it is running successfully. Also, segment profile nodes are connected to view the following cluster chart. The standardisation is set and 6 clusters will be generated.

7.2.2 Model 2 - Neural Network (5L 5N HPNN)

Neural networks are a type of machine learning model that are inspired by the structure and function of the human brain. They consist of layers of interconnected nodes, also known as neurons, which can learn complex patterns in data through a process called training. Referring to the Figure above, it displays the construction of a neural network. To choose useful inputs, an external procedure called the forward regression model has been chosen.

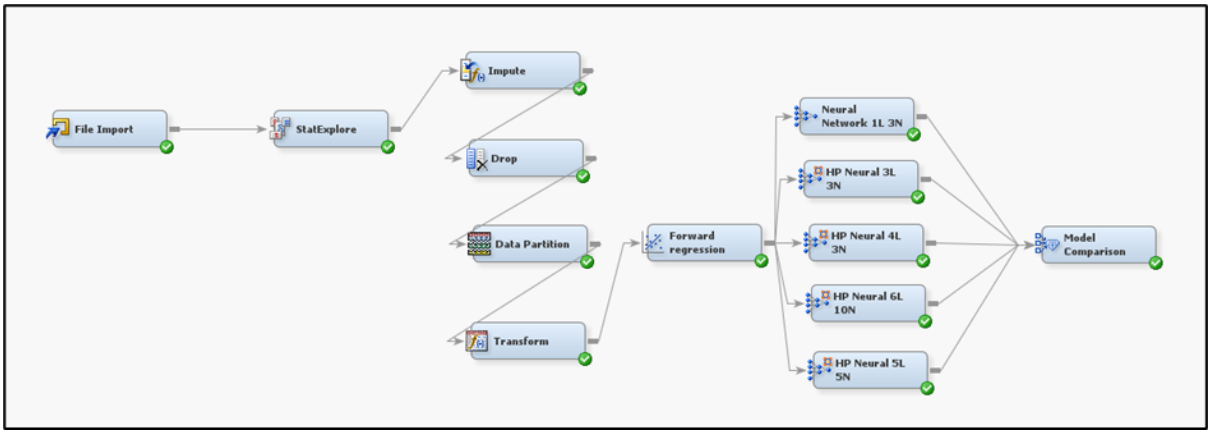


Figure 32: HP Neural Network with 5L 5N Model

Based on the above figure, a 5-layer, 5-neuron HP neural network was constructed. The reason for specifying 5 neurons is from the formula, sqrt (input layer nodes * output layer nodes). In this project, the input layer node is 23 and the output layer node is 1. Since the square root of 23 is 4.7, it can be considered as 5.

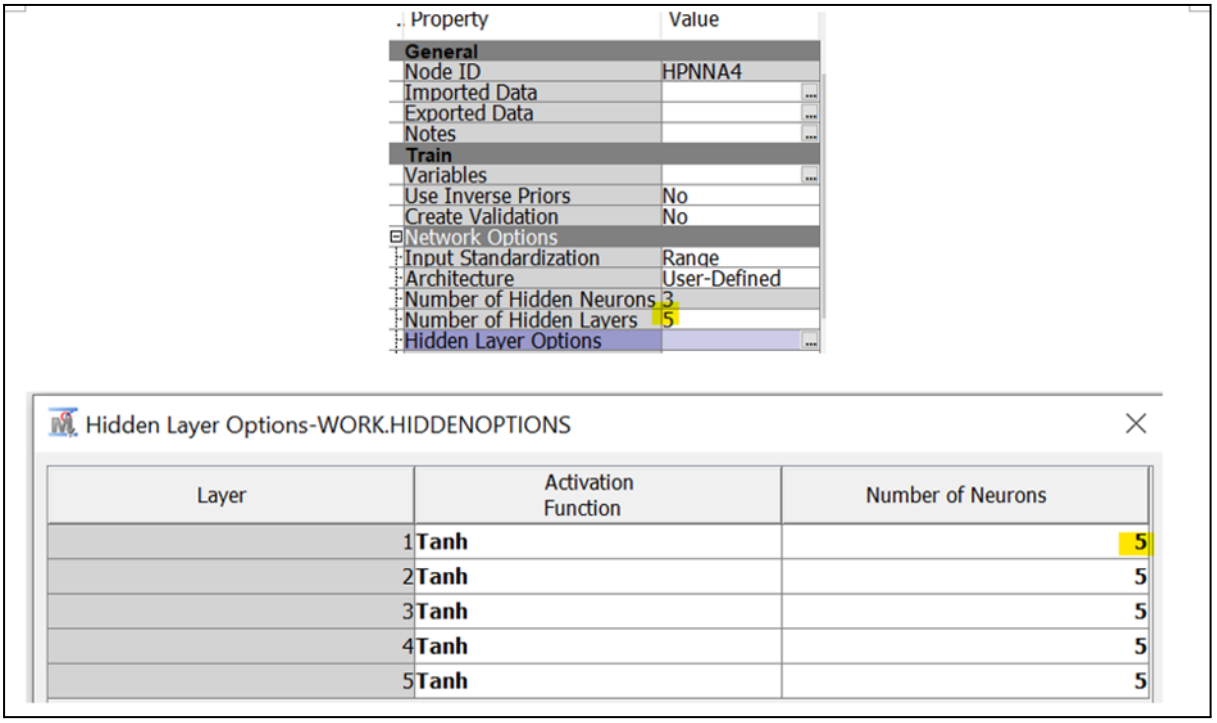
```
                                Optimization Results

Iterations                                  76  Function Calls                    87
Jacobian Calls                              79  Active Constraints                 0
Objective Function                50.310642145  Max Abs Gradient Element  43.378076009
Lambda                           0.0000519022  Actual Over Pred Change  0.2113835171
Radius                           5.9703872468

Convergence criterion (FCONV=0.0001) satisfied.

NOTE: At least one element of the gradient is greater than 1e-3.
```

Figure 33: Optimization Result of 1L 3N Neural Network Model

The Figure above shows the optimization results. In the 1-layer 3 neurons model, the preliminary training has been deactivated because there has been no additional pre-processing

to optimise the neural network. The 1000 iterations and 5 minutes options were selected in the Optimization tab. The others neural networks were constructed to do comparison between neural network model as follows: --

HP Neural: 5 Hidden Layers & 5 Hidden Neurons



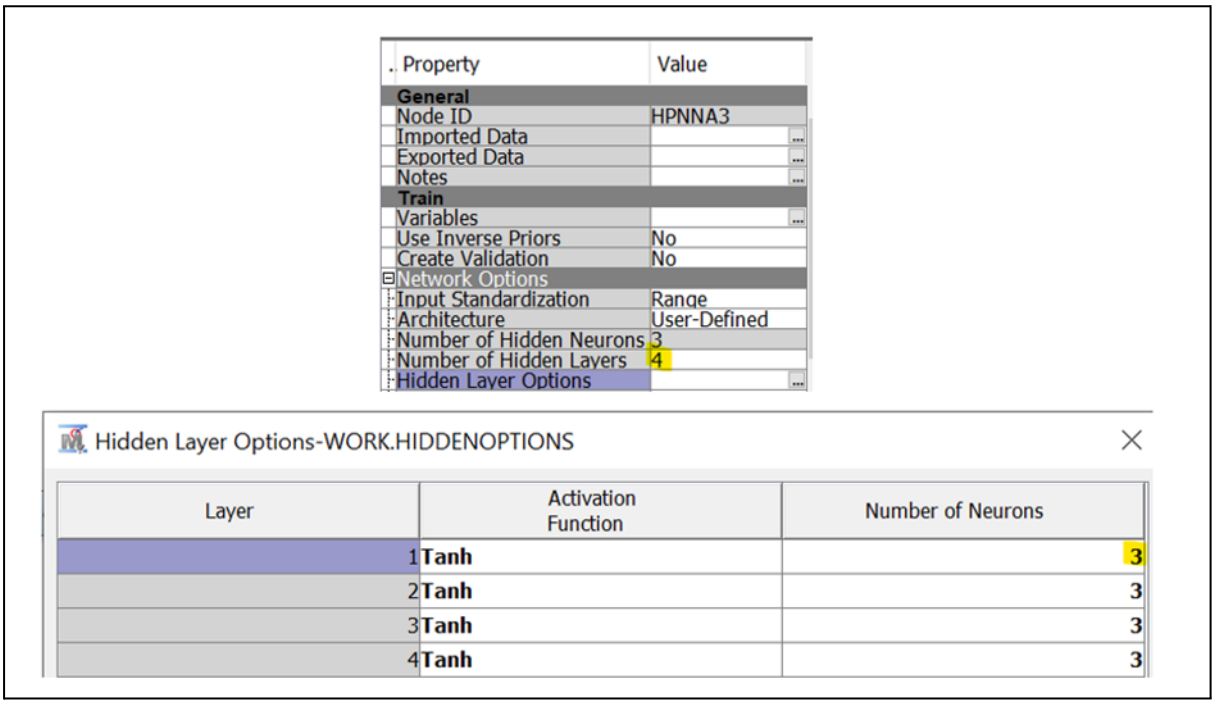
Property	Value
General	
Node ID	HPNNA4
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	User-Defined
Number of Hidden Neurons	3
Number of Hidden Layers	5
Hidden Layer Options	

Hidden Layer Options-WORK.HIDDENOPTIONS

Layer	Activation Function	Number of Neurons
1	Tanh	5
2	Tanh	5
3	Tanh	5
4	Tanh	5
5	Tanh	5

Figure 34: Constructing HP Neural Network with 5L 5N

HP Neural: 4 Hidden Layers & 3 Hidden Neurons



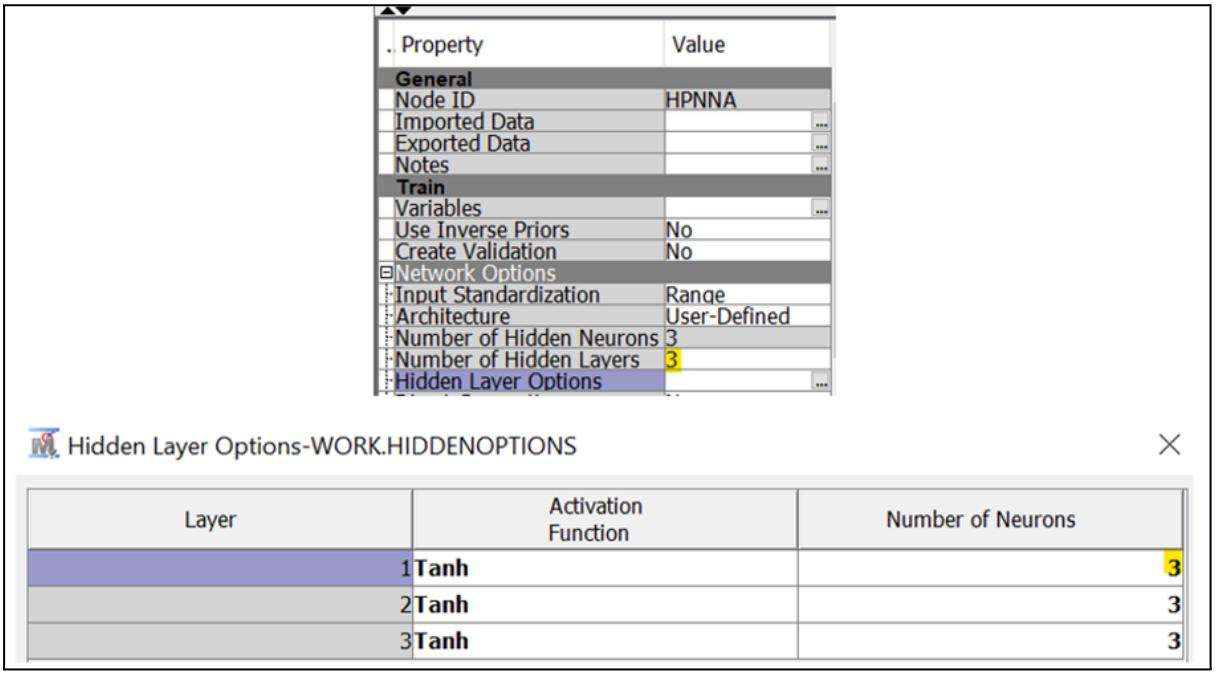
Property	Value
General	
Node ID	HPNNA3
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	User-Defined
Number of Hidden Neurons	3
Number of Hidden Layers	4
Hidden Layer Options	

Hidden Layer Options-WORK.HIDDENOPTIONS

Layer	Activation Function	Number of Neurons
1	Tanh	3
2	Tanh	3
3	Tanh	3
4	Tanh	3

Figure 35: Constructing HP Neural Network with 4L 3N

HP Neural: 3 Hidden Layers & 3 Hidden Neurons



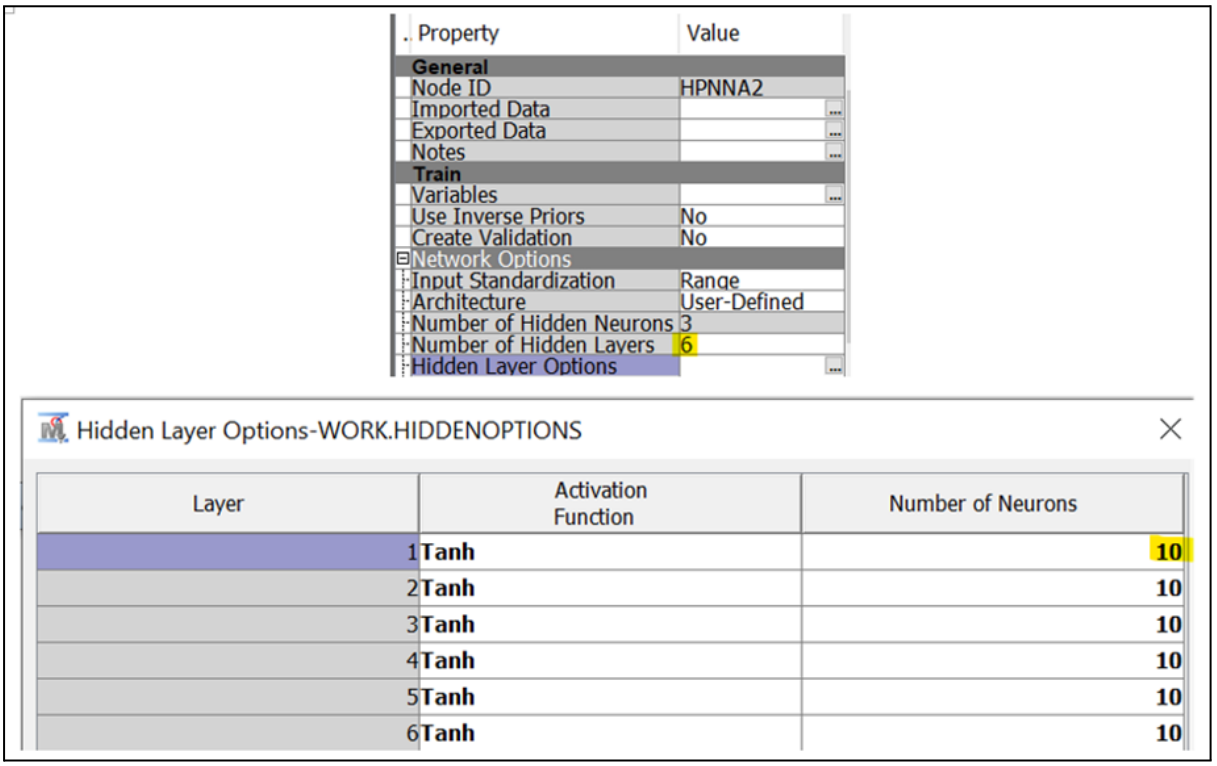
Property	Value
General	
Node ID	HPNNA
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	User-Defined
Number of Hidden Neurons	3
Number of Hidden Layers	3
Hidden Layer Options	

Hidden Layer Options-WORK.HIDDENOPTIONS

Layer	Activation Function	Number of Neurons
1	Tanh	3
2	Tanh	3
3	Tanh	3

Figure 36: Constructing HP Neural Network with 3L 3N

HP Neural 6 Hidden Layers & 10 Hidden Neurons



Property	Value
General	
Node ID	HPNNA2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Use Inverse Priors	No
Create Validation	No
Network Options	
Input Standardization	Range
Architecture	User-Defined
Number of Hidden Neurons	3
Number of Hidden Layers	6
Hidden Layer Options	

Hidden Layer Options-WORK.HIDDENOPTIONS

Layer	Activation Function	Number of Neurons
1	Tanh	10
2	Tanh	10
3	Tanh	10
4	Tanh	10
5	Tanh	10
6	Tanh	10

Figure 37: Constructing HP Neural Network with 6L 10N

7.3 Tan Yong Hong TP063960

The main model category that I've chosen to predict an individual's work-life balance score based on the available input variables is regressions. This is because regression models can help me to identify the most important variables and the relationship between variables. From there, individuals can use the evidence-based output of regression models to identify the parts they need to improve to improve a better work-life balance.

7.3.1 Model 1 & 2- Regressions & HP Regressions

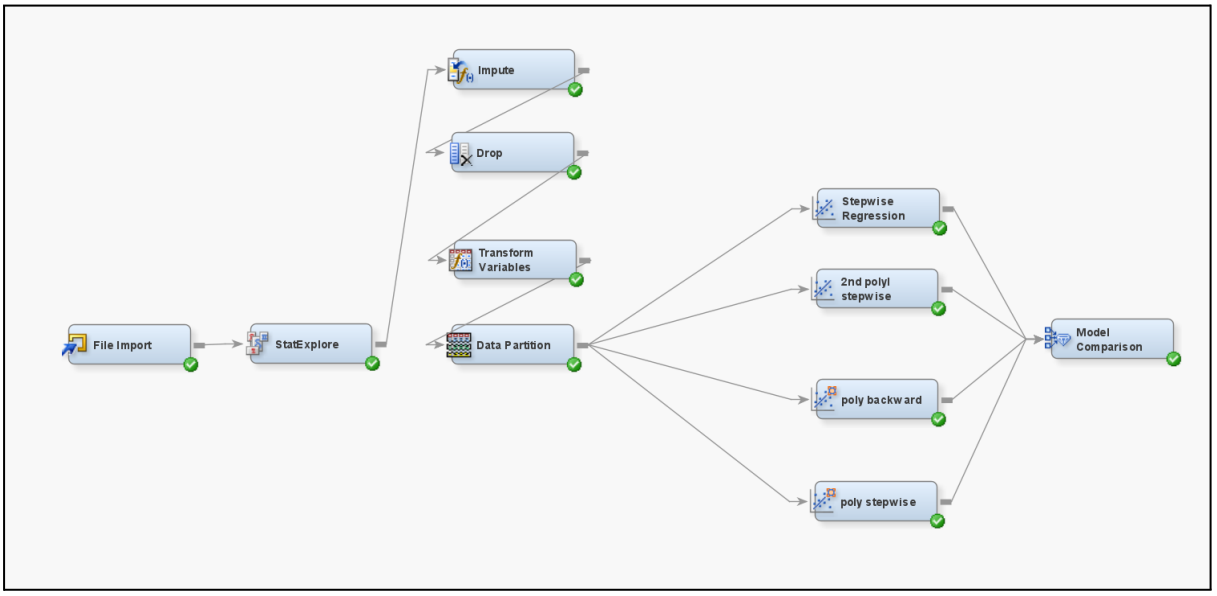


Figure 38: SAS Diagram of The Construction & Optimization of Regression Models

Regression models can provide functions that literate the relationship between the independent variable(s) and a target variable. Depending on the data type of the target variable, the prediction formula that's implemented can vary. In this case, Work-Life Balance Score, our target variable is an interval variable, hence linear regression is being used as the prediction formula, whereas logistic regression will be used when the target variable is categorical.

As demonstrated in Figure 32, 4 regression models were constructed to dictate the 2 most precise regression models, 2 using usual regression nodes, while the other 2 using HP regression nodes, which can be checked by comparing Average Square Error (ASE) values, the lower the better. In all regression models, there are no preliminary training involved as pre-processing has been completed prior to this step and Validation Error has been selected as

the criterion to evaluate each model. The properties of each regression model are demonstrated one after one.

Normal Regressions - Stepwise Regression & 2nd Degree Polynomial Stepwise Regression

.. Property	Value
General	
Node ID	Reg5
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interaction	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	...

Figure 39: Configuration of Stepwise Regression

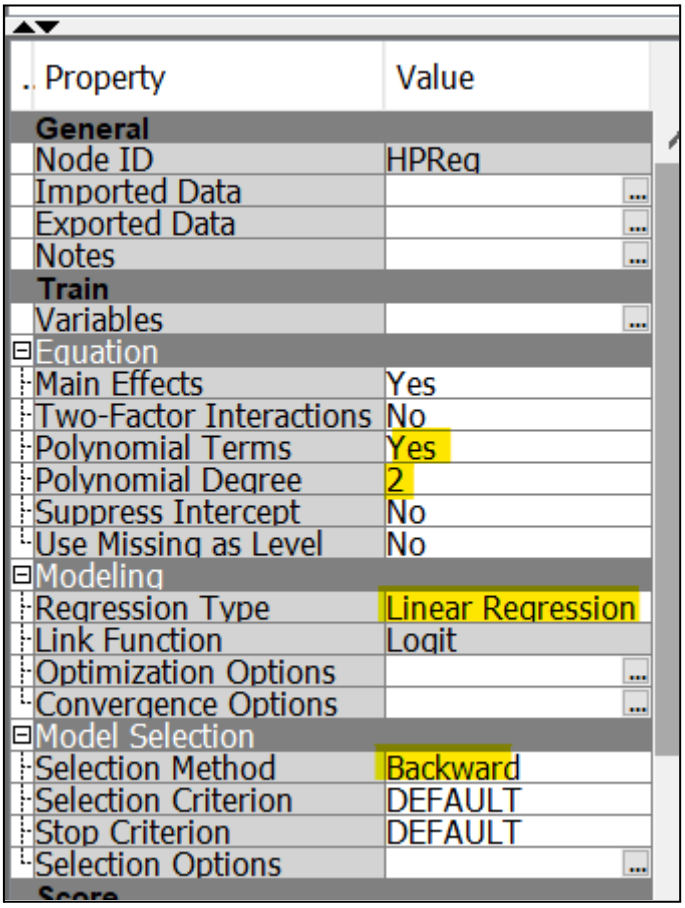
Key properties of the stepwise regression applied can be viewed in Figure 33, with polynomial terms set to no, linear regression being the regression type along with Logit link function and setting selection model as stepwise.

Property	Value
General	
Node ID	Reg2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Yes
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Linear Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Validation Error
Use Selection Defaults	Yes
Selection Options	

Figure 40: Configuration of 2nd Degree Polynomial Stepwise Regression

Key properties of the 2nd degree polynomial stepwise regression applied can be viewed in Figure 34, with polynomial terms set to yes and its degree to 2, linear regression being the regression type along with Logit link function and stepwise being the selection model.

HP Regressions - 2nd Degree Polynomial Backward HP Regression & 2nd Degree Polynomial Stepwise HP Regression



Property	Value
General	
Node ID	HPReg
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Yes
Polynomial Degree	2
Suppress Intercept	No
Use Missing as Level	No
Modeling	
Regression Type	Linear Regression
Link Function	Logit
Optimization Options	
Convergence Options	
Model Selection	
Selection Method	Backward
Selection Criterion	DEFAULT
Stop Criterion	DEFAULT
Selection Options	

Figure 41: Configuration of 2nd Degree Polynomial Backward HP Regression

Key properties of the 2nd degree polynomial backward HP regression applied can be viewed in Figure 35, with polynomial terms set to yes and its degree to 2, linear regression being the regression type along with Logit link function and backward being the selection model.

Property	Value
General	
Node ID	HPReg2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	Yes
Polynomial Degree	2
Suppress Intercept	No
Use Missing as Level	No
Modeling	
Regression Type	Linear Regression
Link Function	Logit
Optimization Options	
Convergence Options	
Model Selection	
Selection Method	Stepwise
Selection Criterion	DEFAULT
Stop Criterion	DEFAULT
Selection Options	

Figure 42: Configuration of 2nd Degree Polynomial Stepwise HP Regression

Key properties of the 2nd degree polynomial stepwise HP regression applied can be viewed in Figure 36, with polynomial terms set to yes and its degree to 2, linear regression being the regression type along with Logit link function and stepwise being the selection model.

8.0 Model Validation

8.1 Wong Kit Yee TP061080

8.1.1 Model 1 - Regression

Regression Models Comparison

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label ▲	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function
Y	Reg7	Reg7	3rd Poly Stepwise Regression	WORK ...		50.10447	50127.95	50.283	50.283
	Reg4	Reg4	2nd Poly Forward Regression	WORK ...		50.24321	50181.81	50.43219	50.43219
	Reg5	Reg5	3rd Poly Forward Regression	WORK ...		50.2793	50152.74	50.30187	50.30187
	Reg3	Reg3	Forward Regression	WORK ...		50.3678	50303.02	51.06452	51.06452
	Reg6	Reg6	Stepwise Regression	WORK ...		50.3678	50303.02	51.06452	51.06452

Figure 43: Regression Models Comparison

The fit statistics presented indicate that the 3rd Degree Poly Stepwise Regression Model has the best performance with the lowest valid ASE value. It is considered the most suitable model for further analysis.

Fit Statistics & Average Squared Error Plot of 3rd Degree Poly Stepwise Regression Model

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
WORK LIFE...		AIC	Akaike's Information Criterion	50127.95	.
WORK LIFE...		ASE	Average Squared Error	50.283	50.10447
WORK LIFE...		AVERR	Average Error Function	50.283	50.10447
WORK LIFE...		DFE	Degrees of Freedom for Error	12744	.
WORK LIFE...		DFM	Model Degrees of Freedom	34	.
WORK LIFE...		DFT	Total Degrees of Freedom	12778	.
WORK LIFE...		DIV	Divisor for ASE	12778	3194
WORK LIFE...		ERR	Error Function	642516.1	160033.7
WORK LIFE...		FPE	Final Prediction Error	50.5513	.
WORK LIFE...		MAX	Maximum Absolute Error	17.73122	18.05232
WORK LIFE...		MSE	Mean Square Error	50.41715	50.10447
WORK LIFE...		NOBS	Sum of Frequencies	12778	3194
WORK LIFE...		NW	Number of Estimate Weights	34	.
WORK LIFE...		RASE	Root Average Sum of Squares	7.09105	7.078451
WORK LIFE...		RFPE	Root Final Prediction Error	7.109944	.
WORK LIFE...		RMSE	Root Mean Squared Error	7.100503	7.078451
WORK LIFE...		SBC	Schwarz's Bayesian Criterion	50381.43	.
WORK LIFE...		SSE	Sum of Squared Errors	642516.1	160033.7
WORK LIFE...		SUMW	Sum of Case Weights Times Freq	12778	3194

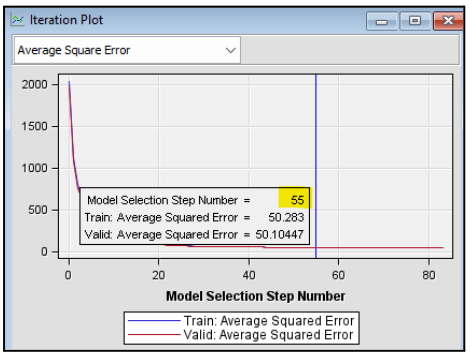


Figure 44: Fit Statistics & Average Squared Error Plot of 3rd Degree Poly Stepwise Regression Model

As seen from the fit statistics table, the 3rd Degree Poly Stepwise Regression Model is having a best average squared error of 0.02081 ((sqrt(50.1)) / (820-480)). This rate was yielded during the 55th iteration during the training, and it was able to predict the outcomes with a 2.08% error rate, in other words, a 97.92% success rate. There is no overfitting.

8.1.2 Model 2 - Neural Network

Neural Network Models Comparison

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Total Degrees of Freedom	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights	Train: Akaike's Information Criterion	Train: Schwarz's Bayesian Criterion	Train: Average Squared Error
Y	HPNNA4	HPNNA4	5L 10N Sine HP Neural Network	WORK ...		50.10585	.	.	.	.	.	.	50.69546
	HPNNA2	HPNNA2	3L 5N Sine HP Neural Network	WORK ...		50.16824	.	.	.	.	.	.	50.78601
	HPNNA3	HPNNA3	5L 5N Sine HP Neural Network	WORK ...		50.17163	.	.	.	.	.	.	50.58707
	HPNNA	HPNNA	3L 10N Sine HP Neural Network	WORK ...		50.19053	.	.	.	.	.	.	50.54604
	Neural	Neural	1L 5N Neural Network	WORK ...		51.71013	12778	12657	121	121	50863.16	51765.27	52.54064

Figure 45: Neural Network Models Comparison

The fit statistics presented indicate that the HP Neural Network with 5 Hidden Layers and 10 Hidden Neurons on each hidden layer produced the best performance with the lowest valid ASE value.

Fit Statistics & Average Squared Error Plot of HP Neural Network with 5 Hidden Layers and 10 Hidden Neurons

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
WORK LIFE BALANCE SCORE		ASE	Average S...	50.69546	50.10585
WORK LIFE BALANCE SCORE		DIV	Divisor for...	12778	3194
WORK LIFE BALANCE SCORE		MAX	Maximum ...	18.23692	17.14355
WORK LIFE BALANCE SCORE		NOBS	Sum of Fr...	12778	3194
WORK LIFE BALANCE SCORE		RASE	Root Aver...	7.120074	7.078548
WORK LIFE BALANCE SCORE		SSE	Sum of Sq...	647786.6	160038.1

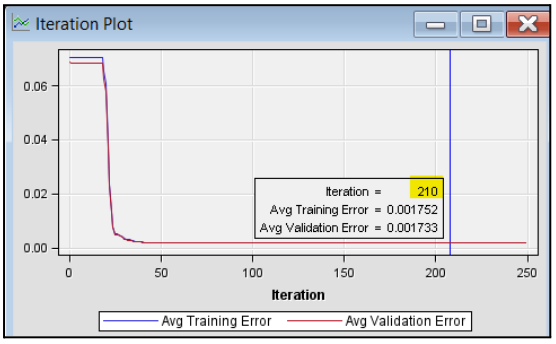


Figure 46: Fit Statistics & Average Squared Error Plot of HP Neural Network (5L 10N)

This rate was yielded during the 210th iteration during the training. There is no overfitting.

