**Section B: Possible Insight on the Speed Dating Dataset**

There are many different exploratory ideas one can get within this speed dating dataset. The idea here is to focus more on the chances of getting a match with a speed date partner and go on a second date.

**Those possible insight are:**

**Insight 1:**

Do extroverts have higher chance to get second dates compared to introverts?

The general public misconception that we all have is that, when we thought of introverts we think of shyness, social awkwardness or the lack of sociability. The same misconception happens we think of extroverts as outgoing and loves to be in a large crowd.

There is no such thing as pure introverts or pure extroverts in this world. A person is consist of both extroverts and introverts characteristics. Which means people who has more introverts characteristic is not necessarily shy when meeting new people.

For this dataset, we define extroverts as peoples who have more interest in activities that have higher chance to trigger interaction with a large crowd of people, introverts as people who have more interest in activities that emphasizes personal time and both their frequency of going out per week.

**Insight 2:**

Do people who go on dates frequently have higher chance to get second date than those who does not?

For the attribute go on dates frequently in the dataset have a fairly vague meaning, it can mean how frequent that a person went on dates with different person, or it can also mean that how frequent a person went on dates with the same person.

People who goes on date frequently would have slightly more advantage and experience in terms of what to expect in dating, they may even comes up with interesting topic or pickup lines that may interest their speed dating partners to leave a good impression that is nice enough to land them a chance for second date.

For this insight, we only consider the case where by the people went on dates with the same person each time which is an advantage for this speed dating case.

**Section C: Type of data mining technique**

For second insight, we think that classification would be relevant as the technique for extracting the rules from the training data that will predict the categorical labels of the “match” column in the data set. This process is also known as supervised learning because the class label is known for each of the training samples.

At the initial step, the original data set will be treated with pre-processing tasks to deal with possible data quality issues like duplication of records and missing values. After getting the clean and tidy data, the next step is to split the data into 2/3 where it will be used as the training samples and the remaining 1/3 as the testing samples. Random sampling without replacement is applied in this case as the technique for choosing both the training and testing data sets. In addition to that, the original data set comes with 195 columns of attributes, however the main focus for this discussion is to build a model that will use the dater’s frequency of going out on a date and several basic information like gender and age as the predictor columns to predict the class label for the “match” column (1=yes, and 0=no).

These predictor columns are named as iid, gender, age and date in the data set.

Table below shows the example of the training data where the records have been labelled with 0 or 1 in the “match” column (the meaning of each columns is described at below).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| iid | gender | age | date | match |
| 1 | 0 | 21 | 4 | 0 |
| 2 | 1 | 21 | 1 | 1 |

iid Unique subject number

gender Gender of the subject

* 0= Female
* 1= Male

age Age of the subject

date Frequency of going dates

* 1= Several times a week
* 2= Twice a week
* 3= Once a week
* 4= Twice a month
* 5= Once a month
* 6= Several times a year
* 7= Almost never

match Whether the dating match or not

* 0= Yes
* 1= No

The goal of classification is to create a set of classification rules called a classifier or model that will do the work of classifying or categorizing a given data record using the classification algorithm. Therefore, the choice of algorithm is very important depending on the goal and the nature structure of the data set. In this task, we are going to learn a classifier model using the Naive Bayes algorithm. The reason of choosing this algorithm is because it is one of the most efficient and effective algorithm for supervised learning. It assumes that the features are independent. Each feature contributes to the probability to identify the classification independently. For instance, a fruit can be classified as an apple based on the color, circle shape and the diameter. Although the features exist relationship among themselves, the features contribute independently in the probability to identify the fruit is apple (Ray, 2015).

Just to recall, in the process of building a classifier, the attribute to be predicted are known in the training set, but not known in the testing set. The predictor columns are used to determine the value of the column “match” in this case. After the classifier model has been generated, the testing set will be used to test the classifier for accuracy, robustness, and speed. To visualize the performance of this classifier, we can adopt receiver operating characteristic (ROC) to plot a two-dimensional graph in which false positive rate is plotted on the x-axis and true positive rate is plotted on the y-axis. ROC is suitable in this case because our classifier is a binary system, which mean that the result of the output is either 0 or 1.

**Reference:**

Ray, S. (2015). *6 easy steps to learn naive bayes algorithm (with code in python)*. Retrieved from <https://analyticsvidhya.com/blog/2015/09/naive-bayes-explained/>

Fawcett, T. (2003). *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers.* Retrieved from http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf