PREDICTING SPEED DATING

A MACHINE LEARNING APPROACH

Joël van Run

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2057225

THESIS COMMITEE

Dr. P. Hendrix (Supervisor)
Dr. R Alhama (Second reader)

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Sciences &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 13, 2023

WORD COUNT

8712

Table of Contents

Data Source, Code & Ethics Statement

- Work on this thesis did not involve collecting data from human participants or animals.
- The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis.
- The author of this thesis acknowledges that they do not have any legal claim to this data or code.
- The code used in this thesis is not publicly available.

Abstract

The aim of this research is to predict the outcome of a speed date between two individuals using Machine Learning techniques. The used dataset is retrieved from Columbia University (2004). The data contains captured instances of arranged and structured speed dates between students of Columbia University. A priori and a posteriori features are captured before, during and after the speed dates. The objective is discovering which features have most predictive power, which model performs best and whether the usage of Unsupervised Learning can reveal types of people and if they are a good indicator for matchmaking. Additionally, the model performances are compared for both sexes. This thesis uses Logistic Regression, Support Vector Machines, Random Forests, XGBoost and Multi-Layer Perceptron as Supervised Learning algorithms. The Unsupervised Learning algorithms used are K-means Clustering, Hierarchical Clustering and Gaussian Mixture Model. Performance of the models are tested using accuracy, AUC and F1-score. The results show that XGBoost performs best. There are no significant model performance differences between males and females. Partner and participant related preference features yield most predictive power. Finally, although including the clustering technique does cause the predictive performance to drop, it simplifies the model and reduces the high dimensionality of the data, resulting several positive consequences. This thesis differentiates itself from previous research by using and comparing models with higher complexity, looking into model performances for both sexes and adding the additional clustering step to the data science pipeline.

Chapter 1. Problem Statement & Research Goal

1.1 Context

The mentality of our society concerning relationships and dating has changed at rapid pace throughout the past decade. The rise of social media and other technological advancements in the field of communication are partially responsible for this. Nowadays, as opposed to earlier times, we have the possibility to scour a tremendous amount of applications on which we are just one click, like or swipe away to get in touch with anyone that piques our interest.

Due to this convenience and the amount of people we see on a daily basis on our newsfeeds or dating apps, we have become spoiled and developed a distorted reality. This reality being one where people make the outside world believe that their everyday lives are close to perfect. Online profiles contain only perfect pictures and highlights of one’s year, almost fooling everyone to believe that they do not have lesser days. Gillath (2016) states that this caused our generation to develop relational disposability. This so-called relational disposability is a term which describes the current generation's tendency to dispose relationships as if they are disposable and easily replaceable objects. The vast amount of attention we receive in our digital world creates the illusion that we have numerous amount of options and we develop unrealistically high expectations of relationships (Gillath, 2016). Why should we compromise with a partner if certain personal traits or looks do not fully meet our desires? After all, we tend to believe that there are many perfect replacements out there.

In our society in which we are becoming increasingly individualistic and in which we tend to forget to look into deeper layers because of the abundance of stimuli around us, it is of utmost importance to make deeper connections with people. A connection that is not just based on the looks of an individual or how many online followers they have, but a connection based on a much wider range of aspects.

Is it truly the case that we have developed a superficial view on dating and are we really only interested in the appearance of the individual? Or do other features still play an equally important role in the chemistry between two individuals? These are questions this thesis aims to answer.

This thesis combines the domains of data science and speed dating to develop a deeper knowledge of various aspects of speed dating. The study focuses on capturing patterns and critical features in speed dating data by using several data science techniques and algorithms. Ultimately, the outcome of a speed date between two individuals is predicted using machine- and deep learning models. These models can help individuals to look beyond the current generation's superficial dating approach and to find them a suitable, qualitative partner. A healthy, fundamentally well-established relationship vastly improves the quality of life of any person and strengthens the unification of our society.

Additionally, the information retrieved from critical feature analysis helps individuals to get a deeper understanding of which topics and aspects of dating are considered to be important when interacting with someone in a romantic setting. This information can be utilized by the individuals to be more successful at speed dating. The usage of clustering techniques exposes certain types of people based on their feature combinations, these clustered types of people form an interesting foundation for future cross sectional work, such as the psychology domain.

Previously, several studies combined data science with speed dating. These studies are discussed in chapter 2. This study aims to fill the methodological- and scientific gaps of prior work such as the lack of deep learning techniques, a comprehensive variety of supervised model comparisons, a between gender model comparison and a feature importance analysis for both a priori and a posteriori features. Most importantly, this thesis introduces a blended usage of supervised- (SL) and unsupervised learning (USL) methods to expose clusters of people based on their features.

1.2 Research Strategy

The main research question this study focuses on is: *To what extent can the outcome of a speed date be predicted using unsupervised and supervised learning techniques?* To develop a framework and consider elements connected to this research question, four sub research questions are established. These questions are elaborated on below.

The first sub research question: *Which features have most predictive power for the outcome of a speed date?* focuses on gaining insight into which features are most important and yield most predictive power. Within this sub research question, both feature importances of the best performing initial model and hybrid model, as explained in the next paragraph, are looked into. The results are compared. The primary reason researching this is that the predictive power of the models is directly related to the information contained within the features. Therefore, gaining insight in which features play a major role is important.

The second sub research question is: *Can clusters and types of people be discovered from the speed dating data and are they a good indicator for matchmaking?* This sub research question introduces a pioneering combination of SL and USL in speed dating. Several feature combinations which are potentially representative of the personality of an individual are used as input for USL models. The result of this are clusters of people grouped together, based on their feature inputs. These clusters are subsequently used as a new feature for the SL models, ultimately simplifying the complex models while maintaining predictive power. The model which includes both USL and SL is hereinafter referred to as the hybrid model. The model which only uses SL is hereinafter referred to as the initial model. The second sub research question is further elaborated on in section 2.3.

The third sub research question: *How does the performance of a set of selected machine learning algorithms compare to the performance of a baseline logistic regression model for the speed dating data?* focuses on the between-model comparison for all initial SL models, giving insight to which SL model yields best predictive performance.
The final sub research question: *How does the performance of machine learning models for the speed dating data compare for females and males?* focuses on how the best performing initial and hybrid model compare for both males and females.

1.3 Findings

Of all SL algorithms used to predict the outcome of a speed date between two individuals, XGBoost proves to perform best. XGBoost yields an accuracy of approximately 0.94. F1 and AUC evaluation metrics are slightly higher than the performance of the random forest and multi-layer perceptron algorithms. Including clustering techniques as additional step in the data science pipeline allows the models to be simplified and to reduce the high dimensionality of the data, while still maintaining acceptable predictive power. Whether this additional step is desirable, depends on the problem at hand. When the aim of the study lies on maximizing predictive performance, one should not include this step. Contrary, if the purpose lies on the interpretability, simplicity and reduced time or computational power are at hand, one might tolerate the drop in predictive performance and consider the model which includes the clustering.

For both of the final models, the most significant features are relatively equivalent. The a priori partner and participant preference related features such as shared interests, intelligence and attractiveness yield most predictive power. Regarding the a posteriori features, an individual who is considered to be attractive, fun and has similar interests as their dating partner has a higher chance to land a match. Concerning the model performances for males and females, no notable differences were captured.

Chapter 2. Literature Review

This chapter discusses relevant literature on speed dating and data science. Section 2.1 discusses prior work on the Columbia University speed dating dataset. Sections 2.2 and 2.3 discuss other relevant work, namely the mentality and gender differences concerning dating behavior and types of people.

2.1 Prior Work Speed Dating

Prior to this study, other studies used the Columbia University speed dating dataset, a dataset which contains many captured instances of arranged and structured speed dates between students of the university. The dataset contains a numerous amount of a priori and a posteriori features captured before, during and after the speed dates. The dataset is further elaborated on in section 4.1. These prior studies also performed several machine learning algorithms on the data. The processes and outcomes of the studies are briefly elaborated below.

The first study is solely focused on the a priori features of the dataset. The main research question focuses on whether RF or SVM (with either linear, polynomial or radial basis function kernels) performs best. The author uses a limited set of features; therefore part of the information is omitted. After the hyperparameter tuning and selection of feature combinations using the max AUC metric, the SVM RBF Kernel yielded an AUC value of 0.66, the Random Forest (RF) model performed best with an AUC value of 0.71 (Los, 2017). The author states that dimensionality reduction did not improve the performance of any of the two models.

Concerning the features, the attributes age and race are the most important for prediction. Concludingly, Los (2017) states that overall, the more we know about an individual, the better we can predict their intentions and actions. Considering this, the question arises which features have most predictive power for a more extended set of features in comparison to the study by Los (2017). As the author omitted many features, there is a detrimental loss of information, which is a shortcoming of the study. As a result, sub research question 1: "*Which features have most predictive power for the outcome of a speed date?*" is established.

The second study by Yunfan (2019) is focused on predicting the attractiveness of male participants using several machine learning algorithms. The first machine learning algorithm that the author uses is logistic regression. Due to the multiple R-square of 34.6% and a much lower adjusted R-square of 27% the author states that there are too many variables that cause an overfit and the model should be improved.
The second and third model that are used in the study of Yunfan (2019) are decision trees and random forests, which hardly improve performance as the accuracy on the test set are respectively 0.633 for the decision tree and 0.658 for the RF model. Finally, the author uses the XGBoost algorithm which performs significantly better than the other models, yielding an AUC performance of 0.702 on the test set (Yunfan, 2019). The inclusion of a more complex model like the XGBoost algorithm concludes to be a valuable addition to the methodology. A striking aspect however, is the fact that the performance of the RF and XGboost models of Yunfan (2019) are both lower than the RF model performance by Los (2017). As Los (2019) omitted many features, the models by Yunfan (2019) are expected to outperform the models of Los (2017), as valuable information is lost by omitting them. This indicates that there is a potential methodological or experimental setup error in the study by Yunfan (2019).

The third study which uses the Columbia University speed dating dataset, is a study that focuses on ethicality. Although the paper includes another machine learning technique, which was not mentioned before – the Gaussian Naive Bayes algorithm, this technique is not discussed further into detail or taken into consideration due to its high bias component. In the study, preferential fairness, which calls into question the ethicality of recommendations generated by machine learning algorithms (Paraschakis & Bengt, 2020) is the common theme.

The study raises the point of attention that religious and racial bias are very common phenomena when using machine learning methods to make predictions, and this should be taken into consideration as it might be discriminating to some extent.

The study of Eastwick et al. (2022) states that partner preference effects are insignificant. This statement is contradicted and refuted by the outcomes of the aforementioned studies by Yunfan (2019 and Los (2017), where the preference matching effects turned out to have a significant predictive power and proved to explain a fair portion of the variance and its R-squared metric.

Although previous studies used a range of machine learning techniques to predict romantic interactions and the outcomes of (speed) dating, a clear comparison between previously used models on the speed dating dataset has yet to be made. The study of Yunfan (2019) uses logistic regression as a baseline, and decision trees and RF are compared to the baseline model. The study by Los (2017) makes a comparison between SVM and RF and focuses solely on the a priori attributes of the dataset, omitting many possibly valuable attributes. Both studies conclude that the RF models outperform the decision tree models and thus, decision trees will not be used in this study. Additionally, none of the previous studies used models with a significant higher complexity such as the multi-layer perceptron artificial neural network deep learning algorithm, which is likely to perform well on the dataset due to its capability of capturing complex interactions. Besides this, using the model allows for a qualitative basis of comparison in contrast to the more simplistic models and answering the question whether increasing complexity does in fact increase the predictive power of a model. Consequentially, sub research question 3: "*How does the performance of a set of selected machine learning algorithms compare to the performance of a baseline logistic regression model for the speed dating data?*" is established.

2.2 Mentality & Behavior Of The Sexes

As generally known, men and woman have a different mentality concerning dating, love and romantic interactions. This difference can largely be explained by the influence which the sex hormones and primal instincts of the sexes have on the behavior of an individual (Anders, Steiger, & Goldey, 2015). According to the study of Anders, Steiger & Goldey (2015), elevated levels of testosterone do not just increase masculinity, dominance and aggressiveness, but also the competitiveness of a male. Contrary, elevated estrogen levels promote maternal and risk averse behavior (Bridges, 2014). Additionally, the study of Anders, Steiger & Goldey (2015) states that competitiveness in females is selectively discouraged via gender norms and the societal mentality towards desired sex-related behavior in a country greatly stimulates the physiologic behavior.

The question that arises from the aforementioned is whether these differences in mentality and behavior for the sexes have an actual significant effect on the final predictive performance of the models. Although the study of Yunfan (2019) specifically focuses on predicting the attractiveness of male participants, no prior study made model performance comparisons for both sexes thus far. As a result, sub research question 4: "*How does the performance of machine learning models for the speed dating data compare for females and males?*" is established.

2.3 Types of People

When two individuals discuss the topic of dating, a question which frequently arises is which type of person the individual is most attracted to. Although I have heard people speak about many different types and all sorts of appellations, I began to wonder whether this simplifying human romantic interaction and dating approach is justified. This very question was addressed in as study of Seidman, a professor of psychology at Albright College. The study of Seidman (2019) states that this simplification of human romantic interaction is indeed justified and the much-discussed types are an actual existing phenomenon. Seidman (2019) states that in particular the personal traits and characteristics of an individual's exes show many similarities and although we tend to repel the idea of the great resemblance due to negative emotions connected to the former relationship, it is in fact the case that we repeatedly tend to fall for the same type of person.

Although the current literature on speed dating unfortunately lacks any usage of types of people, the dataset in question does have many features which could potentially be used to reveal types of people. To achieve this, several USL techniques can be applied to cluster individuals which are alike concerning several combinations of features. These resulting clusters have the potential capability of exposing interesting and possibly predictive inferences when using the newly created clusters as a predictor. By using the clusters as a new predictor, the number of predictors can be decreased, simplifying the models and decreasing the required computational power and time needed to run the models, while still maintaining the information and the predictive power that lies within the omitted and transformed predictors. As a result, sub research question 2: *To what extend can clusters and types of people be discovered from the speed dating data and are they a good indicator for matchmaking?* is established.

Chapter 3. Methodology

This chapter describes the methodology of this study. The methodological steps and their justification are further explained into detail in the referred sections. First, the accompanying data science workflow which visually describes the methodology and experimental setup is presented to give a clear overview of the data science pipeline.

3.1 Data science Workflow

Figure 1 shows a visual representation of the data science workflow this study uses. The first step of the process is the data cleaning and preprocessing phase. After the data is cleaned and preprocessed, the data is split into a training and test set. Then, the data is rebalanced and normalized, as doing this before the split leads to data leakage.

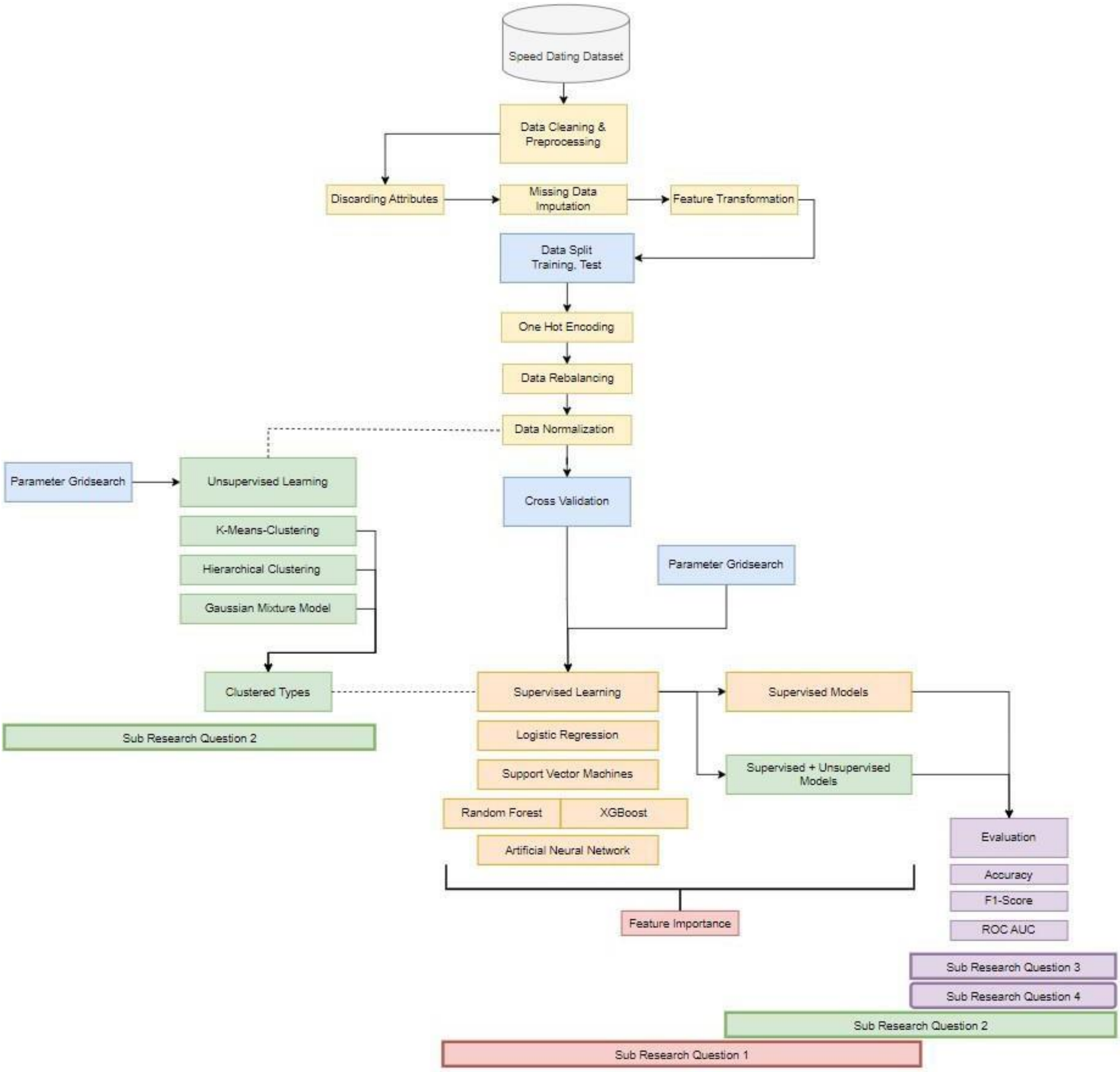


Figure 1 - The data science workflow.

At this point, the split dataset is used as input for all three USL methods, ultimately yielding several clusters as outcome. The clustering is based on three combinations of replaced initial features which are potentially representative of the type a person is, this is further elaborated on in sections 2.3 and 3.3.2.

After the newly formed clusters are extracted, the initial data is cloned and altered by replacing the aforementioned feature combinations with the clusters. As a result, 3x3 new datasets (K-Means, Hierarchical, Gaussian * 3 feature combinations) are formed.

The SL algorithms are then run on the initial dataset, using gridsearch as parameter optimization and the cross-validation technique to extract all available and representative information from the dataset. The overall results are evaluated using several evaluation metrics, answering sub research question 3. Additionally, sub research question 4 is answered by repeating the process and evaluating the model performances for males and females. Sub research question 1 is answered by looking into the feature importance per model.

Finally, the best performing SL model of the previous step is run using the 3x3 altered datasets. As a result, 3x3 models are trained for the best performing SL method. These results are evaluated and compared with the initial SL method using the same evaluation metrics as the previous step, providing an extensive answer to sub research question 2.

3.2 Data mining

In order to extract the desired and valuable information from the data, various data mining methods are performed on the initial dataset. The first step is discarding the attributes which contain unnecessary information for this study. The second and third step in the data mining process are missing data imputation using median imputation and feature transformation. The initial features are transformed to improve the interpretability of clusters, which is described in section 4.3.2 and section 4.3.5. Furthermore, the data is split into a training and test set. The categorical features are one hot encoded; the data is rebalanced using SMOTE and normalized to achieve better predictions as is described in sections 4.3.3 and 4.3.4.

3.3 Machine Learning

After performing the data mining methods, the resulting dataset is used as an input for several machine learning algorithms. These algorithms can be divided into two subcategories, the SL algorithms and the USL algorithms.

3.3.1 Supervised Learning

As stated and justified in section 2.3, this study uses logistic regression as baseline model. RF, XGBoost, SVM and Multi-Layer Perceptron are included as additional and comparative models. The application and usage of the SL models is further elaborated on in section 4.3.5. The SL models are not explained or discussed into detail as they are considered to be known.

3.3.2 Unsupervised Learning

Clustering is based on three combinations of replaced initial features which are potentially representative of the type a person is. The first model only includes a replacement of the interests and hobbies features, the second and third model include an additional replacement and transformation of more features in addition to the former model. Table 1 shows a phased replacement of the features.

	Phase I	Phase II	Phase III	Total Replaced
Model 1	Interests, Hobbies (17)	-	-	17
Model 2	Interests, Hobbies (17)	Confidence, Date, Go Out (3)	-	20
Model 3	Interests, Hobbies (17)	Confidence, Date, Go Out (3)	Field of study, Future Career (2)	22

Table 1 - Three unsupervised learning clustering feature replacement combinations.

Each of the three clustering models include all three feature replacement models. Additionally, the data is scaled for the Euclidian distance algorithm to fulfil its purpose and to ensure there is no unbalanced data within the hyperspace. All models are trained on the training data and applied on the test set.

K-Means Clustering

The first USL model used is K-Means clustering. The K-Means cluster partitions all the observations of a dataset into k clusters, where k is the hyperparameter of the algorithm. Each of the clusters has a cluster centroid and the datapoints are assigned to a certain cluster which has the nearest mean cluster centroid (Nvidia, 2022). The simplicity, flexibility and generalizability of K-Means clustering technique creates a good all round clustering model while still converging relatively fast and requiring a low amount of computational power.

Hierarchical Clustering

The second USL used is the agglomerative hierarchical clustering model. Each of the observations starts with their own cluster, but as a cluster moves up the hierarchy, the clusters merge in a greedy manner (Nvidia, 2022). Just like K-means clustering and Gaussian Mixture Model, hierarchical clustering uses the Euclidean distance measure. Although hierarchical clustering generally takes longer to converge, it often yields better results in comparison to the simpler K-Means algorithm when the data becomes more extensive.

Gaussian Mixture Model

The third USL used to cluster types of people is the Gaussian Mixture Model (GMM). The GMM is a probabilistic model which assumes all datapoints are generated from a mixture of a finite number of Gaussian distributions with unknown parameters (Scikit-Learn, 2022). The major difference between the former two models and the GMM is that GMM uses soft assigning whereat the former two use hard assignment. Hard assignment means that the algorithm at that point in time is certain about the data point's cluster membership, whereat the soft assignment is probability based (Brilliant.org, 2022). A datapoint clustered by the GMM algorithm might for instance have a 60% likelihood to belong to cluster I, 20% to cluster II and 20% to the remaining clusters, the algorithm takes care of the uncertainty of this assignment. Generally, although the GMM algorithm usually takes significantly longer to converge in contrast to the hierarchical algorithm, it does potentially yield better results in some cases.

The justification of the USL algorithms are also discussed in section 2.3. Additionally, section 4.3.5 describes the USL algorithms application and experimental setup in detail.

Chapter 4 Experimental Setup

4.1 Description of the dataset

The dataset that this research uses is publicly available and provided by Columbia University, New York (2002). The dataset contains 4189 speed dates, each taking 4 minutes and all participants were Columbia University students. According to the meta data description written by the authors of the research Fisman & Lyenger (2002) "The entirety of dates took place over a series of 21 waves between 2002-2004 and the students were matched with someone of the opposite sex. At the end of each date, both individuals were asked if they want to meet again, in case of reciprocal liking, a 'match' was registered and the contact details were exchanged. Before attending, they filled in a pre-registration questionnaire to state their demographics, self-perceived traits, and preferences. In particular, attendees could express how important it was for them to date people of their own race or religion". In total the dataset consists of 8378 observations and 195 columns, making it an extensive and interesting dataset to discover interactions between variables (Fisman & Lyenger, 2002).

4.2 Descriptive statistics / Exploratory data analysis

In order to acquire a clear understanding of the dataset and to identify how variables relate to the target variable, exploratory data analysis is conducted.

4.2.1 Goal & Intentions

Participants of the speed dating study were asked about their intentions their participation. Although the division of males and females concerning the classes (1) fun night out, (2) to say I did it, and (3) other, show a relatively equal distribution, there are some notable differences concerning the other three classes. The first and most striking result is the difference in the (To get a date) class whereat males scored 10.35%, over double the female score of 4.71%. Concerning the (Meet New People) class, males scored 33.02% versus a score of 40.77% for females and (Looking for a relationship) saw a score of 4.10% for males versus 3.08% for females. This endorses the results concluded by the studies of Anders, Steiger & Goldey (2015) and Bridges (2014), whereat it is stated that men are more competitive, dominant and straight to the point in contrast to the more maternal and reticent females. Appendix 1 shows a visual representation of the goal division amongst the sexes.

4.2.2 Decision & Match

Finally, regarding the decision and resulting match attributes, the positive class of the overall decision rate is represented in 42% of all instances, a relatively balanced outcome. However, when it comes down to the outcomes of males versus females, there are some striking differences. The positive class of male participants scored 47.4% of instances, the females scored 37%. According to these statistics, it is safe to conclude that females of the speed dating study are far pickier than males. The match attribute is a combination of all instances whereat the participant as well as their partner both labelled the date with a positive class, and thus they would like to go on another date with each other. Concerning the match attribute, a total of 16% of the dates are labelled with the positive class. A visual representation of the decision and match attributes are shown in figure 2. An overview of the features and their description can be found in appendix 3.

