WHAT DO PEOPLE LOOK FOR IN A PARTNER? SPEED DATING ANALYSIS

CS989: BIG DATA FUNDAMENTALS

BELOSLAVA PETROVA BOSILKOVA
REG. NO 201869525

Word count: 3151

1. Introduction

Participants in speed dating events meet with several potential partners in a short space of time, usually under 5 minutes. At the end of these short meetings, participants mark down their opinions of each other and indicate whether they would like to arrange a second date (Speed Dater, 2018).

The aim of this report is to analyse how accurately participants in speed dating events can predict what attributes they will look for in the opposite gender. Furthermore, these desired attributes will be compared with the actual attributes which participants assigned to their matches and non-matches.

Section 2 will commence with a description of the dataset used in this study. Afterwards, an identification of key challenges to be addressed will be explained in Section 3. Further exploration with the help of visualisation analysis will be introduced in Section 4 and upon obtaining a better understanding of the variables, agglomerative clustering methods will be presented in Section 5 followed by decision tree classifiers in Section 6. Finally, Section 7 will contain a detailed reflection on the methods used and Section 8 will conclude the report.

2. Data description

The “Speed Dating Experiment” dataset was obtained from Kaggle.com and was compiled by Ray Fisman and Sheena Iyengar. The information was collected from several speed dating events conducted on an experimental basis between 2002 and 2004. According to the key provided, the experiment consisted of 21 waves where both men and women were asked to give a brief description of themselves and state their preferences based on six criteria, namely Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. During the event, participants were asked to rate each of their dates based on the provided criteria and state whether they would like to arrange another meeting with them (Kaggle, 2016).

3. Identification and description of key challenges

The fact that information was collected both before and after participants met allows for valuable observations into their behaviour. Specifically, whether men and women differ in their preferences for attributes in the opposite gender. Moreover, comparisons can be made between participants’ attribute preferences prior to attending the events versus what they appear to be attracted to as indicated by their decisions regarding matches. Whether or not higher ratings in each attribute correlates with higher percentage of matches for both men and women will also be revealed.

4. Descriptive Statistics

4.1 Data preparation

To begin with, importing the data using the necessary packages resulted in a decoding error.

```
UnicodeDecodeError: 'utf-8' codec can't decode byte 0x8e in position 17:
invalid start byte
```

Figure 1: Python error while importing the data set from Microsoft Excel

The following code was used to solve this problem (Stack Overflow, 2015):

```
df = pd.read_csv('C:\\Speed_Dating_Data.csv', encoding="ISO-8859-1")
fields = df.columns
```

Figure 2: Python code obtained from to fix the decoding error from Figure 1

As most of the information provided in this dataset was not required for the analysis, the desirable data was extracted. To differentiate between participants, the gender column containing the binary variables 0 for females and 1 for males was selected. Furthermore, to compare the predisposed desired attribute distribution of the participants with the attribute ratings they gave after each date, the columns corresponding to these variables were also selected. Additionally, the decision column which indicated participants' preference regarding which partners they intended to see again, containing 0 for no and 1 for yes, was included. Finally, empty cells and cells containing non-numerical values were removed to ensure that the analysis was not inadvertently skewed (Stack Overflow, 2016).

For the purpose of differentiating between males and females, the selected data was split by gender with the summary statistics for both are presented in Appendices 3 and 4. Then the attribute ratings, given on a scale of 1-10, which participants submitted after meeting with their partners were turned into percentages. This allowed for comparisons to be made with the desired attributes which participants outlined at the outset of the dating events, as these were already given in percentages in the initial dataset. To obtain the average attribute distribution values for the “Yes” responses, all rows containing 0 in the decision column were removed. Likewise, rows containing 1 in the decision column were removed for the “No” responses.

4.2 Female Participants

Figures 3, 4 and 5 represent three radar charts which visualise comparisons of desired attributes prior to participating in the experiments and the average attribute distributions associated with “Yes” and “No” responses as percentages of the total ratings. The code for the radar charts was obtained from The Python Graph Gallery (2017). Detailed description of each percentage distribution is presented in Appendices 5 to 7. As evident by Figure 3, it appears the females settled significantly in the “Attractiveness” and “Intelligence” categories when deciding on their partners. There were also minor reductions in their expectations for “Fun” and “Sincerity”. On the contrary, they seem to place more importance on “Ambition” and “Shared Interests”. This would lead us to believe that the main attributes which yield successful matches from males’ perspective are “Ambition” and “Shared Interests”.

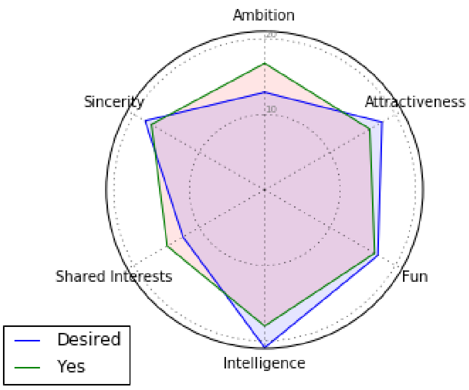


Figure 3: Radar Chart comparing average values of desired attributes and average values of attributes associated with “Yes” responses from female participants.

However, looking at Figure 3 in isolation could be misleading, as Figure 4, comparing desired attributes to “No” votes, contains quite a few similarities. Namely, a reduced focus in “Attractiveness”, “Fun” and “Intelligence” and increase in “Ambition”. Thus, a direct comparison between “Yes” and “No” votes may be the most revealing in terms of what women truly look for in a potential partner.

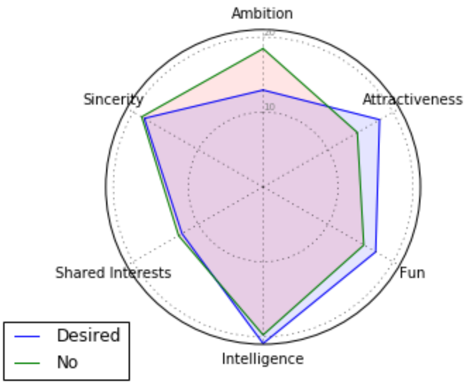


Figure 4: Radar Chart comparing the distribution of average desired attributes and the average attributes associated with "No" responses from female participants.

