**Data Analysis**

There are 6 moderated variables which are “Age”, “Gender”, “Marital Status”, “Education Level”, “Work Industry”, “Work Position”.

Age

|  |  |
| --- | --- |
|  |  |

The preliminary analysis done for the “Age” Variable indicates a fairly even distribution between the age range of “< 25 years”, “26 – 40 years”, “41 – 55 years” and “above 55 years”. This is primarily because the survey were targeted for university students and adults whom have working experience. The descriptive statistics indicates that 29% of respondents were below 25 years old, 33.22% were around 26 – 40 years and 30.07% were around 41- 55 years. Based on the percentage result, it is safe to assume that there won’t be any bias based on age group, as the percentage for each age group is roughly the same and is evenly distributed.

When comparing age groups with the targeted variable of “In the next six months, do you plan to purchase anything using the E-payment mode?”, the graph indicates that majority of each respondents in each age group are highly likely to adopt e-payment features.

Gender

|  |  |
| --- | --- |
|  |  |

Marital Status

|  |  |
| --- | --- |
|  |  |

Education Level

|  |  |
| --- | --- |
|  |  |

Work Industry

|  |  |
| --- | --- |
|  |  |

Work Position

|  |  |
| --- | --- |
|  |  |

**Correlation Among Features**

Correlation analysis is a technique of analyzing the linear relationship between two variables. The two variables can be independent or dependent based on the strength of the relationship computed. We define the strength of the correlation as correlation coefficient. Correlation coefficient may be derived from various formula such as Pearson’s Correlation coefficient, Spearman’s Correlation coefficient and many more.

The importance of correlation analysis iis to determine the strength of relationship between two variables. Normally, it provides inside when trying to analyze causality in a dataset. Based on the strength and value of the correlation coefficient, we are able to determine whether the variable positively or negatively affect the predictor attribute.

One of a Hypothesis we commonly have is that attributes that rank higher will have higher effect in classification accuracy. Therefore, correlation analysis shall be conducted to identify the important attributes that increases the classification accuracy and filter out irrelevant features. The features that are selected may contain an underlying factor, that we shall investigate in our project as well.

To undergo such experimentation, we use two correlation coefficient methods, that is Pearson’s Correlation and Spearman’s Correlation Coefficient.

**Why do we use Correlation?**

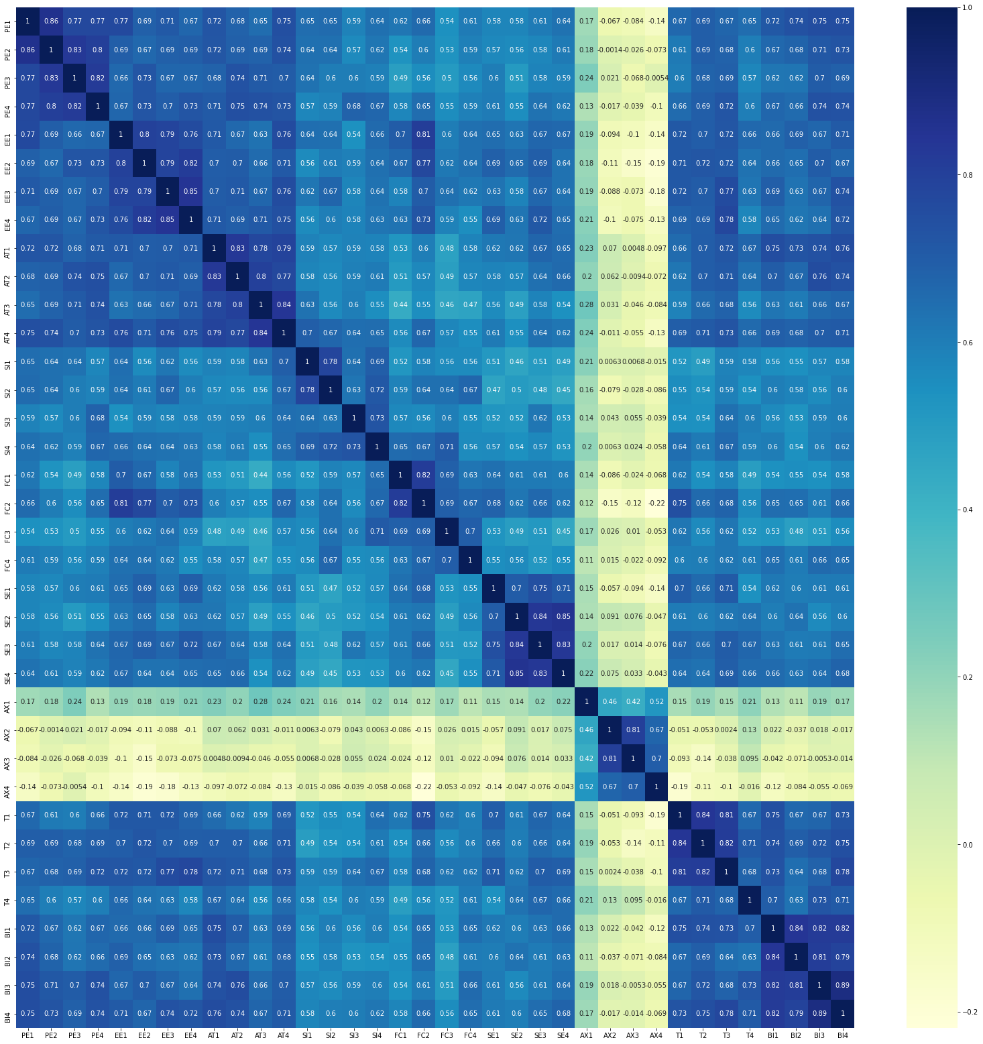
The main issue is that if we use another metric like information gain, we would not be able to select features that have high correlation. Assuming our initial hypothesis is correct, it would mean that highly correlated features are the features that have the underlying meaning.

Diagram

Description automatically generated

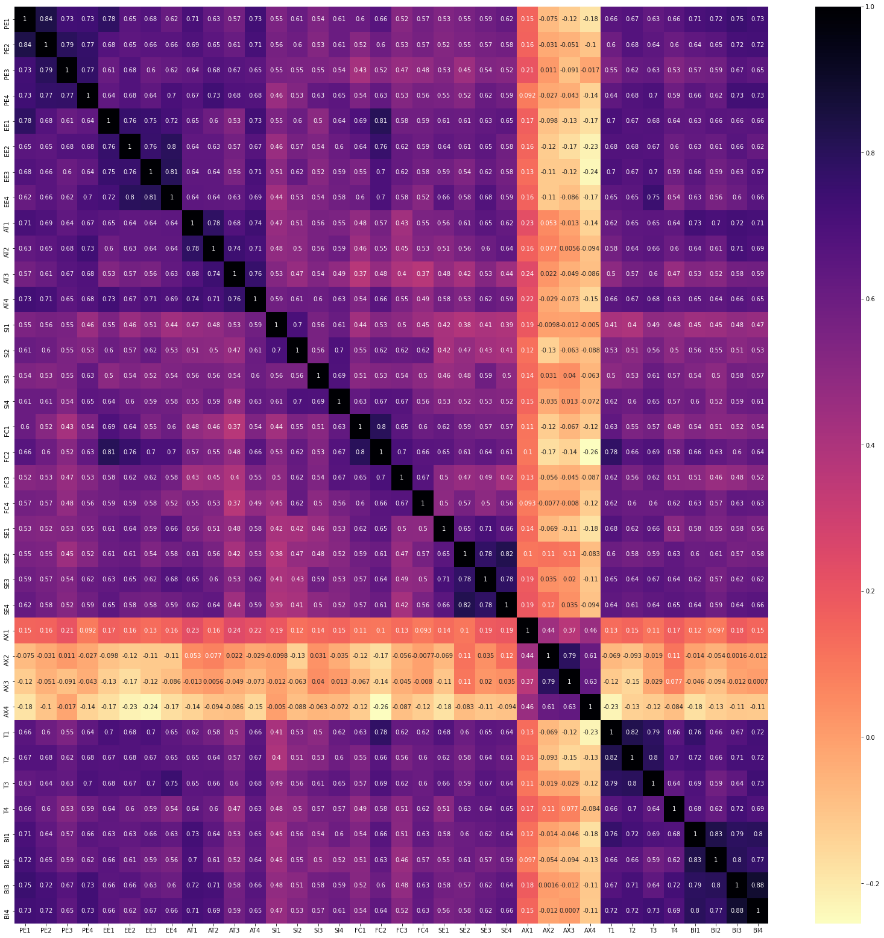
**Pearson’s Correlation Coefficient**

The table of computed result of UTAUT Factor Variables using Pearson’s Correlation:



**Spearman’s Correlation Coefficient**

The table of computed result of UTAUT Factor Variables using Spearman’s Correlation:



**Correlation Network Graph**

Network graphs is a mathematical structure to show relations between points in a less statistical manner. The graph visualizes how subjects are interconnected with each other. Network graphs allows us to study relations between discrete objects or actors. In this scenario, we will be using network graph to visualize the relationship between the UTAUT Factors in a more aesthethic pleasing manner.

Network graphs are made up of nodes and edges. Nodes or vertices are the discrete entities of the graph or dataset. Edges or links are relations among the nodes. Weighted edges can be used to represent correlation coefficient.

Chart, line chart

Description automatically generated

The most important measure in a network is centrality. Centrality is linked up with ideas of prominence, status, social capital or importa

Centralness

Degree

Closeness

Betweeness

Graph is constructed using

Network graph (force directed graph) is a mathematical structure (graph) to show relations between points in an aesthetically-pleasing way. The graph visualizes how subjects are interconnected with each other. Entities are displayed as nodes and the relationship between them are displayed with lines. The graph is force directed by assigning a weight (force) from the node edges and the other interconnected nodes get assigned a weighted factor. The graph simulates the weight as forces in a physical system, where the forces have impact on the nodes and find the best position on the chart’s plotting area. The Network Graph has various use case such as display relations between people, roads, companies, and products.

From the above graph, the central node is the most important

**Feature Selection Methods**

Feature selection can be classified into filter, wraper and embedded methods. The performance of feature selection method is usually evaluated by the machine learning algorithm. The feature selection method should have high learning accuracy but less computational overhead.

Feature selection is a technique that has an ability to decrease the umber of attributes by eliminating the least significant features. However, the issue with feature selection is finding the optimum features.

In this study we shall stu

**Filter Methods Data**

* **Info Gain**
* **Correlation Feature Selection**
* **Chi-Sqaure Analysis**

**Wrapper Methods**

* **Forward-Selection**
* **Backward-Selection**

**Potential Target Variables**