Link Graph of HP Neural Network with 5 Hidden Layers and 10 Hidden Neurons

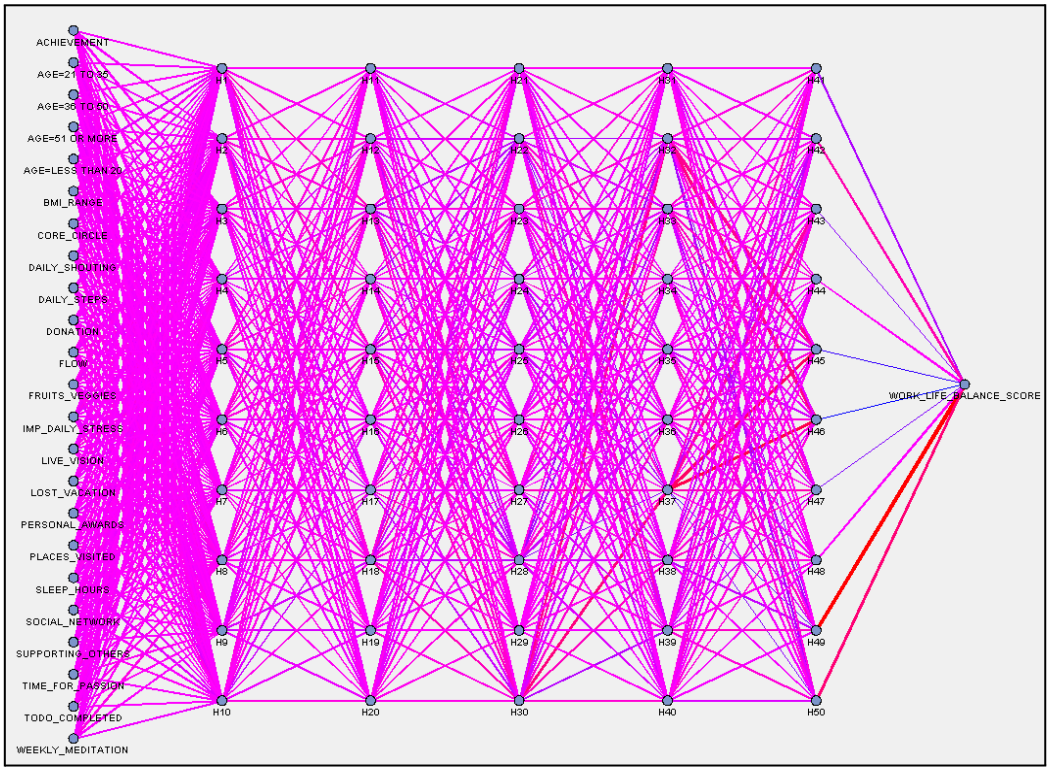


Figure 47: Link Graph of HP Neural Network (5L 10N)

After determining the best neural network model, the **Surrogate Tree** is used to explain the relationships between the input variables and the model's output in the HP Neural Network.

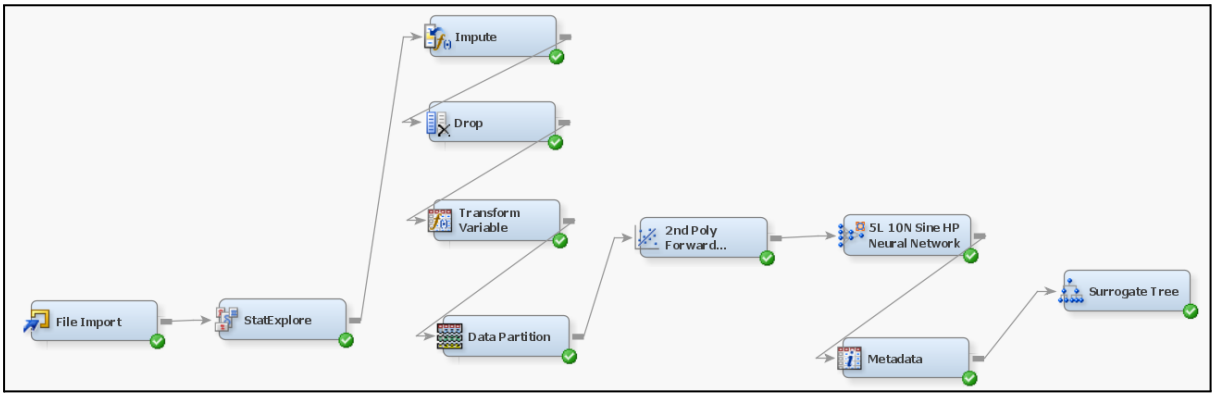


Figure 48: HP Neural Network (5L 10N) added with Surrogate Tree

Metadata

Variables - Meta

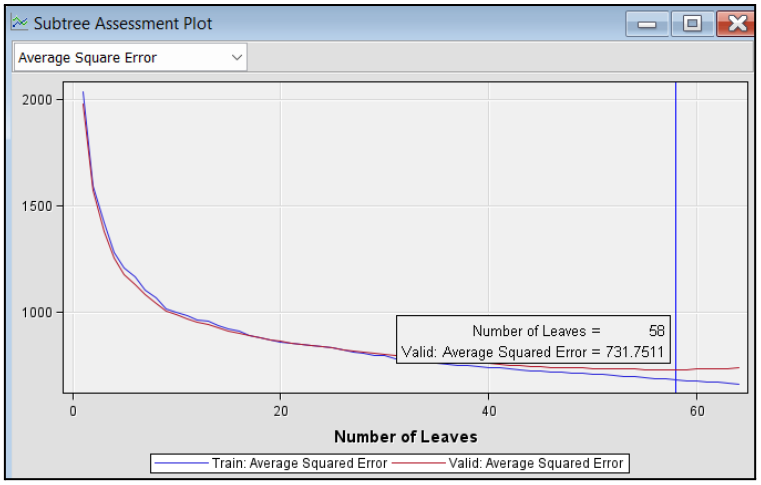
(none) | not | Equal to

Columns: Label | Mining | Basic

Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report	Model
ACHIEVEMENT	N	Default	Input	Default	Interval	Default	Default	Default	
AGE	N	Default	Input	Default	Nominal	Default	Default	Default	
BMI_RANGE	N	Default	Input	Default	Interval	Default	Default	Default	
CORE_CIRCLE	N	Default	Input	Default	Interval	Default	Default	Default	
DAILY_SHOUTING	N	Default	Input	Default	Interval	Default	Default	Default	
DAILY_STEPS	N	Default	Input	Default	Interval	Default	Default	Default	
DAILY_STRESS	Y	Default	Rejected	Default	Interval	Default	Default	Default	
DONATION	N	Default	Input	Default	Interval	Default	Default	Default	
FLOW	N	Default	Input	Default	Interval	Default	Default	Default	
FRUITS_VEGGIES	N	Default	Input	Default	Interval	Default	Default	Default	
GENDER	Y	Default	Rejected	Default	Nominal	Default	Default	Default	
IMP_DAILY_STRESS	N	Default	Input	Default	Interval	Default	Default	Default	
LIVE_VISION	N	Default	Input	Default	Interval	Default	Default	Default	
LOST_VACATION	N	Default	Input	Default	Interval	Default	Default	Default	
PERSONAL_AWARDS	N	Default	Input	Default	Interval	Default	Default	Default	
PLACES_VISITED	N	Default	Input	Default	Interval	Default	Default	Default	
P_WORK_LIFE_BALANCE_SCORE	N	Default	Prediction	Default	Interval	Default	Default	Default	
R_WORK_LIFE_BALANCE_SCORE	N	Default	Residual	Default	Interval	Default	Default	Default	
SLEEP_HOURS	N	Default	Input	Default	Interval	Default	Default	Default	
SOCIAL_NETWORK	N	Default	Input	Default	Interval	Default	Default	Default	
SUFFICIENT_INCOME	Y	Default	Rejected	Default	Interval	Default	Default	Default	
SUPPORTING_OTHERS	N	Default	Input	Default	Interval	Default	Default	Default	
TIME_FOR_PASSION	N	Default	Input	Default	Interval	Default	Default	Default	
TODO_COMPLETED	N	Default	Input	Default	Interval	Default	Default	Default	
VAR1	N	Default	Time ID	Default	Interval	Default	Default	Default	
WEEKLY_MEDITATION	N	Default	Input	Default	Interval	Default	Default	Default	
WORK_LIFE_BALANCE_SCORE	N	Default	Target	Default	Interval	Default	Default	Default	HPNNA4
WARN	N	Default	Assessment	Default	Nominal	Default	Default	Default	
dataobs	N	Default	ID	Default	Interval	Default	Default	Default	

Figure 49: Metadata of HP Neural Network (5L 10N)

Surrogate Tree



Subtree	
Method	N
Number of Leaves	58
Assessment Measure	Average Square Error
Assessment Fraction	0.25

Figure 50: Configuration of Surrogate Tree of HP Neural Network (5L 10N)

The Surrogate Tree has been set up with the chosen construction method for its subtrees determined by the N method. The total number of leaves specified for the tree is 58, and the assessment metric used to evaluate the performance of the subtree is the average square error.

8.2 Chee Wei Keng TP063861

8.2.1 Model 1 – Clustering

According to the Pathak (2020), the distance from cluster points is one similarity or dissimilarity metric used to analyse clusters. The clustering technique has been effective if it can group like observations together and separate dissimilar observations apart.

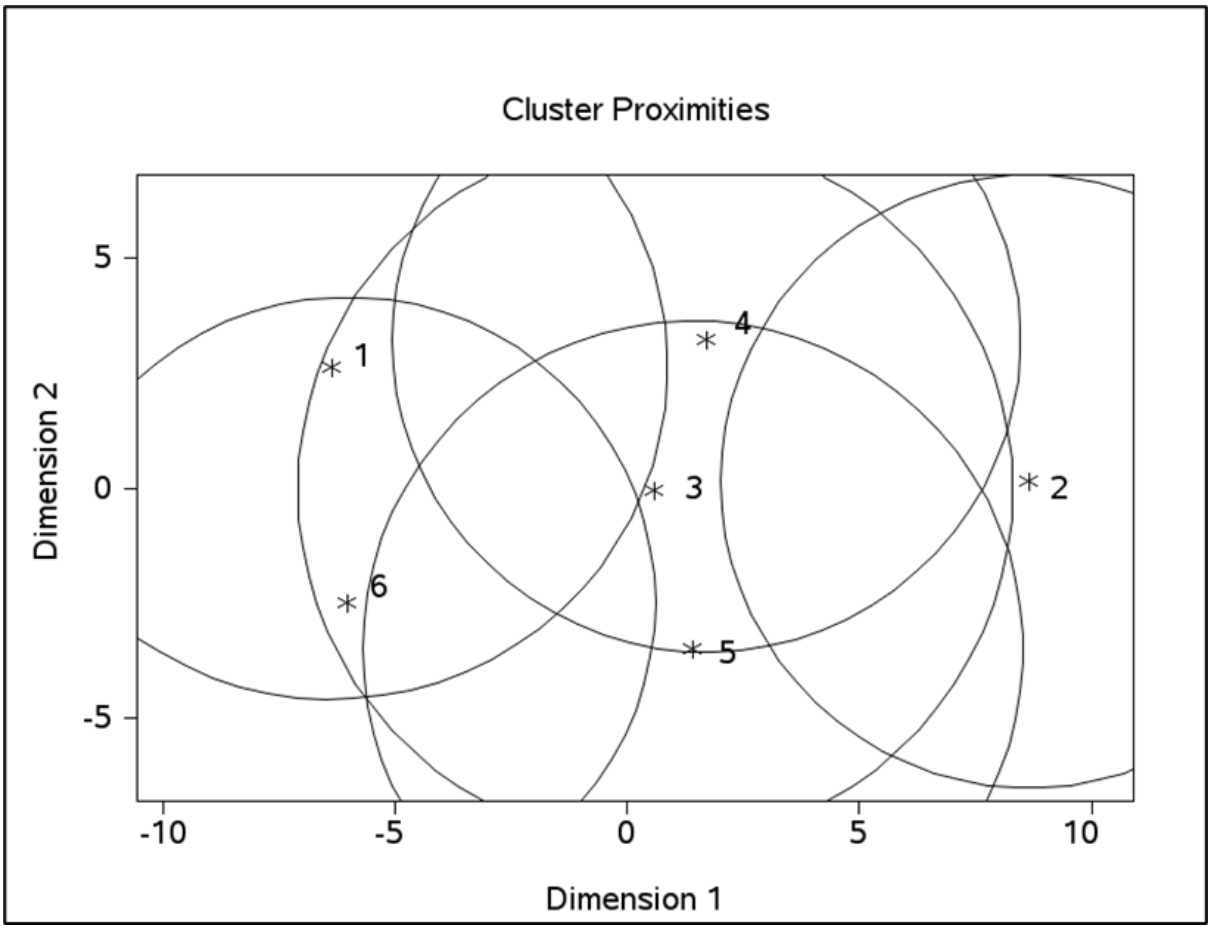


Figure 51: Cluster Distance Plot

Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
1	1725	0.8439...	7.22079	6	2.360219
2	2365	0.8211...	6.64983	4	2.948286
3	2148	0.8026...	7.704991	5	2.154607
4	2268	0.8523...	6.769009	3	2.34801
5	2321	0.7982...	7.118088	3	2.154607
6	1951	0.78426	6.625298	1	2.360219

Figure 52: Mean Statistics of Cluster Model

Therefore, the distance between cluster 3 and cluster 5 and the nearest cluster is 2.155, which is a minimum value. On the other hand, cluster 2 is the furthest from the other clusters with a distance of 2.948.

8.2.2 Model 2 – Neural Network

To find the neural network that performed the best regarding generating the lowest ASE numbers, five different layers of neural network were developed primarily. Below is the result of the model comparison.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Y	HPNNA4	HPNNA4	HP Neural 5L 5N	WORK LIFE BALANCE SCORE		50.1902
	HPNNA3	HPNNA3	HP Neural 4L 3N	WORK LIFE BALANCE SCORE		50.19495
	Neural	Neural	Neural Network 1L 3N	WORK LIFE BALANCE SCORE		50.25339
	HPNNA	HPNNA	HP Neural 3L 3N	WORK LIFE BALANCE SCORE		50.28997
	HPNNA2	HPNNA2	HP Neural 6L 10N	WORK LIFE BALANCE SCORE		50.29238

Figure 53: Statistics of Model Comparison

As a result, 5 hidden layers with 5 hidden neurons HP neural is selected to identify the characteristics of individuals with high work life balance score more effectively as it had the lowest ASE value of 50.1902.

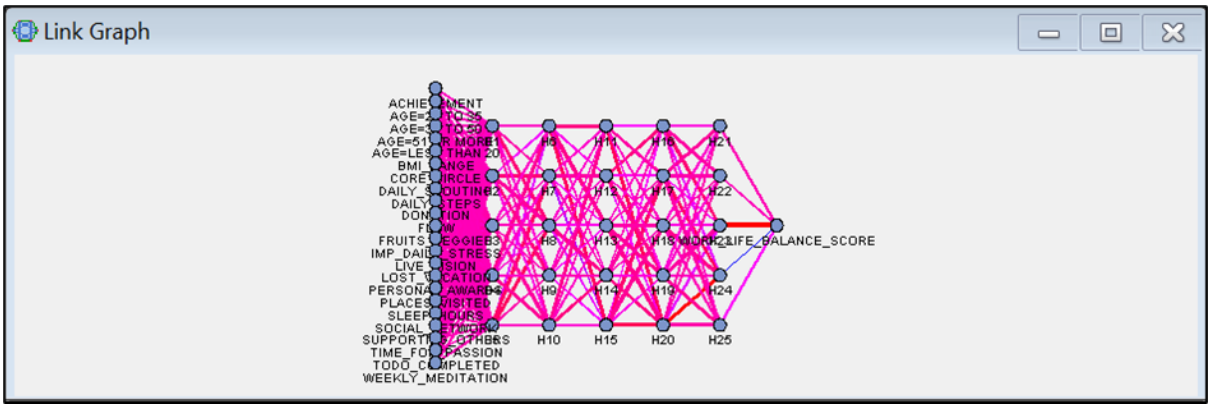


Figure 54: Link Graph

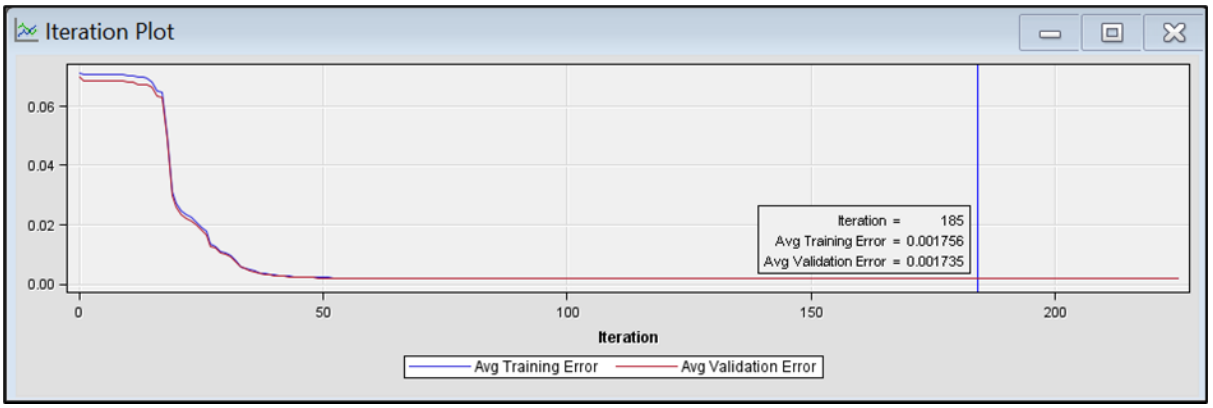


Figure 55: Iteration Plot

Since the target variable is WORK_LIFE_BALANCE_SCORE in this project. It is an interval variable ranging from 420 to 820. The accuracy can be calculated by square rooting the ASE then the result of ASE divided with the difference of the interval variable range or (calculated using the formula [{sqrt(50.1902)}/(820-480)]), thus the accuracy of this model is 98%.

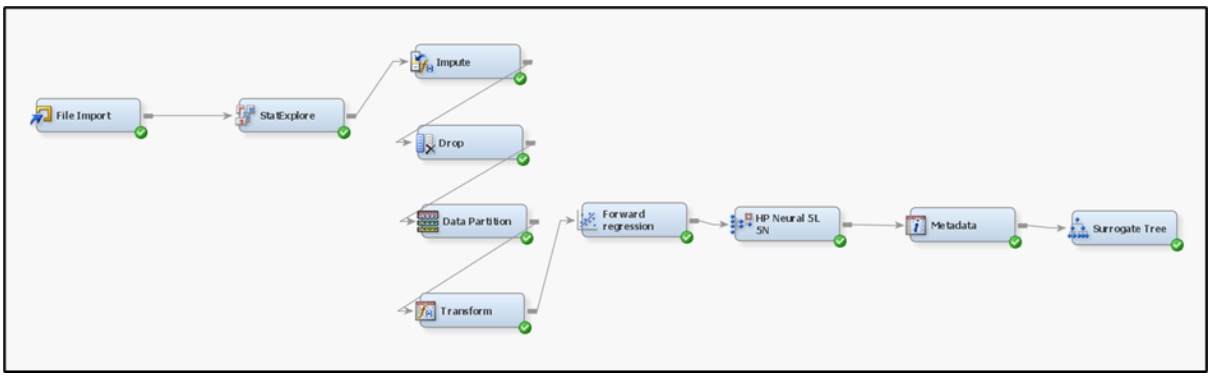


Figure 56: Add Metadata and Surrogate Tree node for 5L5N HP Neural Network Model

Variables - Meta

(none) not Equal to

Columns: Label Mining Basic Statistics

Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report	Model
ACHIEVEMENT	N	Default	Input	Default	Interval	Default	Default	Default	
AGE	N	Default	Input	Default	Nominal	Default	Default	Default	
BMI_RANGE	N	Default	Input	Default	Interval	Default	Default	Default	
CORE_CIRCLE	N	Default	Input	Default	Interval	Default	Default	Default	
DAILY_SHOUTING	N	Default	Input	Default	Interval	Default	Default	Default	
DAILY_STEPS	N	Default	Input	Default	Interval	Default	Default	Default	
DAILY_STRESS	Y	Default	Rejected	Default	Interval	Default	Default	Default	
DONATION	N	Default	Input	Default	Interval	Default	Default	Default	
FLOW	N	Default	Input	Default	Interval	Default	Default	Default	
FRUITS_VEGGIES	N	Default	Input	Default	Interval	Default	Default	Default	
GENDER	Y	Default	Rejected	Default	Nominal	Default	Default	Default	
IMP_DAILY_STRESS	N	Default	Input	Default	Interval	Default	Default	Default	
LIVE_VISION	N	Default	Input	Default	Interval	Default	Default	Default	
LOST_VACATION	N	Default	Input	Default	Interval	Default	Default	Default	
PERSONAL_AWARDS	N	Default	Input	Default	Interval	Default	Default	Default	
PLACES_VISITED	N	Default	Input	Default	Interval	Default	Default	Default	
P_WORK_LIFE_BALANCE_SCORE	N	Default	Prediction	Default	Interval	Default	Default	Default	
R_WORK_LIFE_BALANCE_SCORE	N	Default	Residual	Default	Interval	Default	Default	Default	
SLEEP_HOURS	N	Default	Input	Default	Interval	Default	Default	Default	
SOCIAL_NETWORK	N	Default	Input	Default	Interval	Default	Default	Default	
SUFFICIENT_INCOME	Y	Default	Rejected	Default	Interval	Default	Default	Default	
SUPPORTING_OTHERS	N	Default	Input	Default	Interval	Default	Default	Default	
TIME_FOR_PASSION	N	Default	Input	Default	Interval	Default	Default	Default	
TODO_COMPLETED	N	Default	Input	Default	Interval	Default	Default	Default	
VAR1	N	Default	Time ID	Default	Interval	Default	Default	Default	
WEEKLY_MEDITATION	N	Default	Input	Default	Interval	Default	Default	Default	
WORK_LIFE_BALANCE_SCORE	N	Default	Target	Default	Interval	Default	Default	Default	HPNNA4
_WARN	N	Default	Assessment	Default	Nominal	Default	Default	Default	
dataobs	N	Default	ID	Default	Interval	Default	Default	Default	

Figure 57: Metadata

Subtree	
Method	N
Number of Leaves	60
Assessment Measure	Average Square Error
Assessment Fraction	0.25

Figure 58: Surrogate Tree

Metadata and surrogate tree nodes are added for the best model which is 5 hidden layers 5 hidden neurons HP neural network. The surrogate tree is set to have 60 leaves and measured by average square error.

8.3 Tan Yong Hong TP063960

Comparison of Regression Models

Stepwise Regression & 2nd Degree Polynomial Stepwise Regression & HP Regression Models - 2nd Degree Polynomial Backward HP Regression & 2nd Degree Polynomial Stepwise HP Regression

拟合统计量

Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Average Squared Error
Y	Reg2	Reg2	2nd polyl stepwise	WORK ...		50.14317	50.45874
	HPReg2	HPReg2	poly stepwise	WORK ...		50.27226	49.94625
	Reg	Reg	Stepwise Regression	WORK ...		50.3678	51.06452
	HPReg	HPReg	poly backward	WORK ...		50.41175	49.88736

Figure 59: Regression Models Comparison

The fit statistics above shows that 2nd-degree polynomial stepwise regression model and 2nd degree polynomial stepwise HP regression model have the best performances out of the four with the 1st and 2nd lowest valid ASE value. But, for the 2nd-degree polynomial stepwise HP regression model, its train ASE is higher than its valid ASE, meaning that it's overfitting, therefore the model with the 3rd lowest valid ASE, which is stepwise regression will be joining 2nd-degree polynomial stepwise regression to be applied for further analysis.

8.3.1 Model 1 - 2nd Degree Polynomial Stepwise Regression

Fit Statistics & ASE Iteration Plot of 2nd Degree Polynomial Stepwise Regression

拟合统计量

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
WORK_LIFE_BALANCE_SCORE		AIC	Akaike's Information Criterion	50162.53	.
WORK_LIFE_BALANCE_SCORE		ASE	Average Squared Error	50.45874	50.14317
WORK_LIFE_BALANCE_SCORE		AVERR	Average Error Function	50.45874	50.14317
WORK_LIFE_BALANCE_SCORE		DFE	Degrees of Freedom for Error	12749	.
WORK_LIFE_BALANCE_SCORE		DFM	Model Degrees of Freedom	29	.
WORK_LIFE_BALANCE_SCORE		DFT	Total Degrees of Freedom	12778	.
WORK_LIFE_BALANCE_SCORE		DIV	Divisor for ASE	12778	3194
WORK_LIFE_BALANCE_SCORE		ERR	Error Function	644761.8	160157.3
WORK_LIFE_BALANCE_SCORE		FPE	Final Prediction Error	50.6883	.
WORK_LIFE_BALANCE_SCORE		MAX	Maximum Absolute Error	18.28283	18.72372
WORK_LIFE_BALANCE_SCORE		MSE	Mean Square Error	50.57352	50.14317
WORK_LIFE_BALANCE_SCORE		NOBS	Sum of Frequencies	12778	3194
WORK_LIFE_BALANCE_SCORE		NW	Number of Estimate Weights	29	.
WORK_LIFE_BALANCE_SCORE		RASE	Root Average Sum of Squares	7.103432	7.081184
WORK_LIFE_BALANCE_SCORE		RFPE	Root Final Prediction Error	7.119571	.
WORK_LIFE_BALANCE_SCORE		RMSE	Root Mean Squared Error	7.111506	7.081184
WORK_LIFE_BALANCE_SCORE		SBC	Schwarz's Bayesian Criterion	50378.74	.
WORK_LIFE_BALANCE_SCORE		SSE	Sum of Squared Errors	644761.8	160157.3
WORK_LIFE_BALANCE_SCORE		SUMW	Sum of Case Weights Times Freq	12778	3194

Figure 60: Fit Statistics of 2nd Degree Polynomial Stepwise Regression

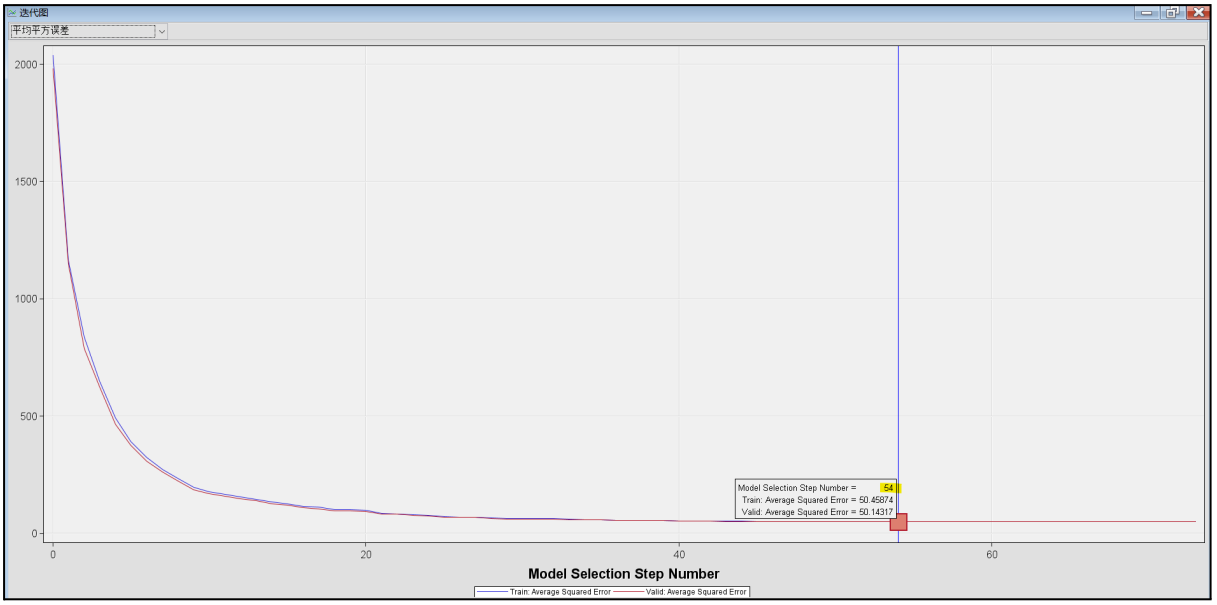


Figure 61: ASE Iteration Plot of 2nd Degree Polynomial Stepwise Regression

According to the fit statistics table above, the 2nd Degree Polynomial Stepwise Regression Model is having a best average squared error of 0.02081, which is calculated by using the square root of 50.1, which is the validation ASE, divided by the range of the target variable, Work-Life Balance Score, which is (820-480). Only till the 54th iteration, the training process was able to yield this rate. This rate indicates that the error rate of the outcomes predicted is 2.08% and according to the ASE graph and the fact that validation ASE is smaller than train ASE, there are no signs of overfitting.

8.3.2 Model 2 - Stepwise Regression

Fit Statistics & ASE Iteration Plot of Stepwise Regression

拟合统计量

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
WORK_LIFE_BA...		AIC	Akaike's Informati...	50303.02	.
WORK_LIFE_BA...		ASE	Average Squared ...	51.06452	50.3678
WORK_LIFE_BA...		AVERR	Average Error Fun...	51.06452	50.3678
WORK_LIFE_BA...		DFE	Degrees of Freed...	12755	.
WORK_LIFE_BA...		DFM	Model Degrees of ...	23	.
WORK_LIFE_BA...		DFT	Total Degrees of ...	12778	.
WORK_LIFE_BA...		DIV	Divisor for ASE	12778	3194
WORK_LIFE_BA...		ERR	Error Function	652502.4	160874.7
WORK_LIFE_BA...		FPE	Final Prediction Er...	51.24868	.
WORK_LIFE_BA...		MAX	Maximum Absolut...	18.33359	18.87151
WORK_LIFE_BA...		MSE	Mean Square Error	51.1566	50.3678
WORK_LIFE_BA...		NOBS	Sum of Frequencies	12778	3194
WORK_LIFE_BA...		NW	Number of Estima...	23	.
WORK_LIFE_BA...		RASE	Root Average Su...	7.145944	7.097027
WORK_LIFE_BA...		RFPE	Root Final Predicti...	7.158818	.
WORK_LIFE_BA...		RMSE	Root Mean Squar...	7.152384	7.097027
WORK_LIFE_BA...		SBC	Schwarz's Bayesi...	50474.5	.
WORK_LIFE_BA...		SSE	Sum of Squared E...	652502.4	160874.7
WORK_LIFE_BA...		SUMW	Sum of Case Wei...	12778	3194

Figure 62: Fit Statistics of stepwise regression

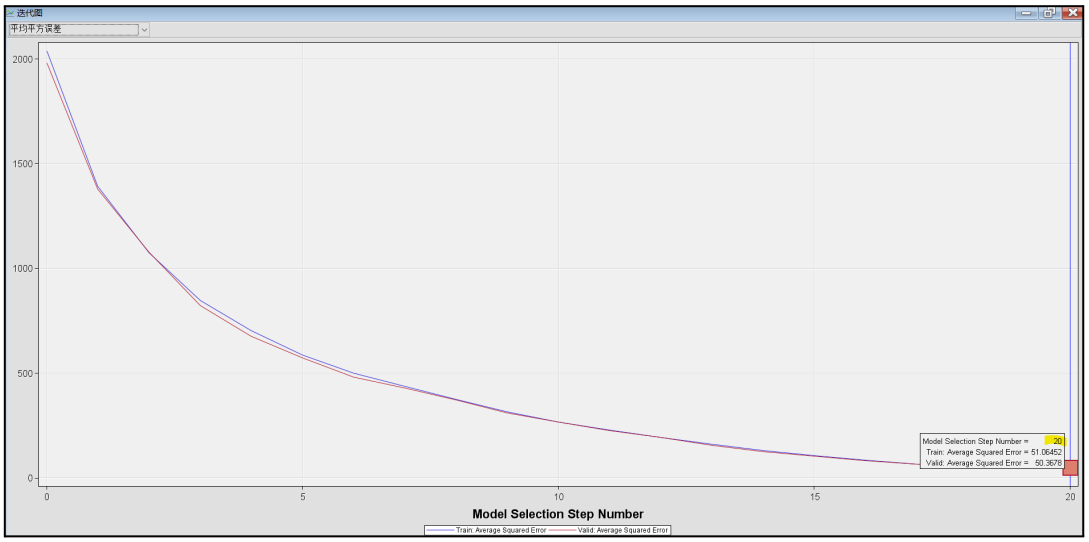


Figure 63: ASE Iteration Plot of stepwise regression

According to the fit statistics table above, the stepwise regression model is having a best average squared error of 0.02087, which is calculated by using the square root of 50.37, which the validation ASE, divided by the range of the target variable, Work-Life Balance Score, which is (820-480). Only till the 20th iteration, the training process was able to yield this rate. This rate indicates that the error rate of the outcomes predicted is 2.087% and according to the ASE graph and the fact that validation ASE is smaller than train ASE, there are no signs of overfitting.

9.0 Critical Interpretation of Outcomes

9.1 Wong Kit Yee TP061080

9.1.1 Model 1 - Regression

Summary of Stepwise Selection

Summary of Stepwise Selection

Step	Entered	Effect Removed	DF	Number In	F Value	Pr > F	Validation Error Rate
1	SUPPORTING_OTHERS*TODO_COMPLETED*WEEKLY_MEDITATION		1	1	10479.4	<.0001	3452281
2	CORE_CIRCLE*DAILY_STEPS*PLACES_VISITED		1	2	5042.07	<.0001	2394926
3	BMI_RANGE*IMP_DAILY_STRESS		1	3	2949.81	<.0001	2008588
4	PERSONAL_AWARDS*SLEEP_HOURS*SOCIAL_NETWORK		1	4	4113.69	<.0001	1501950
5	FRUITS_VEGGIES*LIVE_VISION		1	5	3021.12	<.0001	1217410
6	DONATION		1	6	2428.88	<.0001	1036246
7	TIME_FOR_PASSION		1	7	1806.05	<.0001	896931
8	BMI_RANGE*LOST_VACATION		1	8	2243.30	<.0001	749994
9	ACHIEVEMENT		1	9	1708.60	<.0001	667676
10	BMI_RANGE*DAILY_SHOUTING		1	10	1573.61	<.0001	589746
11	TODO_COMPLETED		1	11	1216.29	<.0001	542080
12	SUPPORTING_OTHERS*WEEKLY_MEDITATION		1	12	1868.56	<.0001	476802
13	FLOW*FRUITS_VEGGIES		1	13	1080.40	<.0001	435044
14	PLACES_VISITED		1	14	828.89	<.0001	408888
15	CORE_CIRCLE*DAILY_STEPS		1	15	2839.92	<.0001	332842
16	BMI_RANGE*IMP_DAILY_STRESS*SOCIAL_NETWORK		1	16	439.22	<.0001	325146
17	PERSONAL_AWARDS*SLEEP_HOURS		1	17	606.02	<.0001	316369
18	SOCIAL_NETWORK		1	18	592.44	<.0001	303362
19	DAILY_STEPS*WEEKLY_MEDITATION		1	19	444.05	<.0001	291861
20	CORE_CIRCLE*SUPPORTING_OTHERS		1	20	510.85	<.0001	279776
21	LIVE_VISION		1	21	439.52	<.0001	268449
22	FRUITS_VEGGIES		1	22	1173.43	<.0001	243343
23	FLOW		1	23	463.42	<.0001	234970
24		FRUITS_VEGGIES*LIVE_VISION	1	22	1.76	0.1843	235160
25	BMI_RANGE*BMI_RANGE*LOST_VACATION		1	23	372.34	<.0001	228721
26	BMI_RANGE*BMI_RANGE*BMI_RANGE		1	24	1039.31	<.0001	210322
27	BMI_RANGE*BMI_RANGE*IMP_DAILY_STRESS		1	25	929.81	<.0001	195881
28	DAILY_SHOUTING		1	26	440.15	<.0001	189325
29		BMI_RANGE*DAILY_SHOUTING	1	25	0.00	0.9793	189325
30	IMP_DAILY_STRESS*SLEEP_HOURS*SUPPORTING_OTHERS		1	26	231.72	<.0001	188084
31	WEEKLY_MEDITATION		1	27	303.36	<.0001	184925
32	DAILY_STEPS		1	28	552.35	<.0001	177047
33	CORE_CIRCLE		1	29	488.18	<.0001	174036
34		CORE_CIRCLE*DAILY_STEPS	1	28	0.04	0.8372	174014
35		BMI_RANGE*IMP_DAILY_STRESS*SOCIAL_NETWORK	1	27	1.28	0.2584	174007
36		FLOW*FRUITS_VEGGIES	1	26	3.42	0.0643	174026
37	SUPPORTING_OTHERS		1	27	517.26	<.0001	168421
38		SUPPORTING_OTHERS*TODO_COMPLETED*WEEKLY_MEDITATION	1	26	0.33	0.5639	168447
39		DAILY_STEPS*WEEKLY_MEDITATION	1	25	1.01	0.3145	168517
40		CORE_CIRCLE*SUPPORTING_OTHERS	1	24	3.33	0.0682	168319
41	AGE		3	25	72.59	<.0001	164835
42	SLEEP_HOURS		1	26	105.12	<.0001	164648
43		IMP_DAILY_STRESS*SLEEP_HOURS*SUPPORTING_OTHERS	1	25	3.63	0.0569	164552
44	PERSONAL_AWARDS*SLEEP_HOURS*SLEEP_HOURS		1	26	293.79	<.0001	161240
45	PLACES_VISITED*PLACES_VISITED*TODO_COMPLETED		1	27	34.49	<.0001	160790
46		CORE_CIRCLE*DAILY_STEPS*PLACES_VISITED	1	26	0.19	0.6599	160766
47		PERSONAL_AWARDS*SLEEP_HOURS*SOCIAL_NETWORK	1	25	2.53	0.1121	160760
48	DAILY_STEPS*DAILY_STEPS		1	26	30.76	<.0001	160528
49	LIVE_VISION*LIVE_VISION		1	27	23.46	<.0001	160719
50		SUPPORTING_OTHERS*WEEKLY_MEDITATION	1	26	3.44	0.0638	160707
51	DONATION*PERSONAL_AWARDS*WEEKLY_MEDITATION		1	27	18.71	<.0001	160899
52	PERSONAL_AWARDS		1	28	17.51	<.0001	160460
53	IMP_DAILY_STRESS*SLEEP_HOURS*SOCIAL_NETWORK		1	29	15.56	<.0001	160251
54	BMI_RANGE*FLOW*TIME_FOR_PASSION		1	30	13.14	0.0003	160188
55	SLEEP_HOURS*SLEEP_HOURS*SLEEP_HOURS		1	31	12.10	0.0005	160034
56		PERSONAL_AWARDS*SLEEP_HOURS	1	30	0.00	0.9753	160035
57		PERSONAL_AWARDS*SLEEP_HOURS*SLEEP_HOURS	1	29	0.37	0.5428	160053
58	DAILY_SHOUTING*DAILY_SHOUTING		1	30	12.82	0.0003	160317
59	SOCIAL_NETWORK*SOCIAL_NETWORK*SOCIAL_NETWORK		1	31	11.54	0.0007	160294
60	DAILY_SHOUTING*DONATION*SLEEP_HOURS		1	32	10.32	0.0013	160656
61	DONATION*DONATION*IMP_DAILY_STRESS		1	33	10.13	0.0015	160766
62	DAILY_STEPS*DAILY_STEPS*TIME_FOR_PASSION		1	34	9.76	0.0018	160846
63	BMI_RANGE*FRUITS_VEGGIES*WEEKLY_MEDITATION		1	35	9.92	0.0016	160989
64	LIVE_VISION*LIVE_VISION*LIVE_VISION		1	36	8.30	0.0040	161089
65	BMI_RANGE*LIVE_VISION*PLACES_VISITED		1	37	8.11	0.0044	161219
66	ACHIEVEMENT*DAILY_STEPS*SUPPORTING_OTHERS		1	38	7.82	0.0052	161468
67	FRUITS_VEGGIES*PLACES_VISITED*TODO_COMPLETED		1	39	7.73	0.0054	161679
68	BMI_RANGE*CORE_CIRCLE*CORE_CIRCLE		1	40	7.10	0.0077	161616
69	ACHIEVEMENT*LIVE_VISION*SUPPORTING_OTHERS		1	41	6.90	0.0086	161814
70	ACHIEVEMENT*FRUITS_VEGGIES*TIME_FOR_PASSION		1	42	7.67	0.0056	162029
71	DAILY_SHOUTING*SOCIAL_NETWORK*WEEKLY_MEDITATION		1	43	6.96	0.0083	162311
72	BMI_RANGE*LOST_VACATION*TIME_FOR_PASSION		1	44	6.10	0.0136	162229
73	LOST_VACATION*PLACES_VISITED*TIME_FOR_PASSION		1	45	6.29	0.0121	162419
74	DAILY_SHOUTING*LIVE_VISION*LIVE_VISION		1	46	5.26	0.0219	162302
75	CORE_CIRCLE*DAILY_SHOUTING*TODO_COMPLETED		1	47	6.75	0.0094	162482
76	FRUITS_VEGGIES*LOST_VACATION*LOST_VACATION		1	48	4.95	0.0261	162628
77	FLOW*FLOW*IMP_DAILY_STRESS		1	49	4.33	0.0374	162641
78	ACHIEVEMENT*FLOW*IMP_DAILY_STRESS		1	50	4.20	0.0405	162625
79	FRUITS_VEGGIES*FRUITS_VEGGIES*SOCIAL_NETWORK		1	51	4.24	0.0395	162793
80	SLEEP_HOURS*SUPPORTING_OTHERS*TODO_COMPLETED		1	52	3.93	0.0475	162870
81	DONATION*SOCIAL_NETWORK*TODO_COMPLETED		1	53	4.66	0.0309	162999
82	CORE_CIRCLE*WEEKLY_MEDITATION*WEEKLY_MEDITATION		1	54	3.98	0.0461	162971
83	DONATION*SOCIAL_NETWORK*WEEKLY_MEDITATION		1	55	4.76	0.0292	163087

The selected model, based on the error rate for the validation data, is the model trained in Step 55. It consists of the following effects:

Figure 64: Summary of Stepwise Selection

A high F-value generally suggests a more favourable fit of the model. From the figure above, when the predictor variables "SUPPORTING_OTHERS*TODO_COMPLETE* WEEKLY_MEDIDATION" entered, they demonstrated the highest degree of contribution to the model, as evidenced by their respective F-values. This means that they have a strong association with the response variable and increase the accuracy of the model's predictions.

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	548.3	1.6965	323.17	<.0001
ACHIEVEMENT		1	1.7042	0.0283	60.12	<.0001
AGE	21 to 35	1	-0.8567	0.1029	-8.33	<.0001
AGE	36 to 50	1	0.6011	0.1103	5.45	<.0001
AGE	51 or more	1	1.6474	0.1289	12.78	<.0001
CORE_CIRCLE		1	1.7378	0.0252	69.03	<.0001
DAILY_SHOUTING		1	-1.7151	0.0250	-68.61	<.0001
DAILY_STEPS		1	2.4513	0.0998	24.56	<.0001
DONATION		1	3.6348	0.0531	68.49	<.0001
FLOW		1	1.5344	0.0426	36.03	<.0001
FRUITS_VEGGIES		1	3.5925	0.0482	74.54	<.0001
LIVE_VISION		1	2.1202	0.0708	29.93	<.0001
PERSONAL_AWARDS		1	1.9552	0.3583	5.46	<.0001
PLACES_VISITED		1	2.0755	0.0418	49.67	<.0001
SLEEP_HOURS		1	3.1177	0.3475	8.97	<.0001
SOCIAL_NETWORK		1	1.7235	0.0442	39.03	<.0001
SUPPORTING_OTHERS		1	1.6542	0.0240	68.81	<.0001
TIME_FOR_PASSION		1	1.5192	0.0383	39.69	<.0001
TODO_COMPLETED		1	2.0793	0.0333	62.50	<.0001
WEEKLY_MEDITATION		1	1.7363	0.0277	62.71	<.0001
BMI_RANGE*IMP_DAILY_STRESS		1	-6.5198	0.1837	-35.50	<.0001
BMI_RANGE*LOST_VACATION		1	-2.5764	0.0477	-53.98	<.0001
DAILY_STEPS*DAILY_STEPS		1	-0.0424	0.00836	-5.07	<.0001
LIVE_VISION*LIVE_VISION		1	-0.0323	0.00664	-4.87	<.0001
PERSONAL_AWARDS*SLEEP_HOURS		1	-0.00324	0.1043	-0.03	0.9753
BMI_RANGE*BMI_RANGE*BMI_RANGE		1	-2.5103	0.0451	-55.67	<.0001
BMI_RANGE*BMI_RANGE*IMP_DAILY_STRESS		1	2.2127	0.0842	26.28	<.0001
BMI_RANGE*BMI_RANGE*LOST_VACATION		1	0.8424	0.0264	31.90	<.0001
BMI_RANGE*FLOW*TIME_FOR_PASSION		1	0.0172	0.00473	3.64	0.0003
DONATION*PERSONAL_AWARDS*WEEKLY_MEDITATION		1	-0.00447	0.000991	-4.51	<.0001
IMP_DAILY_STRESS*SLEEP_HOURS*SOCIAL_NETWORK		1	0.00742	0.00190	3.91	<.0001
PERSONAL_AWARDS*SLEEP_HOURS*SLEEP_HOURS		1	-0.00050	0.00756	-0.07	0.9478
PLACES_VISITED*PLACES_VISITED*TODO_COMPLETED		1	-0.00265	0.000540	-4.90	<.0001
SLEEP_HOURS*SLEEP_HOURS*SLEEP_HOURS		1	-0.00799	0.00230	-3.48	0.0005

Figure 65: Analysis of Maximum Likelihood Estimates @ step 55

The intercept is a regression model parameter that signifies the dependent variable value when all the independent predictor variables in the model are equal to zero. Essentially, the intercept establishes the initial value of the dependent variable and facilitates comparison between various predictor variable categories or groups.

The estimate value is the most likely value of the parameter given the available data and the assumptions of the model. At step 55, the intercept parameter has an estimated value of 548.3. Besides, a smaller standard error generally implies a more precise and dependable estimate. Parameter "PLACES_VISITED*PLACED_VISITED*TODO_COMPLETE" is having the least standard error which is only 0.000540.

Type 3 Analysis of Effects

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
ACHIEVEMENT	1	182201.016	3613.87	<.0001
AGE	3	12314.1752	81.42	<.0001
CORE_CIRCLE	1	240267.918	4765.60	<.0001
DAILY_SHOUTING	1	237361.101	4707.94	<.0001
DAILY_STEPS	1	30400.9563	602.99	<.0001
DONATION	1	236528.565	4691.43	<.0001
FLOW	1	65467.2976	1298.51	<.0001
FRUITS_VEGGIES	1	280135.368	5556.35	<.0001
LIVE_VISION	1	45166.2440	895.85	<.0001
PERSONAL_AWARDS	1	1501.6158	29.78	<.0001
PLACES_VISITED	1	124375.017	2466.92	<.0001
SLEEP_HOURS	1	4057.5226	80.48	<.0001
SOCIAL_NETWORK	1	76791.2727	1523.12	<.0001
SUPPORTING_OTHERS	1	238741.835	4735.33	<.0001
TIME_FOR_PASSION	1	79421.8888	1575.30	<.0001
TODO_COMPLETED	1	196925.652	3905.93	<.0001
WEEKLY_MEDITATION	1	198267.878	3932.55	<.0001
BMI_RANGE*IMP_DAILY_STRESS	1	63536.0360	1260.21	<.0001
BMI_RANGE*LOST_VACATION	1	146913.060	2913.95	<.0001
DAILY_STEPS*DAILY_STEPS	1	1297.9103	25.74	<.0001
LIVE_VISION*LIVE_VISION	1	1194.5416	23.69	<.0001
PERSONAL_AWARDS*SLEEP_HOURS	1	0.0485	0.00	0.9753
BMI_RANGE*BMI_RANGE*BMI_RANGE	1	156256.830	3099.28	<.0001
BMI_RANGE*BMI_RANGE*IMP_DAILY_STRESS	1	34828.7105	690.81	<.0001
BMI_RANGE*BMI_RANGE*LOST_VACATION	1	51308.5261	1017.68	<.0001
BMI_RANGE*FLOW*TIME_FOR_PASSION	1	668.2615	13.25	0.0003
DONATION*PERSONAL_AWARDS*WEEKLY_MEDITATION	1	1023.9555	20.31	<.0001
IMP_DAILY_STRESS*SLEEP_HOURS*SOCIAL_NETWORK	1	769.0147	15.25	<.0001
PERSONAL_AWARDS*SLEEP_HOURS*SLEEP_HOURS	1	0.2162	0.00	0.9478
PLACES_VISITED*PLACES_VISITED*TODO_COMPLETED	1	1208.5284	23.97	<.0001
SLEEP_HOURS*SLEEP_HOURS*SLEEP_HOURS	1	609.9480	12.10	0.0005

Figure 66: Type 3 Analysis of Effects

The influence of each predictor variable is examined in type 3 analysis while adjusting for the impacts of all other predictor variables in the model. The "Pr > F" value is used to assess the significance of each predictor variable's effect on the response variable. “PERSONAL_AWARDS*SLEEP_HOURS” effect is producing the highest “Pr > F” value. Apart from that, most of the predictor variables have "Pr > F" values less than 0.05.

Analysis of Variance

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	33	25407088	769912	15270.8	<.0001
Error	12744	642516	50.417147		
Corrected Total	12777	26049604			

Figure 67: Analysis of Variance

In Analysis of Variance, "error" refers to the difference between the observed value of the dependent variable and the value predicted by the model. On the other hand, "corrected total" refers to the overall variance in the dependent variable.

Model Fit Statistics

Model Fit Statistics

R-Square	0.9753	Adj R-Sq	0.9753
AIC	50127.9484	BIC	50128.5844
SBC	50381.4347	C(p)	329.4001

Figure 68: Model Fit Statistics

R-square, which is also known as coefficient, is a metric that indicates how accurately the model fits the data. An R-squared value of 0.9753 shows that the model explains a large proportion (97.53%) of the total variation in the dependent variable, indicating a good fit between the model and the data.

9.1.2 Model 2 - Neural Network

Surrogate Tree

Based on the figure below which indicates details of node 127, it is showing the ***highest work life balance score*** predicted by the 3rd Degree Polynomial Stepwise Regression Model with a total number of 109 observations.

```
*-----------------------------------------------------------*
 Node = 127
*-----------------------------------------------------------*
if WEEKLY_MEDITATION >= 7.5 or MISSING
AND SUPPORTING_OTHERS >= 4.5 or MISSING
AND PLACES_VISITED >= 4.5 or MISSING
AND LIVE_VISION >= 9.5
AND FRUITS_VEGGIES >= 4.5
AND ACHIEVEMENT >= 6.5
then
 Tree Node Identifier   = 127
 Number of Observations = 109
 Predicted: WORK_LIFE_BALANCE_SCORE = 764.82385321
```

Figure 69: Surrogate Tree Node Rules with the highest value of work life balance score

It is shown that:

An individual who engages in self-reflection for more than 7 times in a week,
AND has assisted 4 or more people in achieving a better life,
AND has visited more than 4 new places,
AND has maintained a clear life vision for at least 9 years,
AND consumes more than 4 servings of fruits or vegetables on a daily basis,
AND takes pride in more than 6 significant achievements,
is more probable to achieve a work-life balance score greater than 764.

```
*-------------------------------------------------------------*
 Node = 118
*-------------------------------------------------------------*
if TODO_COMPLETED >= 6.5 or MISSING
AND TIME_FOR_PASSION >= 2.5 or MISSING
AND SUPPORTING_OTHERS >= 4.5 or MISSING
AND PLACES_VISITED >= 4.5 or MISSING
AND BMI_RANGE < 1.5 or MISSING
AND ACHIEVEMENT < 6.5 or MISSING
then
 Tree Node Identifier   = 118
 Number of Observations = 674
 Predicted: WORK_LIFE_BALANCE_SCORE = 715.76854599
```

Figure 70: Surrogate Tree Node Rules with the highest number of observations

Based on the figure above which indicates details of node 118, it is showing the work life balance score predicted with the ***highest number of observations*** by the 3rd Degree Polynomial Stepwise Regression Model.

<u>It is shown that:</u>

An individual who has the ability to complete 6 or more to-do lists each week,
AND devotes at least 2 hours daily to their passion,
AND has assisted 4 or more people in achieving a better life,
AND has visited at least 4 new places,
AND has a BMI range below 1.5,
AND takes pride in more than 6 significant achievements,
is having a higher likelihood of achieving a work-life balance score greater than 715.

9.2 Chee Wei Keng TP063861

9.2.1 Model 1 - Clustering

Mean Statistics

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id ▲	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	ACHIEVEMENT	CORE_CIRCLE	DAILY_SHOUTING	DAILY_STEPS	DONATION	FLOW	FRUITS_VEGGIES	Imputed DAILY_STRESS
0.81126	0.006395	.	1	2572	0.809801	7.095135	2	2.783551	1.800156	3.276439	3.584759	3.998056	1.4514	1.532271	1.999611	3.290047
0.81126	0.006395	.	2	1814	0.800315	7.387537	3	2.309305	3.881477	5.80871	2.261852	5.539691	1.49559	3.871555	2.276736	2.548512
0.81126	0.006395	.	3	1772	0.811787	6.853819	4	2.137352	4.26298	5.68228	1.883183	4.170429	3.656321	3.35553	2.803047	1.9193
0.81126	0.006395	.	4	2170	0.781271	6.171754	3	2.137352	3.065438	5.282028	2.395853	7.114286	2.567281	2.206452	3.795392	2.462212
0.81126	0.006395	.	5	2439	0.810322	6.581344	3	2.746922	6.905289	7.555966	2.448954	7.373514	3.778188	5.470275	3.700287	2.401394
0.81126	0.006395	.	6	2011	0.855218	6.615432	2	2.364873	4.168573	5.599204	4.810542	5.855296	3.378419	2.818498	2.913476	3.952159

	LIVE_VISION	LOST_VACATION	PERSONAL_AWARDS	PLACES_VISITED	SLEEP_HOURS	SOCIAL_NETWORK	SUPPORTING_OTHERS	TIME_FOR_PASSION	TODO_COMPLETED	WEEKLY_MEDITATION	AGE=21 to 35	AGE=36 to 50	AGE=51 or more	AGE=Less than 20
7	1.68196	3.593701	3.420295	2.805988	6.7725...	4.106532	2.738725	1.464619	3.323872	4.746112	0.4506...	0.2741...	0.1496...	0.1255...
2	4.439361	3.019294	4.648842	3.951488	6.9812...	7.962514	4.363286	3.292723	6.201213	4.939912	0.44763	0.2320...	0.1003...	0.2199...
3	4.718962	1.932844	6.441874	5.178894	7.3521...	4.698646	6.607223	4.039503	6.432844	7.641084	0.3504...	0.2381...	0.3544...	0.0569...
2	2.523041	1.337788	5.206452	7.257143	7.4603...	5.896774	4.580645	2.252535	6.051152	7.440092	0.4267...	0.2686...	0.2211...	0.08341
4	6.345223	2.448954	8.147191	7.261993	7.2103...	8.533825	8.364904	6.108241	7.556786	7.726527	0.2997...	0.3165...	0.2796...	0.1041...
9	3.16907	5.239682	6.639483	4.805569	6.4937...	7.829438	7.210343	2.790651	5.306813	4.91099	0.3102...	0.4157...	0.17454	0.0994...

Figure 71: Mean Statistics of Cluster Model

Based upon the Figure above represents the average values of the variables being studied in the cluster model. This information can be used to gain insights into the characteristics of the clusters formed and the distribution of the variables within each cluster.

<u>Segment 1</u>

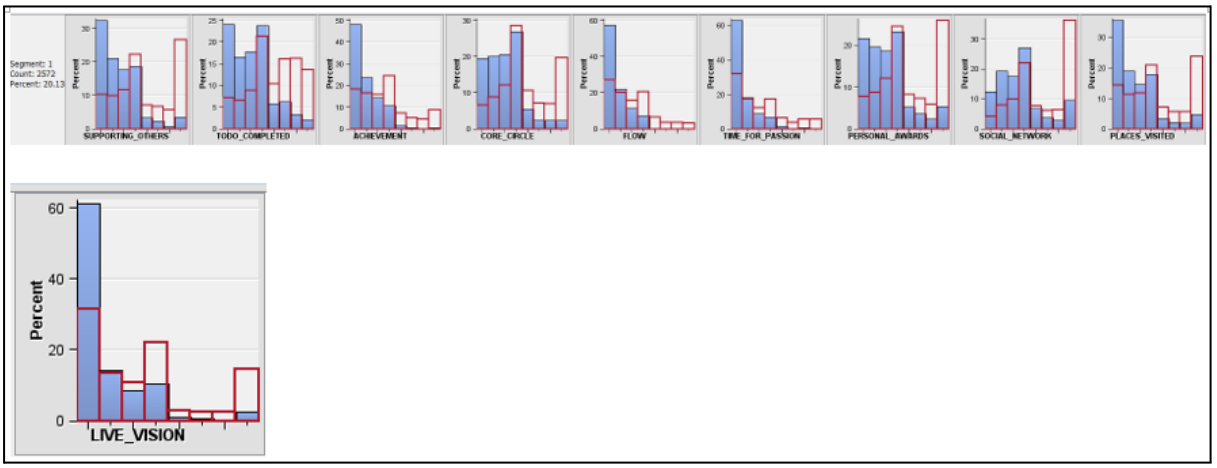


Figure 72: Segment 1

Segment 1 consists of the 2572 number of frequencies in the cluster, reaching 20.13%. The most influential variable in segment 1 is Supporting_Others. From the observation, it can be

determined Supporting_others, Todo_completed, Achievement, Core_circle, Flow, Time_for_Passion variables are slightly lower than average.

Segment 2

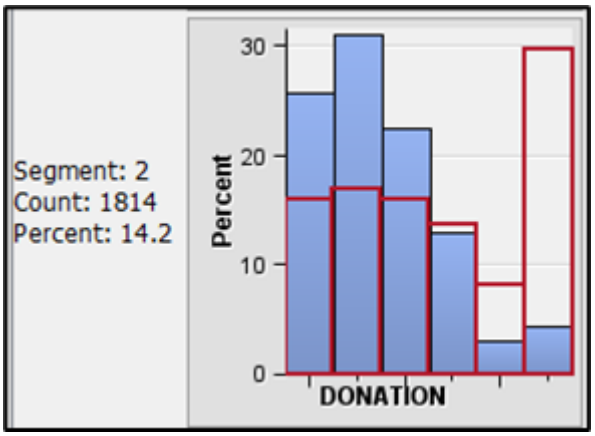


Figure 73: Segment 2

Segment 2 contains 1814 numbers of the frequency in the cluster, accounting for 14.2%. Segment 2 reveals a slightly lower number of people who have donation habits.

Segment 3

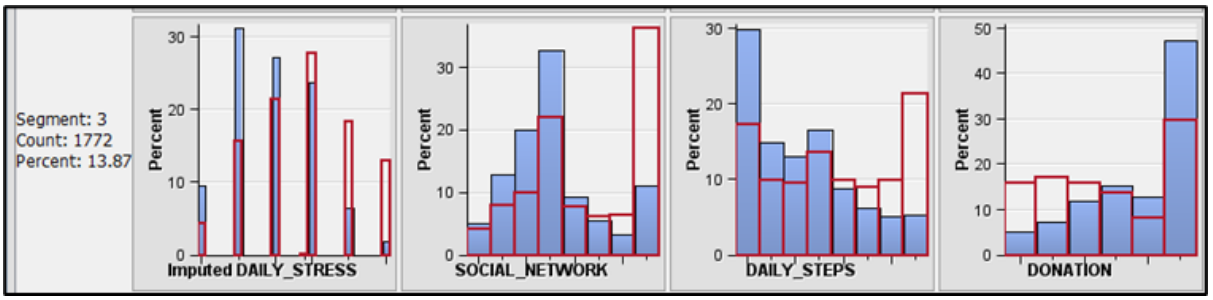


Figure 74: Segment 3

Segment 3 contains 1772 numbers of the frequency in the cluster, accounting for 13.87%. Segment 3 is like an average segment. All the variables are having an average number of people. The most influential variable in this segment is Daily_Stress. In this order, with a slight elevated number of people shown for the Donation variable. All the others are exactly overlapping averages, indicating there is no unique point at which to say this is low or high.

Segment 4

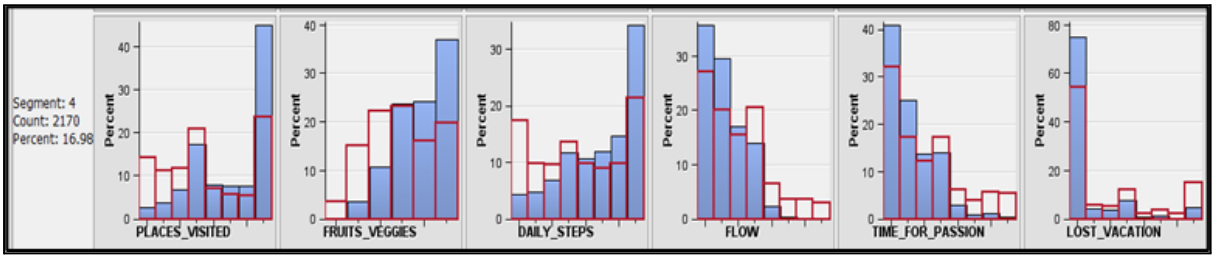


Figure 75: Segment 4

Segment 4 contains 2170 numbers of the frequency in the cluster, accounting for 16.98%. Segment 4 contains a slightly elevated number of people who prefer fruits and vegetables. The most influential variable in this segment is Places_Visited. On the other hand, Flow, Time_For_Passion option reveals a slightly lower number of people than the average. Places_Visited are performed similar to the average numbers.

Segment 5

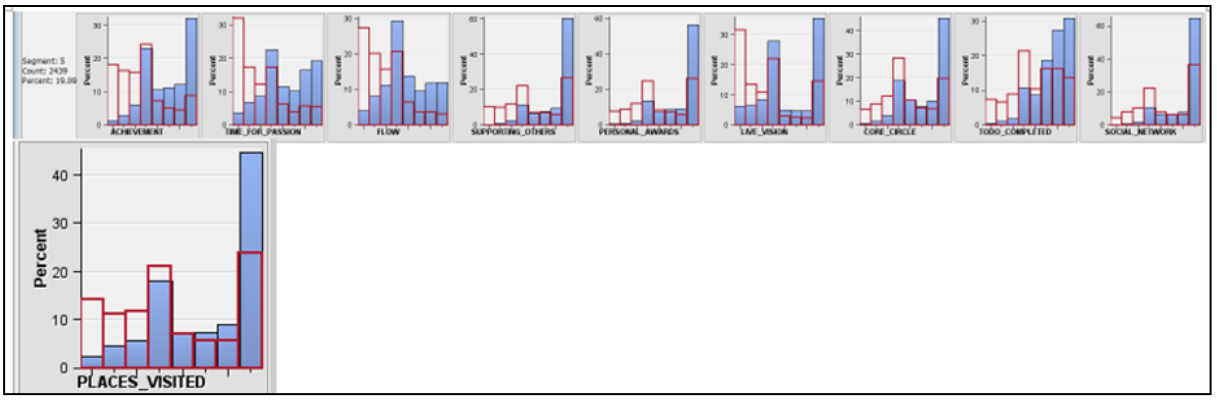


Figure 76: Segment 5

Segment 5 consists of the 2439 number of frequencies in the cluster, reaching 19.09%. The most influential variable in segment 5 is Achievement. The variables are Achievement, Time_For_Passion, Flow, Supporting_Other, Live_Vision, Core_Circle, Todo_Completed, Social_Network and Places_Visited in this segment. From the observation, it can be determined that all the variables are slightly higher than average.

Segment 6

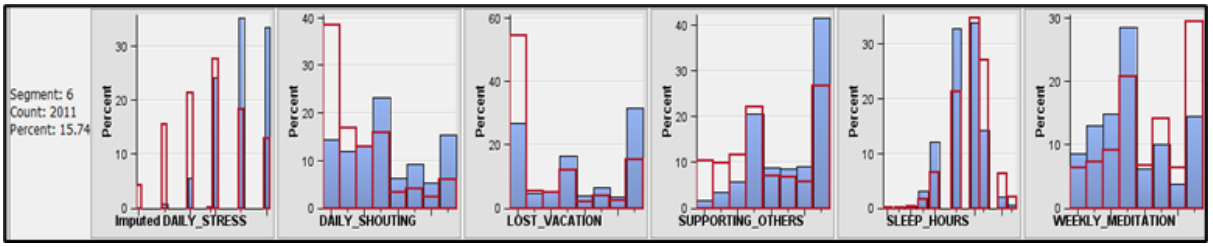


Figure 77: Segment 6

Segment 6 contains the 2011 number of the frequency in the cluster, accounting for 15.74%. Most influential variable in segment 6 is Daily_Stress. It shows a very high number of people that is more than the overall averages. Individuals who are willing to Supporting_Other have a highest average with 7.2103.

9.2.2 Model 2 - Neural Network (5L 5N HPNN)

```
*-------------------------------------------------------------*
 Node = 127
*-------------------------------------------------------------*
if WEEKLY_MEDITATION >= 7.5 or MISSING
AND SUPPORTING_OTHERS >= 4.5 or MISSING
AND PLACES_VISITED >= 4.5 or MISSING
AND LIVE_VISION >= 9.5
AND FRUITS_VEGGIES >= 4.5
AND ACHIEVEMENT >= 6.5
then
 Tree Node Identifier   = 127
 Number of Observations = 109
 Predicted: WORK_LIFE_BALANCE_SCORE = 764.82385321
```

Figure 78: Surrogate Tree Node Rules Observation for highest score

Based on the observation in the above figure, 764.8239 is the highest score predicted by the HP neural network model with 5 hidden layers and 5 hidden neurons for an individual's work life balance. There are 109 individuals with this score.

These people are characterised by practising meditation more than 7 times a week, supporting more than 4 people in their lives, visiting more than 4 destinations, having a clear vision of life for more than 9 years, eating healthy food more than 4 times a day, and achieving more than 6 accomplishments in their lives.

```
*-------------------------------------------------------------*
 Node = 68
*-------------------------------------------------------------*
if WEEKLY_MEDITATION < 0.5
AND TODO_COMPLETED < 4.5
AND SUPPORTING_OTHERS < 4.5
AND PLACES_VISITED < 2.5
AND Imputed DAILY_STRESS >= 4.5
AND CORE_CIRCLE < 2.5
then
 Tree Node Identifier   = 68
 Number of Observations = 23
 Predicted: WORK_LIFE_BALANCE_SCORE = 518.15652174
```

Figure 79: Surrogate Tree Node Rules Observation for lowest scores

Based on the observation in the above figure, 518.1565 is the lowest score predicted by the HP neural network model with 5 hidden layers and 5 hidden neurons for an individual's work life balance. There are 23 individuals with this score.

These people are characterised by meditating less than 0.5 times per week, completing less than 4 to-do lists, supporting less than 4 people in their lives, visiting less than 2 destinations, having more than 4 stressful times per day, and having less than 2 close people. Not having a clear vision, eating unhealthily, and not accomplishing anything in life.

9.3 Tan Yong Hong TP063960

9.3.1 Model 1 - 2nd Degree Polynomial Stepwise Regression

Summary of Regression Selection

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	F Value	Pr > F	Validation Error Rate
1	SUPPORTING_OTHERS*TODO_COMPLETED		1	1	9616.39	<.0001	3668132
2	DAILY_STEPS*PLACES_VISITED		1	2	5005.86	<.0001	2515115
3	BMI_RANGE*IMP_DAILY_STRESS		1	3	3694.25	<.0001	1995897
4	CORE_CIRCLE*PERSONAL_AWARDS		1	4	4009.44	<.0001	1482842
5	FRUITS_VEGGIES*LIVE_VISION		1	5	3334.74	<.0001	1199865
6	SOCIAL_NETWORK*WEEKLY_MEDITATION		1	6	2631.01	<.0001	980005
7	DONATION*SLEEP_HOURS		1	7	2432.79	<.0001	833989
8	DAILY_SHOUTING*LOST_VACATION		1	8	1991.07	<.0001	717770
9	ACHIEVEMENT*TIME_FOR_PASSION		1	9	2518.82	<.0001	594848
10	FLOW*IMP_DAILY_STRESS		1	10	1305.59	<.0001	537499
11	FRUITS_VEGGIES*SLEEP_HOURS		1	11	871.73	<.0001	509345
12	LIVE_VISION		1	12	973.85	<.0001	475295
13	LOST_VACATION*SLEEP_HOURS		1	13	848.85	<.0001	445175
14	BMI_RANGE*DAILY_SHOUTING		1	14	1171.59	<.0001	402225
15	ACHIEVEMENT		1	15	610.91	<.0001	386007
16	TIME_FOR_PASSION		1	16	1181.47	<.0001	352005
17	TODO_COMPLETED		1	17	517.65	<.0001	336518
18	SUPPORTING_OTHERS		1	18	1324.15	<.0001	305936
19		FRUITS_VEGGIES*LIVE_VISION	1	17	2.09	0.1488	305937
20	DAILY_STEPS		1	18	437.97	<.0001	295701
21	PLACES_VISITED		1	19	1698.51	<.0001	267598
22	BMI_RANGE*TODO_COMPLETED		1	20	422.40	<.0001	260829
23	FLOW		1	21	744.78	<.0001	246831
24	PERSONAL_AWARDS		1	22	308.18	<.0001	240748
25	CORE_CIRCLE		1	23	1228.11	<.0001	220544
26	DAILY_SHOUTING		1	24	237.71	<.0001	216670
27		FLOW*IMP_DAILY_STRESS	1	23	0.00	0.9506	216669
28	BMI_RANGE*BMI_RANGE		1	24	540.78	<.0001	206389
29	IMP_DAILY_STRESS		1	25	547.28	<.0001	196518
30		BMI_RANGE*IMP_DAILY_STRESS	1	24	0.04	0.8373	196508
31		BMI_RANGE*TODO_COMPLETED	1	23	0.19	0.6640	196536
32		BMI_RANGE*DAILY_SHOUTING	1	22	0.95	0.3306	196538
33	IMP_DAILY_STRESS*SOCIAL_NETWORK		1	23	408.84	<.0001	190281
34	WEEKLY_MEDITATION		1	24	634.28	<.0001	183249
35		SUPPORTING_OTHERS*TODO_COMPLETED	1	23	1.87	0.1715	183383
36	SOCIAL_NETWORK		1	24	536.96	<.0001	176216
37		SOCIAL_NETWORK*WEEKLY_MEDITATION	1	23	1.92	0.1659	176159
38	LOST_VACATION		1	24	299.30	<.0001	174496
39		DAILY_SHOUTING*LOST_VACATION	1	23	0.56	0.4537	174559
40	DONATION		1	24	488.86	<.0001	167540
41		ACHIEVEMENT*TIME_FOR_PASSION	1	23	3.31	0.0691	167322
42		DONATION*SLEEP_HOURS	1	22	3.67	0.0555	167344
43	AGE		3	23	75.15	<.0001	163745
44	BMI_RANGE*SLEEP_HOURS		1	24	78.08	<.0001	163608
45	FRUITS_VEGGIES		1	25	141.10	<.0001	161407
46	SLEEP_HOURS		1	26	103.77	<.0001	160917
47		LOST_VACATION*SLEEP_HOURS	1	25	0.20	0.6578	160938
48		BMI_RANGE*SLEEP_HOURS	1	24	0.33	0.5649	160956
49		FRUITS_VEGGIES*SLEEP_HOURS	1	23	2.13	0.1443	160966
50	SLEEP_HOURS*SLEEP_HOURS		1	24	43.28	<.0001	160736
51	DAILY_STEPS*DAILY_STEPS		1	25	30.47	<.0001	160389
52		DAILY_STEPS*PLACES_VISITED	1	24	2.09	0.1480	160254
53	LIVE_VISION*LIVE_VISION		1	25	24.20	<.0001	160389
54	PLACES_VISITED*PLACES_VISITED		1	26	19.60	<.0001	160157
55	DAILY_SHOUTING*DAILY_SHOUTING		1	27	13.62	0.0002	160410
56	SOCIAL_NETWORK*SOCIAL_NETWORK		1	28	10.15	0.0014	160384
57	FRUITS_VEGGIES*WEEKLY_MEDITATION		1	29	10.02	0.0015	160367
58	DAILY_STEPS*TIME_FOR_PASSION		1	30	9.43	0.0021	160392
59	PLACES_VISITED*TODO_COMPLETED		1	31	8.45	0.0037	160307
60	SLEEP_HOURS*SOCIAL_NETWORK		1	32	7.86	0.0051	160275
61	DONATION*IMP_DAILY_STRESS		1	33	7.62	0.0058	160190
62	DAILY_SHOUTING*DONATION		1	34	11.41	0.0007	160510
63	FLOW*TIME_FOR_PASSION		1	35	6.00	0.0144	160373
64	PERSONAL_AWARDS*WEEKLY_MEDITATION		1	36	5.90	0.0152	160490
65	LIVE_VISION*PERSONAL_AWARDS		1	37	6.47	0.0110	160528
66	DAILY_SHOUTING*LIVE_VISION		1	38	5.85	0.0156	160441
67	DONATION*PERSONAL_AWARDS		1	39	5.65	0.0174	160491
68	DAILY_STEPS*SUPPORTING_OTHERS		1	40	4.84	0.0278	160679
69	LOST_VACATION*TIME_FOR_PASSION		1	41	4.65	0.0311	160553
70	CORE_CIRCLE*CORE_CIRCLE		1	42	4.84	0.0277	160555
71		CORE_CIRCLE*PERSONAL_AWARDS	1	41	2.51	0.1129	160563
72	ACHIEVEMENT*PERSONAL_AWARDS		1	42	5.04	0.0248	160532
73	FRUITS_VEGGIES*PLACES_VISITED		1	43	4.83	0.0279	160570

Figure 80: Summary of Regression Selection

The stepwise selection was performed on the data to evaluate the importance of the model and influence of individual predictor variables. F-values were calculated, and the model with "Supporting Others*ToDo Completed" had the highest F-value, indicating a strong association with the response variable and increasing the accuracy of the model's predictions. Other predictor variables with high F-values that contributed to the model were "Daily Steps*Places Visited," "BMI RangeImp* Daily Stress," "Core Circle*Personal Awards," "Fruits_Veggies*Live Vision," and "Social Network*Weekly Meditation."

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	549.1	1.2855	427.14	<.0001
ACHIEVEMENT		1	1.6963	0.0284	59.77	<.0001
AGE	21 to 35	1	-0.8320	0.1029	-8.09	<.0001
AGE	36 to 50	1	0.6197	0.1104	5.61	<.0001
AGE	51 or more	1	1.6379	0.1289	12.71	<.0001
CORE_CIRCLE		1	1.8482	0.0487	37.96	<.0001
DAILY_SHOUTING		1	-1.7111	0.0250	-68.38	<.0001
DAILY_STEPS		1	2.4631	0.0999	24.65	<.0001
DONATION		1	3.4706	0.0385	90.11	<.0001
FLOW		1	1.6307	0.0325	50.12	<.0001
FRUITS_VEGGIES		1	3.5914	0.0483	74.40	<.0001
IMP_DAILY_STRESS		1	-4.2103	0.1050	-40.10	<.0001
LIVE_VISION		1	2.1073	0.0707	29.81	<.0001
LOST_VACATION		1	-1.7527	0.0175	-100.00	<.0001
PERSONAL_AWARDS		1	1.9338	0.0453	42.65	<.0001
PLACES_VISITED		1	2.2422	0.0803	27.93	<.0001
SLEEP_HOURS		1	4.1058	0.3542	11.59	<.0001
SOCIAL_NETWORK		1	1.7384	0.0474	36.70	<.0001
SUPPORTING_OTHERS		1	1.6529	0.0241	68.71	<.0001
TIME_FOR_PASSION		1	1.6107	0.0283	57.01	<.0001
TODO_COMPLETED		1	1.9881	0.0275	72.23	<.0001
WEEKLY_MEDITATION		1	1.6678	0.0229	72.89	<.0001
BMI_RANGE*BMI_RANGE		1	-5.7203	0.0448	-127.79	<.0001
CORE_CIRCLE*PERSONAL_AWARDS		1	-0.0190	0.00711	-2.67	0.0076
DAILY_STEPS*DAILY_STEPS		1	-0.0436	0.00837	-5.21	<.0001
IMP_DAILY_STRESS*SOCIAL_NETWORK		1	0.0456	0.0145	3.15	0.0016
LIVE_VISION*LIVE_VISION		1	-0.0310	0.00663	-4.68	<.0001
PLACES_VISITED*PLACES_VISITED		1	-0.0306	0.00691	-4.43	<.0001
SLEEP_HOURS*SLEEP_HOURS		1	-0.1515	0.0256	-5.92	<.0001

Figure 81: Analysis of Maximum Likelihood Estimates

Every predictor variable in the set of data had comparatively small standard errors, indicating that their coefficient estimates were accurate and reliable. In addition, the high t-values for each predictor variable implied that their effects on the dependent variable were both substantial and accurate. The predictor variables all had low p-values, indicating that they had a significant influence on the dependent variable. The 22-degree-of-freedom model revealed that each predictor variable had a significant impact on the dependent variable.

Analysis of Effects

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
ACHIEVEMENT	1	180671.989	3572.46	<.0001
AGE	3	12275.6084	80.91	<.0001
CORE_CIRCLE	1	72856.3754	1440.60	<.0001
DAILY_SHOUTING	1	236486.514	4676.09	<.0001
DAILY_STEPS	1	30733.3657	607.70	<.0001
DONATION	1	410644.933	8119.76	<.0001
FLOW	1	127038.057	2511.95	<.0001
FRUITS_VEGGIES	1	279956.304	5535.63	<.0001
IMP_DAILY_STRESS	1	81325.1190	1608.06	<.0001
LIVE_VISION	1	44955.6631	888.92	<.0001
LOST_VACATION	1	505777.999	10000.8	<.0001
PERSONAL_AWARDS	1	91973.9951	1818.62	<.0001
PLACES_VISITED	1	39450.8932	780.07	<.0001
SLEEP_HOURS	1	6794.8520	134.36	<.0001
SOCIAL_NETWORK	1	68111.4714	1346.78	<.0001
SUPPORTING_OTHERS	1	238779.101	4721.43	<.0001
TIME_FOR_PASSION	1	164366.398	3250.05	<.0001
TODO_COMPLETED	1	263829.427	5216.75	<.0001
WEEKLY_MEDITATION	1	268706.053	5313.18	<.0001
BMI_RANGE*BMI_RANGE	1	825879.390	16330.3	<.0001
CORE_CIRCLE*PERSONAL_AWARDS	1	360.2172	7.12	0.0076
DAILY_STEPS*DAILY_STEPS	1	1373.4095	27.16	<.0001
IMP_DAILY_STRESS*SOCIAL_NETWORK	1	502.0147	9.93	0.0016
LIVE_VISION*LIVE_VISION	1	1105.5488	21.86	<.0001
PLACES_VISITED*PLACES_VISITED	1	991.0842	19.60	<.0001
SLEEP_HOURS*SLEEP_HOURS	1	1769.9661	35.00	<.0001

Figure 82:Analysis of Effects

The findings showed that most predictor variables had "Pr > F" values less than 0.05, indicating their statistical significance. The predictor variable with the highest F-value was "LOST VACATION," with a value of 10000.8, suggesting a massive effect on the response variable. Other predictors with high F-values were "DONATION," "FRUITS VEGGIES," and "WEEKLY MEDITATION," all of which had significant effects on the response variable. Additionally, all predictor variables had p-values less than 0.0001, indicating a major impact on the dependent variable. Lastly, the presence of interaction terms (e.g., BMI_RANGEBMI_RANGE) and "CORE_CIRCLEPERSONAL_AWARDS" with a p-value of 0.0076 indicated a significant interaction effect on the response variable.