Whereas Figure 3 would indicate that "Attractiveness" and "Fun" do not have a large impact on a woman's decision, Figure 5 suggests that these are in fact two of the three attributes which increase the likelihood of a match occurring, the third being "Shared Interests". On the other hand, "Intelligence" and "Sincerity", the two attributes which women considered to be the most important prior to the dating events appear to be less significant. Finally, the fact that "Ambition" commanded a greater percentage of importance after meetings than it did prior, regardless of "Yes" or "No" being indicated, may be due to the ambiguity of the attribute compared to the other attributes which are clearer in their calculation.

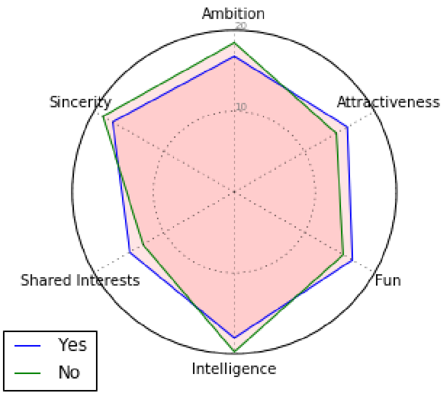


Figure 5: Radar chart comparing average values of attributes associated with "Yes" responses and average values of attributes associated with "No" responses from female participants.

To clear the confusion arising from the fact that the radar chart displays “Intelligence” as a percentage of the total attribute ratings, it is reasonable to compute the average scores relating to “Yes” and “No” ratings without adjustments as depicted in Figure 6.

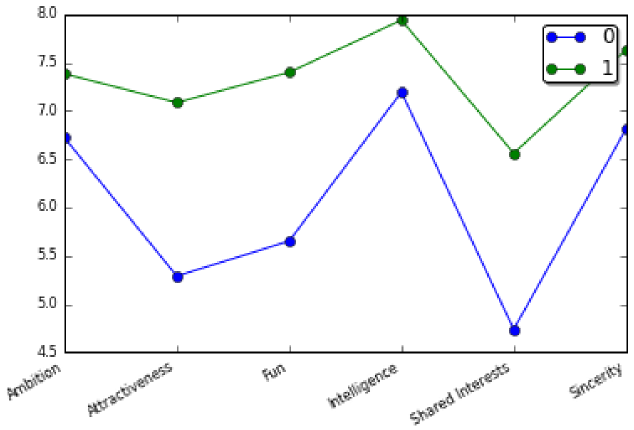


Figure 6: Line Graph comparing the average “No” (corresponding to line 0) and “Yes” (corresponding to line 1) ratings for female participants. The code for the x-axis was obtained from Stack Overflow (2012).

Looking only at Figure 5, it appears that female participants reject more intelligent male participants. However, Figure 6 reveals that females in fact considered males which they voted “Yes” on to be more intelligent than males they voted “No” on. In fact, females considered males who they voted “Yes” on to be more desirable in all six categories. This was not apparent in Figure 5 because the more well-rounded attributes of the “Yes” votes make attributes like “Intelligence” seem less important in comparison.

4.3 Male Participants

The same process was carried out for the preferences and decisions of male participants in Figures 7, 8 and 9. The corresponding percentages for male attributes are shown in Appendices 8-10. Figure 7 reveals that male participants placed a greater emphasis on “Attractiveness” in their desired attributes than female participants did and a lesser emphasis on “Ambition” and “Sincerity”. However, similarly to their female counterparts, males considered “Ambition” and “Shared Interests” to be significantly more important, and “Attractiveness” less so, in their “Yes” decisions.

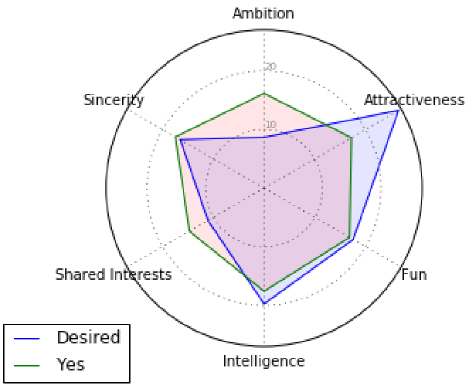


Figure 7: Radar Chart comparing average values of desired attributes and average values of attributes associated with “Yes” response across male participants.

The shape shown in Figure 8 does not differ greatly from Figure 7, with almost equal values throughout. Thus, examining Figure 9 proved to be the most revealing in terms of male decision making.

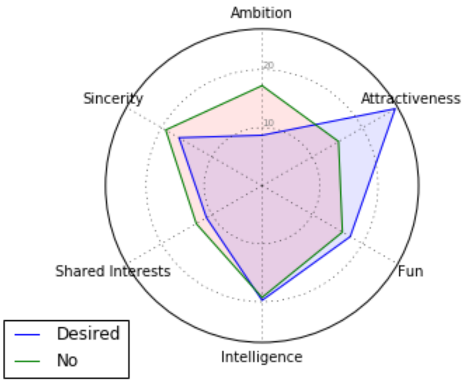


Figure 8: Radar Chart comparing the distribution of average desired attributes and the average attributes across male participants.

Male participants were able to anticipate more accurately which attributes would be more important to them than their female counterparts were, as they correctly predicted “Attractiveness” and “Fun” to be two key characteristics. However, they underestimated the influence that “Shared Interests” would have on their decisions in the same way that females did.

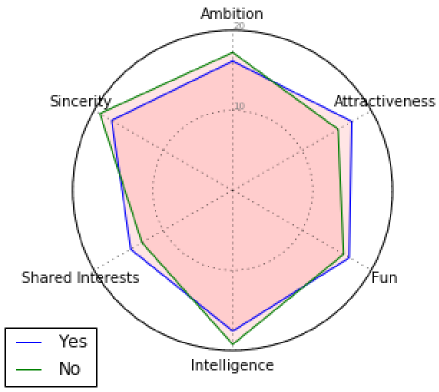


Figure 9: Radar chart comparing average values of attributes associated with “Yes” response and average values of attributes associated with “No” response across male participants.

Figure 8 suggests that males voted “No” for females they regarded as more ambitious, sincere and intelligent as compared to their “Yes” matches. Just like in the previous section, this statement needs to be checked against the line graph (Figure 10) to ensure that this is the case.

As expected, Figure 10 reveals that partners who received “Yes” votes were considered more desirable in all six categories than partners who received “No” votes. The main difference in the ratings which males and females gave for one another lies in the “Attractiveness” category. On average, males rated their partners higher than females did regardless of their decision.

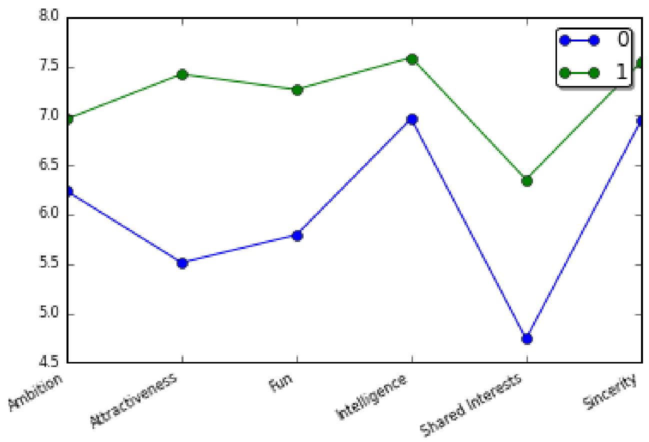


Figure 10: Line Graph comparing the average “No” (corresponding to line 0) and “Yes” (corresponding to line 1) ratings for male participants. The code for the x-axis was obtained from Stack Overflow (2012).

5. Description, rationale, application and findings from one unsupervised analysis method

5.1 Description

This section will attempt to cluster the variables. Clustering is best known as a technique for grouping objects based on similarities. One proven method is Hierarchical Clustering which entails construction of nested clusters through splitting or successive merging of the given data. Better known as a dendrogram, this takes the visual appearance of a tree. A single cluster at the root comprises of all the samples and each of the leaves represents a data point. The intermediate points of the dendrogram represent the proximity of objects from one another and the height indicates the distance between clusters and data points. This analysis will use agglomerative clustering which involves placing each sample in its own cluster at the start. Subsequently, these clusters are joined with the most similar cluster (EMC, 2015), (Igual and Segui, 2017). The selected data was clustered according to the ward linkage method which computes the distance between clusters by minimizing the sum of squared dissimilarities followed by a computing the Euclidean distance between the points (Igual and Segui, 2017). The code used for the clustering was obtained through a video tutorial on Youtube(2017).

5.2 Rationale

Since interpretation is relatively uncomplicated, hierarchical clustering is preferred for the unsupervised method in this analysis. Explaining the basis of forming clusters and their visualisation on various scales is made simple using the dendrogram. Data points can be easily partitioned by choosing a cut at a specific level. Moreover, the hierarchical structure of the output is more informative than the flat structure of the clusters returned by K-means. (Igual and Segui, 2017).

5.3 Application

According to the key, there were 21 waves of participants, each having their own unique identification number. After separating the data by gender, it was possible to calculate average attribute ratings and the average match rates for all participants by using nested loops. The first loop is a straightforward “For Loop” which looks at each wave individually. The nested loop uses two variables, the largest participant ID number and the smallest participant ID number to calculate how many dates participants had. Then average attribute ratings and the percentage of matches that each date received a "Yes" from were calculated. These average values were stacked into a new array, later to be turned into a dataframe for clustering purposes.

5.4 Findings

Applying the Ward's linkage method to the selected data produced the dendrogram displayed in Figure 11. The figure indicates that the threshold measure should be set in the middle, as per Xu and Wunsch (2009), which implies that the optimal number of clusters should be between two and three. For this analysis the number of clusters is set to three to allow for more freedom in modelling the participants' attributes as per Figure 11.

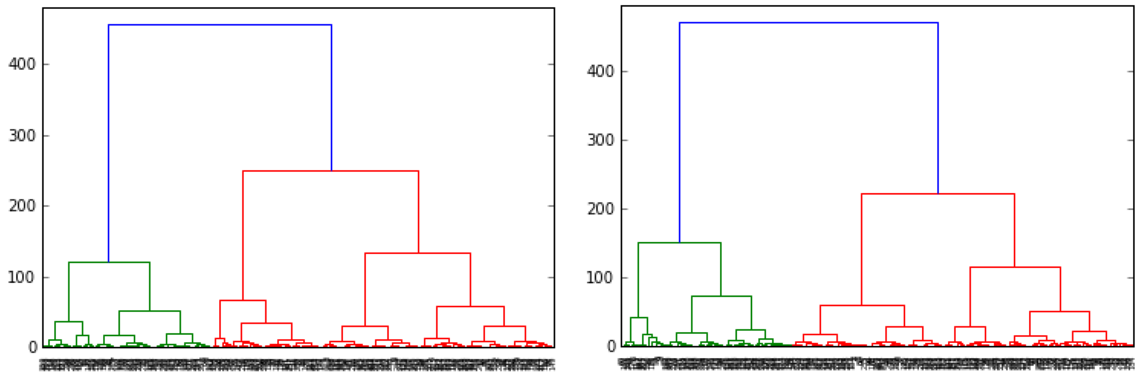


Figure 11: Dendrogram showing female participants' clusters (on the left) and male participants' clusters (on the right) generated from the compounded average attribute data by using the Ward's method.

Figure 12 displays the two-dimensional plot of attribute ratings against the average matches which male participants have received. The red cluster shows participants who have received the lowest percentage of matches. Lower match percentages do not always indicate lower attribute ratings, however, as shown in the plots for "Intelligence", "Sincerity" and "Ambition". These plots illustrate that although higher ratings for these attributes do not hinder matches from occurring, they also do not indicate a significant increase in percentage of matches. On the other hand, as displayed in the "Fun", "Attractiveness" and "Shared Interest" plots, participants within the blue and yellow clusters received both higher ratings and higher percentages of matches than those in the red clusters.

Similarly, clusters showing female attribute ratings and their respective percentages of matches reveal the same upward trends, although less pronounced than in Figure 13, in the “Fun”, “Attractiveness” and “Shared Interests” plots. A noticeable varation in shape from its male counterpart is seen in the “Sincerity” plot, where female atribute ratings are positioned more densely. This suggests that female sincerity was the most consistent gender-attribute combination regardless of match percentage.

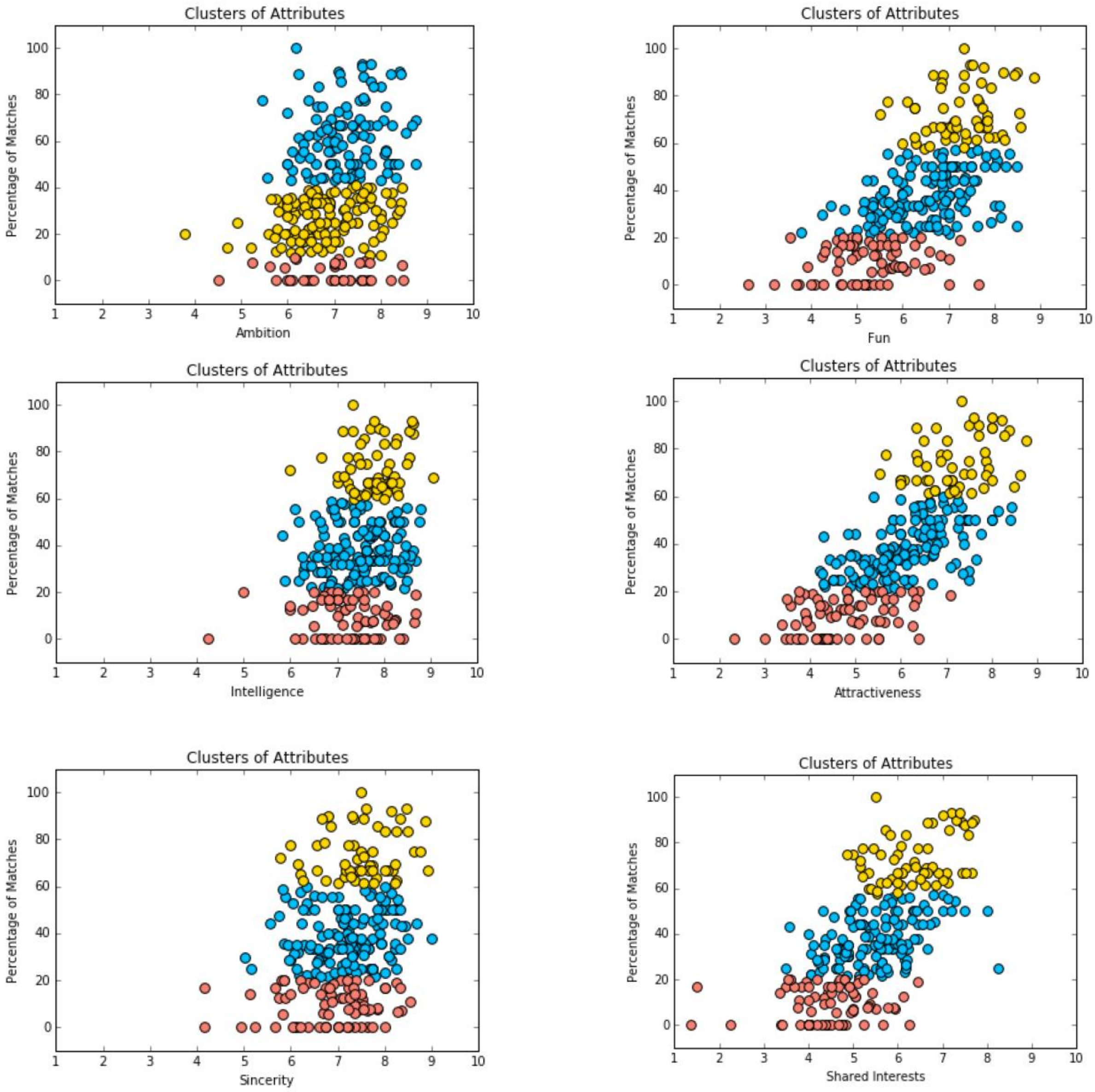


Figure 12: Agglomerative Clustering of attribute ratings against percentage of matches for male participants

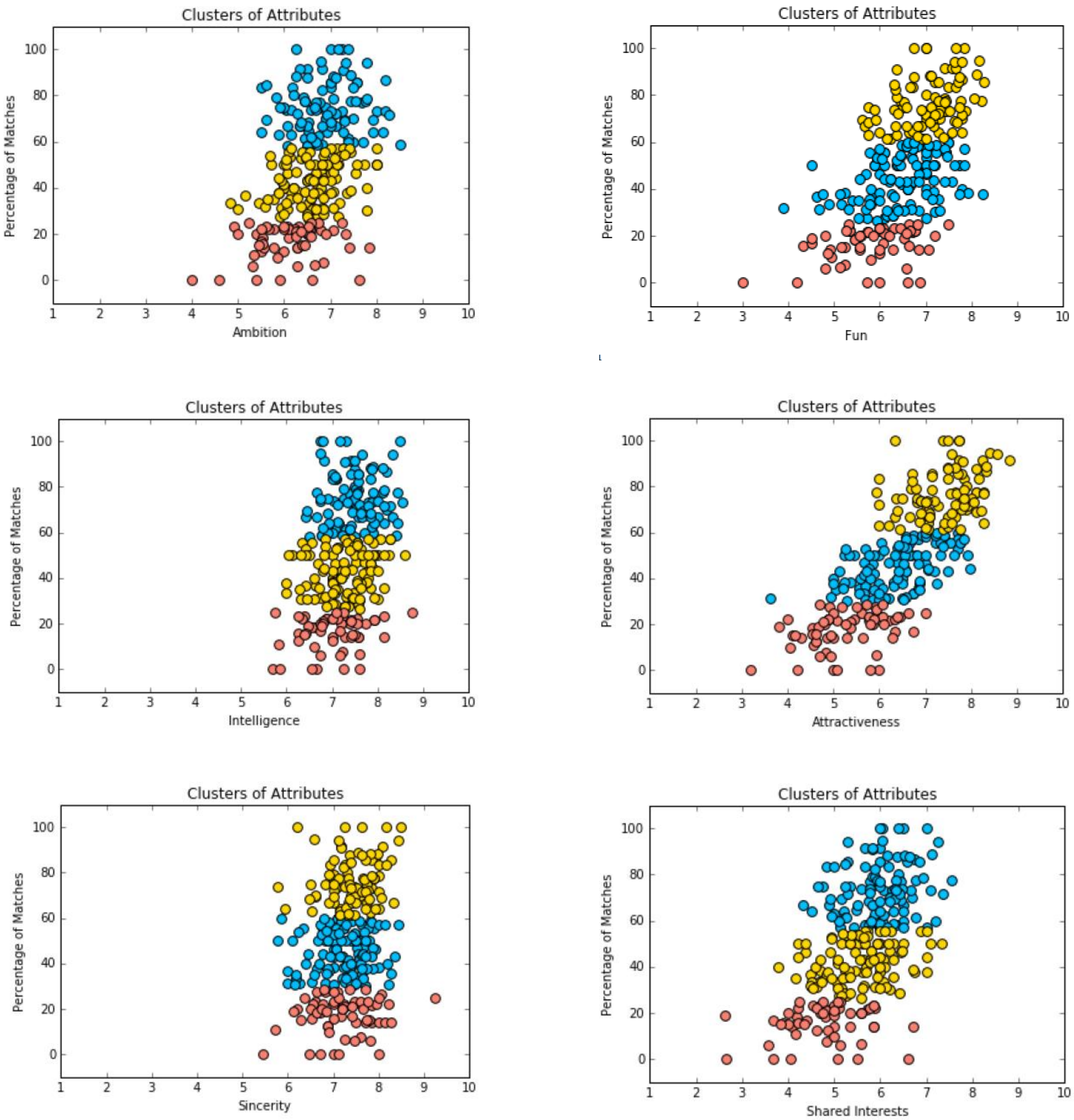


Figure 13: Agglomerative Clustering of attribute ratings against percentage of matches for female participants

6. Description, rationale, application and findings from one other analysis

6.1 Description

A decision tree is a type of supervised learning method which resembles the shape of a tree. The nodes of the tree are computed on the basis of a given attribute, computed by a Gini index, and the corresponding branches show the result of the performed test. In turn, the leaf nodes indicate class labels and the decision is calculated on the base of the computation of all attributes. As with other supervised learning methods, decision trees employ a data set which is split into training and test data. The former is used to fit the given model whereas the latter's purpose is to produce an unbiased prediction based on the same model. In the case of decision trees, the training data is set at the root of the tree and the rest of the data is looped through and partitioned in accordance with the given attributes (Swamynathan, 2017).

6.2 Rationale

Classification trees are easily comprehendible and simple to follow due to their logical nature. Another advantage of using decision trees over other supervised methods is that they remain accurate even when there are missing values in the data set (Hastie, Tibshirani and Friedman, 2016).

6.3 Application

Although decision trees have several advantages, using a large data sets may disrupt the outcome, as the tree may expand to create a new leaf for each observation (Swamynathan, 2017). As such, the attributes were limited to "Attractiveness" and "Ambition", to observe the relationship between the most impactful attribute and one of the attributes with lesser impact.

Data was imported for participant gender, attractiveness, ambition and decisions and then the selected data was separated by gender. For splitting the data into training and test data, X was set to include all rows for the "Attractiveness" and "Ambition" columns while Y contained the binary decision variable with 70:30 ratio.

For the binary tree characteristics, the nodes were computed by the default criterion, the Gini index. The maximum depth of the classification tree was set to three to prevent the tree from growing ltoo leaves and the minimum requirement for samples to appear in a leaf node was set to two. Following this, the model was fit to the data and the predictions for the built model were made.

In order to visualise the decision tree the graphviz package was imported. This creates graphs which are saved as string files and need to be visualised with the help of http://www.webgraphviz.com/. The class names indicate whether a match has occurred, and the feature names were related to the variables. The code was extracted from Dataaspirant (2017).

6.4 Findings

Appendices 11 and 12 display the decision trees for both female and male participants. Classification reports and confusion matrices for male participants and female participants are presented in Appendix 13. Using the Gini index resulted in "Attractiveness" being the initial node. The Gini score, which indicates the split quality by the degree of sample mixing, is closer to its maximum value of 0.5, meaning the variable split was inefficient. Following the branches on the right side of the tree right leads to female participants receiving matches. This was achieved by having an "Attractiveness" rating greater than 7.25 or by having an "Attractiveness" rating greater than 6.75 supplemented by an "Ambition" rating greater than 4.5. Less than 6.75 ratings in "Attractiveness" yielded no matches. The male decision tree was completely identical in match distribution except for the initial node, where "Attractiveness" was split by 6.25 rather than 6.75.

7. Reflection on methods used for analysis

Although the methods used for this analysis provided sufficiently satisfying results, K-means clustering may have provided a more thorough outcome as it is more suitable for larger datasets (Igual and Segui, 2017). As for supervised analysis, the high Gini scores obtained from the decision tree method, although accurate, indicated a low split quality. In this case, a logistic regression might generate predictions more appropriately as it measures the relevancy as well as the direction of the predictor coefficient.