Analysis of Variance & Model Fit Statistics

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	28	25404843	907316	17940.5	<.0001
Error	12749	644762	50.573519		
Corrected Total	12777	26049604			

Model Fit Statistics

R-Square	0.9752	Adj R-Sq	0.9752
AIC	50162.5313	BIC	50164.2212
SBC	50378.7403	C(p)	126.6827

Figure 83: Analysis of Variance & Model Fit Statistics

The high R-square value of 0.9752 indicated a high fit, with a considerable portion of the variance in the dependent variable explained. The low AIC value indicated a good fit for the data.

9.3.2 Model 2 - Stepwise Regression

Summary of Regression Selection

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F	Validation Error Rate
1	ACHIEVEMENT	1	1	5950.55	<.0001	4402500
2	TODO_COMPLETED	1	2	3762.51	<.0001	3444640
3	PLACES_VISITED	1	3	3436.14	<.0001	2631470
4	SUPPORTING_OTHERS	1	4	2608.40	<.0001	2160953
5	IMP_DAILY_STRESS	1	5	2518.78	<.0001	1829059
6	DAILY_STEPS	1	6	2195.45	<.0001	1534481
7	LIVE_VISION	1	7	1848.72	<.0001	1368795
8	BMI_RANGE	1	8	2110.43	<.0001	1191547
9	DONATION	1	9	2418.16	<.0001	995624
10	LOST_VACATION	1	10	2308.82	<.0001	849658
11	TIME_FOR_PASSION	1	11	2165.49	<.0001	725276
12	SOCIAL_NETWORK	1	12	2276.36	<.0001	616432
13	FRUITS_VEGGIES	1	13	2665.60	<.0001	499710
14	PERSONAL_AWARDS	1	14	2764.83	<.0001	406555
15	WEEKLY_MEDITATION	1	15	2882.51	<.0001	335684
16	CORE_CIRCLE	1	16	3078.70	<.0001	267945
17	DAILY_SHOUTING	1	17	3796.49	<.0001	209025
18	FLOW	1	18	2145.36	<.0001	175250
19	SLEEP_HOURS	1	19	1276.02	<.0001	164507
20	AGE	3	20	78.99	<.0001	160875

Figure 84: Summary of Regression Selection

Stepwise selection was performed on the data to assess the importance of the model and the impact of each predictor variable on the response variable. F-values were calculated, and the model with the highest F-value was "Achievement," indicating a strong association with the response variable and contributing significantly to the model's accuracy.

Subsequent steps added "ToDo Completed," "Places Visited," "Supporting Others," "Imp Daily Stress," "Daily Steps," and "Live Vision," all of which had high F-values, improving the model's accuracy. Other predictor variables like "BMI_Range," "Donation," "Lost Vacation," "Time for Passion," "Social Network," "Fruits_Veggies," "Personal Awards," "Weekly Meditation," "Core Circle," "Daily Shouting," "Flow," "Sleep Hours," and "Age" also had high F-values and contributed to the model.

The study concludes that each predictor variable had a significant influence on the dependent variable, improving the model's accuracy and corroborating the results of the previous analysis.

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	568.8	0.5386	1056.05	<.0001
ACHIEVEMENT		1	1.7086	0.0285	60.00	<.0001
AGE	21 to 35	1	-0.7676	0.1031	-7.45	<.0001
AGE	36 to 50	1	0.6193	0.1106	5.60	<.0001
AGE	51 or more	1	1.6269	0.1290	12.61	<.0001
BMI_RANGE		1	-17.1681	0.1350	-127.15	<.0001
CORE_CIRCLE		1	1.7474	0.0253	69.01	<.0001
DAILY_SHOUTING		1	-1.7242	0.0251	-68.59	<.0001
DAILY_STEPS		1	1.9637	0.0236	83.31	<.0001
DONATION		1	3.4795	0.0387	89.89	<.0001
FLOW		1	1.6292	0.0327	49.84	<.0001
FRUITS_VEGGIES		1	3.6180	0.0485	74.63	<.0001
IMP_DAILY_STRESS		1	-3.9397	0.0514	-76.58	<.0001
LIVE_VISION		1	1.7954	0.0218	82.33	<.0001
LOST_VACATION		1	-1.7568	0.0176	-99.70	<.0001
PERSONAL_AWARDS		1	1.8298	0.0237	77.30	<.0001
PLACES_VISITED		1	1.9120	0.0215	88.85	<.0001
SLEEP_HOURS		1	2.0403	0.0549	37.15	<.0001
SOCIAL_NETWORK		1	1.8892	0.0232	81.28	<.0001
SUPPORTING_OTHERS		1	1.6561	0.0242	68.46	<.0001
TIME_FOR_PASSION		1	1.6061	0.0284	56.59	<.0001
TODO_COMPLETED		1	2.0184	0.0275	73.27	<.0001
WEEKLY_MEDITATION		1	1.6507	0.0229	71.98	<.0001

Figure 85: Analysis of Maximum Likelihood Estimates

The analysis of maximum likelihood estimates shows that each predictor variable in the dataset had a significant impact on the dependent variable. The small standard errors for each predictor variable indicate that their coefficient estimates were reliable and accurate. The high t-values for each predictor variable also suggest that their effects on the dependent variable were both substantial and accurate. The low p-values for each predictor variable indicate that they had a significant influence on the dependent variable.

Analysis of Effects

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
ACHIEVEMENT	1	184180.826	3600.33	<.0001
AGE	3	12123.0939	78.99	<.0001
BMI_RANGE	1	826994.897	16165.9	<.0001
CORE_CIRCLE	1	243622.058	4762.28	<.0001
DAILY_SHOUTING	1	240687.575	4704.92	<.0001
DAILY_STEPS	1	355095.135	6941.34	<.0001
DONATION	1	413369.433	8080.47	<.0001
FLOW	1	127064.955	2483.84	<.0001
FRUITS_VEGGIES	1	284910.995	5569.39	<.0001
IMP_DAILY_STRESS	1	299974.046	5863.84	<.0001
LIVE_VISION	1	346773.087	6778.66	<.0001
LOST_VACATION	1	508515.086	9940.36	<.0001
PERSONAL_AWARDS	1	305691.950	5975.61	<.0001
PLACES_VISITED	1	403829.977	7894.00	<.0001
SLEEP_HOURS	1	70589.5788	1379.87	<.0001
SOCIAL_NETWORK	1	337946.067	6606.11	<.0001
SUPPORTING_OTHERS	1	239787.718	4687.33	<.0001
TIME_FOR_PASSION	1	163852.541	3202.96	<.0001
TODO_COMPLETED	1	274647.332	5368.76	<.0001
WEEKLY_MEDITATION	1	265043.487	5181.02	<.0001

Figure 86: Analysis of Effects

The analysis of effects revealed that most predictor variables had a "Pr > F" value less than 0.05, indicating their statistical significance. The predictor variable with the highest F-value was "BMI_RANGE" with a value of 16165.9, implying a significant effect on the response variable. Other predictors with high F-values were "LOST_VACATION," "DONATION," "FRUITS_VEGGIES," "IMP_DAILY_STRESS," "LIVE_VISION," "PLACES_VISITED," "SOCIAL_NETWORK," "SUPPORTING_OTHERS," "TODO_COMPLETED," and "WEEKLY_MEDITATION," all of which had significant effects on the response variable. Additionally, all predictor variables had p-values less than 0.0001, indicating a significant impact on the dependent variable.

Overall, these results suggest that the predictor variables in the model have a significant impact on the response variable and that their estimates are reliable and accurate.

Analysis of Variance & Model Fit Statistics

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	25397102	1154414	22566.3	<.0001
Error	12755	652502	51.156595		
Corrected Total	12777	26049604			

Model Fit Statistics

R-Square	0.9750	Adj R-Sq	0.9749
AIC	50303.0219	BIC	50305.1049
SBC	50474.4980	C(p)	23.0000

Figure 87: Analysis of Variance & Model Fit Statistics

The high R-square value of 0.975 indicated a high fit, with a considerable portion of the variance in the dependent variable explained. The low AIC value indicated a good fit for the data.

10.0 Discussion & Conclusion

10.1 Model Comparison

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error
Y	Reg2	Reg2	3rd Poly Stepwise Regression	WOR...		50.10447	50.283	12778	17.73122	12778	7.09105
	HPNNA	HPNNA	5L 10N Sine HP Neural Network	WOR...		50.10585	50.69546	12778	18.23692	12778	7.120074
	Reg3	Reg3	2nd Poly Stepwise Regression	WOR...		50.14317	50.45874	12778	18.28283	12778	7.103432
	HPNNA2	HPNNA2	HP Neural 5L 5N	WOR...		50.1902	50.82067	12778	18.20567	12778	7.128862
	Reg5	Reg5	Stepwise Regression	WOR...		50.3678	51.06452	12778	18.33359	12778	7.145944

Figure 88: Fit Statistics of Model Comparison

Cluster models are a type of unsupervised learning model that is used to uncover patterns and relationships within a dataset. It does not have a predetermined target variable, and it is not make predictions about the data. Therefore, it is not included in the model comparison.

Based on the fit statistics table for the six models that were selected, the 3rd Degree Polynomial Stepwise Regression model has the lowest Average Squared Error of 0.02081 (calculated using the formula sqrt(50.10477)/(820-480)). This indicates the model is the most suitable one among the other models, with the highest accuracy in predicting an individual's work-life balance score based on various important lifestyle and habit features.

10.2 Ensemble Model

Ensemble models in machine learning are created by merging multiple models together to form a more powerful and reliable predictive model. The main concept behind ensemble modelling is to aggregate the predictions of multiple models in order to achieve improved overall performance compared to using a single model on its own (Poria, S et al., 2020).

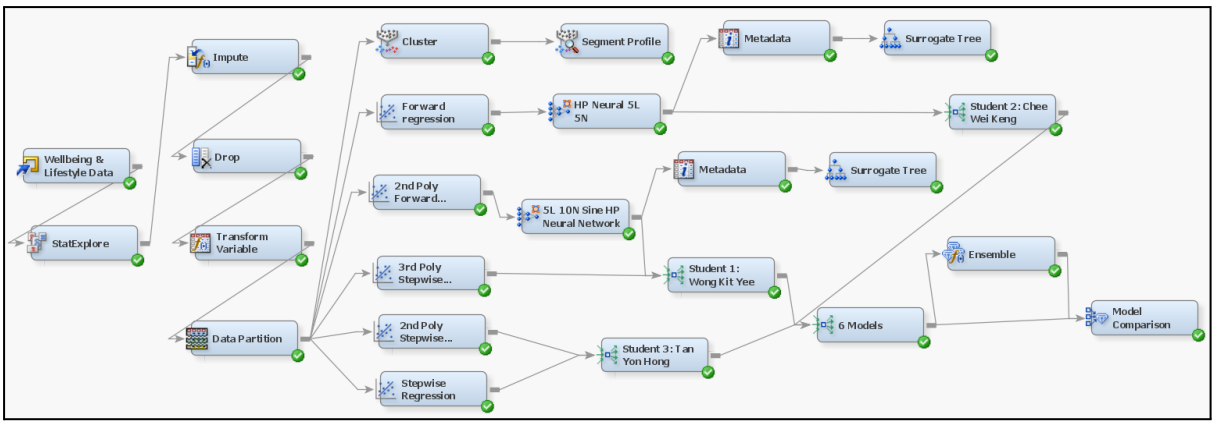


Figure 89: Process Flow Diagram included Ensemble Model

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Train: Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors
Y	Ensmbl	Ensmbl	Ensemble	WORK ...		49.97612	50.45424	12778	17.74406	12778	7.103115	644704.3
	Reg2	Reg2	3rd Poly Stepwise Regression	WORK ...		50.10447	50.283	12778	17.73122	12778	7.09105	642516.1
	HPNNA	HPNNA	5L 10N Sine HP Neural Network	WORK ...		50.10585	50.69546	12778	18.23692	12778	7.120074	647786.6
	Reg3	Reg3	2nd Poly Stepwise Regression	WORK ...		50.14317	50.45874	12778	18.28283	12778	7.103432	644761.8
	HPNNA2	HPNNA2	HP Neural 5L 5N	WORK ...		50.1902	50.82067	12778	18.20567	12778	7.128862	649386.5
	Reg5	Reg5	Stepwise Regression	WORK ...		50.3678	51.06452	12778	18.33359	12778	7.145944	652502.4

Figure 90: Fit Statistics of Model Comparison included Ensemble Model

The fit statistics presented in Figure 90 indicate that the Ensemble model is performing slightly better than the 3rd Degree Poly Stepwise Regression model. It has the lowest Average Squared Error of 0.02079 (calculated using the formula sqrt(49.97612)/(820-480)).

11.0 References

Moulder, J. H. D. (2022, January 13). *10 Reasons Why Work-Life Balance Is Important (In 2022)*. Leadership & Business Coaching for Lawyers | Heather Moulder. https://www.coursecorrectioncoaching.com/reasons-why-work-life-balance-is-important/

Statista. (2022, September 30). *Importance of work-life balance among employees in the U.S. 2018*. https://www.statista.com/statistics/916521/us-economy-importance-work-life-balance/

20+ Alarming Burnout Statistics [2022]: Stress And Lack Of Motivation In The Workplace – Zippia. (2022, December 9). https://www.zippia.com/advice/burnout-statistics/

Hotz, N. (2018, September 10). *What is CRISP-DM? - Data Science Process Alliance*. Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Kenan-Dlagler. (2023, February 11). Being too humble can hurt you | UNC Kenan-Flagler. https://www.kenan-flagler.unc.edu/news/being-too-humble-can-hurt-you/

What is Clustering? | Clustering in Machine Learning. (2019). Google Developers. https://developers.google.com/machine-learning/clustering/overview

Kumar, A., & Chatterjee, P. (2017). "Machine Learning for Clinical Diagnostics." In Handbook of Neural Computation (pp. 497-514). Springer. https://link.springer.com/chapter/10.1007/978-3-319-91479-9_25

Poria, S., Cambria, E., & Hussain, A. (2020). Ensemble learning in machine learning. WIREs Data Mining and Knowledge Discovery, 10(4), e1356. https://doi.org/10.1002/widm.1356

Pathak, M. (2020, November 23). *Quick Guide to Evaluation Metrics for Supervised and Unsupervised Machine Learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/quick-guide-to-evaluation-metrics-for-supervised-and-unsupervised-machine-learning/#:~:text=Clustering%20Performance%20Evaluation%20Metrics&text=Here%20clusters%20are%20evaluated%20based,then%20it%20has%20performed%20well.