8. Conclusion

This report has analysed the behaviour of speed dating event participants. It has compared their desired attributes for potential matches prior to going on any dates with their opinions and decisions regarding fellow speed daters with whom they interacted. Summary statistics, radar charts, pie charts and line graphs were used to visualise the relationships between the variables. The most prominent outcomes from these were firstly that "Ambition" and "Shared Interests" were underrated attributes prior to going on dates by both males and females.

Secondly, participants who received “Yes” decisions from their partners were more well-rounded and more highly rated on average in every attribute. Agglomerative clustering of each of the attributes and percentage of matches has shown that higher ratings in “Attractiveness”, “Fun” and “Shared Interests” lead to significantly higher match percentages. Finally, the decision tree classifiers revealed that neither gender received matches while having low “Attractiveness” ratings. Further investigation utilizing K-means clustering and logistic regression methods would likely return even more useful information.

References

Dataaspirant. (2017). Building Decision Tree Algorithm in Python with scikit learn. [online] Available at: http://dataaspirant.com/2017/02/01/decision-tree-algorithm-python-with-scikit-learn [Accessed 4 Nov. 2018].

EMC (2015). Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. EMC Education Services, pp.117-136.

Hastie, T., Tibshirani, R. and Friedman, J. (2016). The Elements of Statistical Learning. New York: Springer.

Igual, L. and Segui, S. (2017). Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications. Springer, pp.116-120.

Kaggle. (2016). Speed Dating Experiment. [online] Available at: https://www.kaggle.com/annavictoria/speed-dating-experiment [Accessed 4 Nov. 2018].

Speed Dater. (2018). What is Speed Dating? | SpeedDater. [online] Available at: https://www.speeddater.co.uk/sd-news/what-is-speed-dating [Accessed 5 Nov. 2018].

Stack Overflow. (2012). Rotate axis text in python matplotlib. [online] Available at: https://stackoverflow.com/questions/10998621/rotate-axis-text-in-python-matplotlib [Accessed 4 Nov. 2018].

Stack Overflow. (2015). Encoding a csvreader. [online] Available at: https://stackoverflow.com/questions/31444177/encoding-a-csvreader [Accessed 4 Nov. 2018].

Stack Overflow. (2015). How to make a pie chart smaller with matplotlib. [online] Available at: https://stackoverflow.com/questions/29786807/how-to-make-a-pie-chart-smaller-with-matplotlib [Accessed 4 Nov. 2018].

Stack Overflow. (2016). Pandas - dropping rows with missing data not working using .isnull(), notnull(), dropna(). [online] Available at: https://stackoverflow.com/questions/39339935/pandas-dropping-rows-with-missing-data-not-working-using-isnull-notnull [Accessed 4 Nov. 2018].

Swamynathan, M. (2017). Mastering Machine Learning with Python in Six Steps. Apress, pp.176-180.

The Python Graph Gallery. (2017). Radar chart with several individuals. [online] Available at: https://python-graph-gallery.com/391-radar-chart-with-several-individuals/ [Accessed 4 Nov. 2018].

Xu, R. and Wunsch, D. (2009). Clustering. Hoboken, N.J.: Wiley, pp.31-61.

YouTube. (2017). Machine learning with Python and sklearn - Hierarchical Clustering (E-commerce dataset example). [online] Available at: https://www.youtube.com/watch?v=7BPLNOMNIXM [Accessed 5 Nov. 2018].

Appendices

Appendix 1: Python's version

```
3.5.2 |Anaconda 4.2.0 (64-bit)| (default, Jul  5 2016,
11:41:13) [MSC v.1900 64 bit (AMD64)]
```

Development environment: Spyder (Python 3.5)

Appendix 2: Python packages used as part of the analysis

```
#Importing the necessary packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering
import sklearn
from sklearn import tree
from sklearn import metrics
from sklearn import cross_validation
from sklearn.tree import DecisionTreeClassifier
import graphviz
```

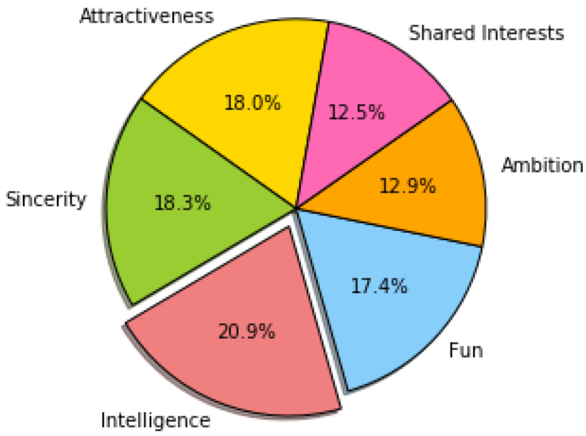
Appendix 3: Summary statistics for female participants

	gender	attr1_1	sinc1_1	intel1_1	fun1_1	amb1_1	shar1_1	attr	sinc	intel	fun	amb	shar	dec
count	3385	3385	3385	3385	3385	3385	3385	3385	3385	3385	3385	3385	3385	3385
mean	0	17.96	18.31	20.90	17.35	12.93	12.50	5.96	7.12	7.47	6.31	6.97	5.42	0.38
std	0	9.31	6.55	6.68	5.44	5.45	5.68	1.99	1.84	1.57	2.04	1.82	2.18	0.48
min	0	0	0	2	0	0	0	0	0	0	0	0	0	0
25%	0	14	15	18	15	10	10	5	6	7	5	6	4	0
50%	0	15.38	20	20	18	15	12.5	6	7	8	6	7	6	0
75%	0	20	20	25	20	16.67	16	7	8	9	8	8	7	1
max	0	90	47	50	40	30	30	10	10	10	10	10	10	1

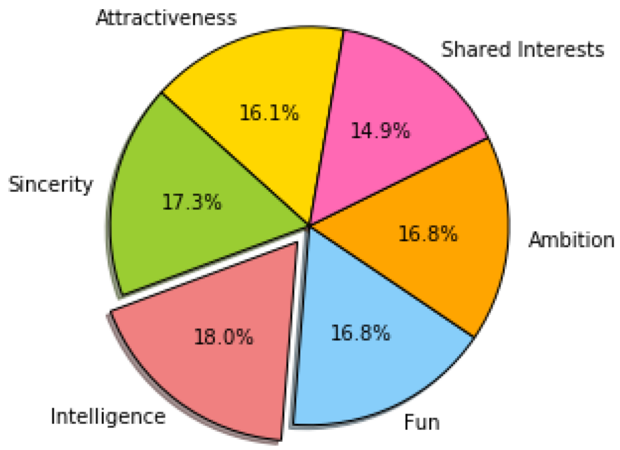
Appendix 4: Summary statistics for male participants

	gender	attr1_1	sinc1_1	intel1_1	fun1_1	amb1_1	shar1_1	attr	sinc	intel	fun	amb	shar	dec
count	3547	3547	3547	3547	3547	3547	3547	3547	3547	3547	3547	3547	3547	3547
mean	1.00	26.52	16.54	19.75	17.56	8.68	11.07	6.45	7.24	7.27	6.51	6.59	5.53	0.49
std	0.00	13.27	7.34	6.70	6.70	5.90	6.68	1.85	1.61	1.50	1.83	1.72	2.09	0.50
min	1	6.67	0	0	0	0	0	0	0	0	0	0	0	0
25%	1	19.57	10.53	17	15	5	5	5	6	6	5	5	4	0
50%	1	23	17.39	20	18	10	10	6	7	7	7	7	5	0
75%	1	30	20	23.26	20	12.5	15.09	8	8	8	8	8	7	1
max	1	100	40	42.86	50	53	30	10	10	10	10	10	10	1

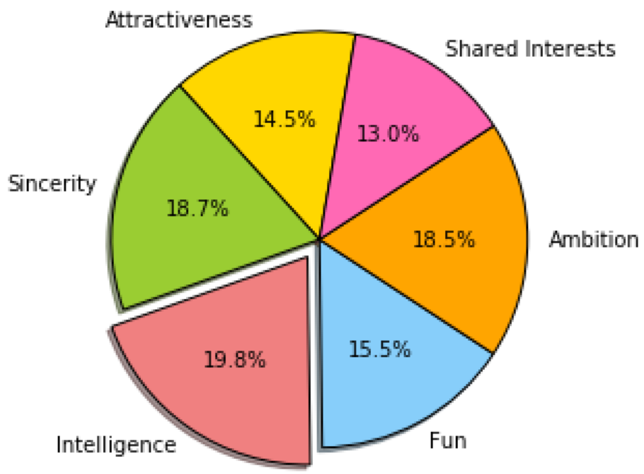
Appendix 5: Pie Chart showing the distribution of average desired attributes for female participants as percentage of the total desired ratings. Code for making the pie chart smaller was obtained from Stack Overflow (2015).



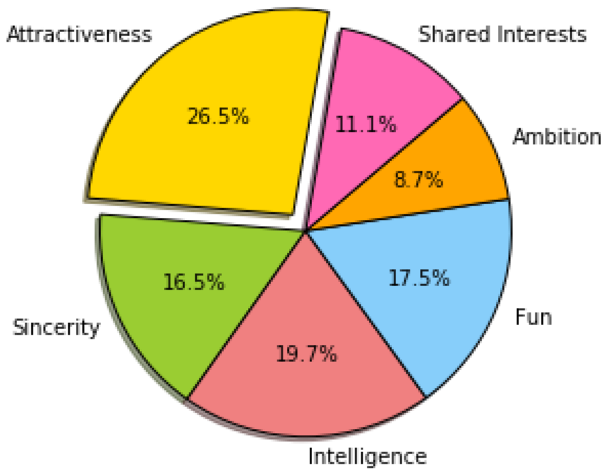
Appendix 6: Pie Chart showing the distribution of the average values of attributes associated with "Yes" response for female participants as a percentage of the total "Yes" ratings. Code for making the pie chart smaller was obtained from Stack Overflow (2015).



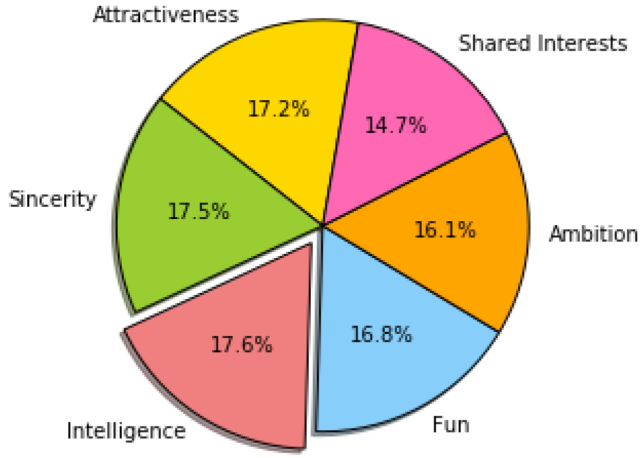
Appendix 7: Pie Chart showing the distribution of the average values of attributes associated with "No" response for female participants as a percentage of the total "No" ratings. Code for making the pie chart smaller was obtained from Stack Overflow (2015).



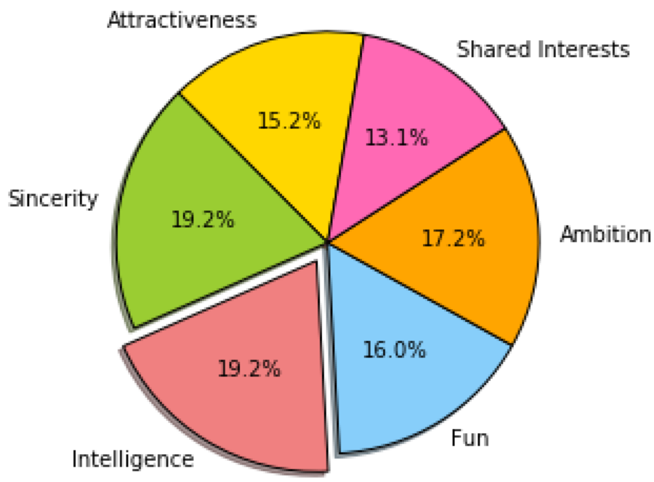
Appendix 8: Pie Chart showing the distribution of average desired attributes for male participants as percentage of the total desired ratings. Code for making the pie chart smaller was obtained from Stack Overflow (2015).



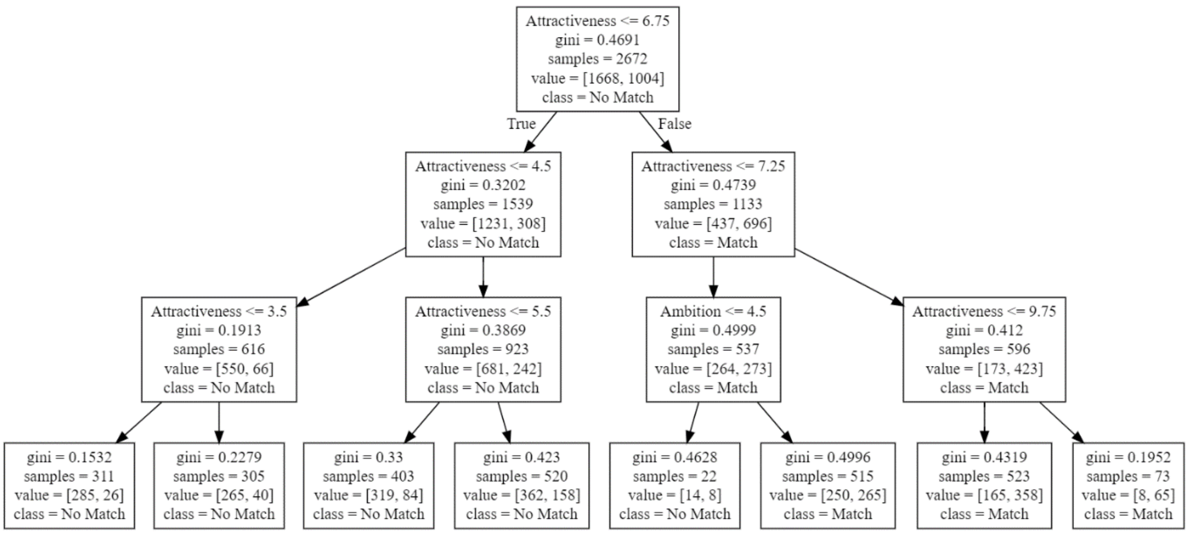
Appendix 9: Pie Chart showing the distribution of the average values of attributes associated with “Yes” response for male participants as a percentage of the total “Yes” ratings. Code for making the pie chart smaller was obtained from Stack Overflow (2015).



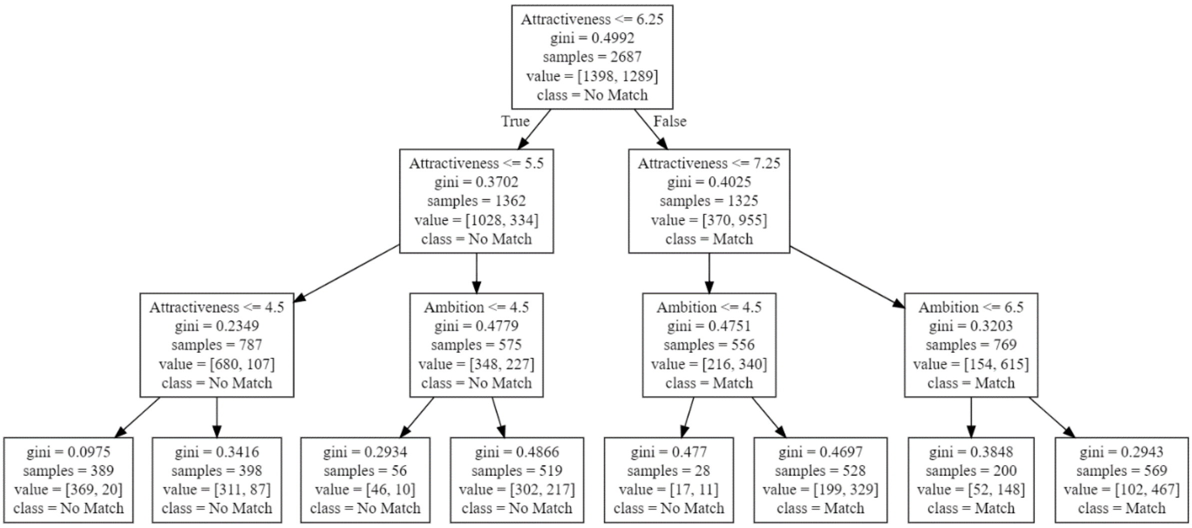
Appendix 10: Pie Chart showing the distribution of the average values of attributes associated with “No” response for male participants as a percentage of the total “No” ratings. Code for making the pie chart smaller was obtained from Stack Overflow (2015).



Appendix 11: Decision Tree for “Attractiveness” and “Ambition” ratings for female participant



Appendix 12: Decision Tree for "Attractiveness" and "Ambition" ratings for male participants



Appendix 13: Classification reports and confusion matrices for male participants (top) and female participants (bottom)

```
             precision    recall  f1-score   support

        0.0       0.75      0.76      0.76       603
        1.0       0.74      0.72      0.73       549

avg / total       0.75      0.75      0.75      1152

[[461 142]
 [151 398]]
             precision    recall  f1-score   support

        0.0       0.81      0.73      0.77       718
        1.0       0.61      0.71      0.66       428

avg / total       0.73      0.72      0.72      1146

[[524 194]
 [125 303]]
```