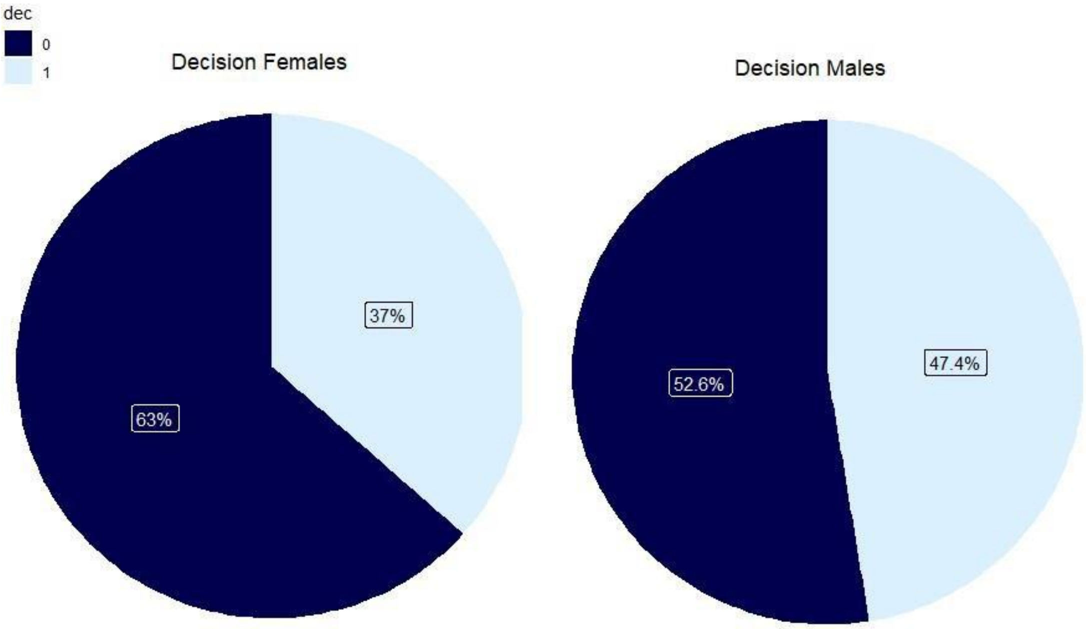


Figure 2 - Visualization of the decision & match attributes for both males and females.

4.3 Cleaning/Preprocessing

4.3.1 Missing data treatment

The speed dating dataset contains a relatively low number of missing values. Although the a priori focused study of Yunfan (2019) omitted all rows of the speed dating dataset that contained missing values, this does result in a loss of potentially valuable information. Therefore, in this study, all rows of the initial dataset are kept, and the missing values are imputed to maintain the maximum informative value.

To deal with the missing values in the dataset, median imputation is applied. The primary reason for this is because the study by Eastwick et al. (2022) performed median imputation and several other missing data techniques, such as multiple imputation, on the speed dating dataset. However, on average these techniques produced a difference in delta r-squared (ΔR_2) of just .0007 in comparison to the more simplistic median imputation method. The difference in performance is negligible and multiple imputation technique is relatively computationally expensive, this study uses the more simplistic median imputation technique. Additionally, irrelevant attributes are discarded from the dataset.

4.3.2 Feature Transformation

One of the initial features are transformed into a new feature. This feature concerns the confidence feature which is based upon the total amount of matches a participant expects to get after all dates. Originally the feature score would be a numeric value somewhere between zero and twenty. However, a large proportion of participants filled in extreme values, either 0 or 20, making it harder to extract valuable information about what the actual value represents. Therefore, the scores are ranked and divided into 3 equal groups - not confident (0-33.3%), neutral (33.3%-66.6%) and confident (66.6%-100%).This transformation allows for easier interpretation of the confidence feature, which as a result makes the clustering done with USL less complex while still maintaining its informative value.

4.3.3 Data Rebalancing

As discussed and shown in the decision and match section of paragraph 4.2, the positive class of the target variable (match) is underrepresented, having a representation of just 16%. In order to deal with this class imbalance, the SMOTE data rebalancing method is used. SMOTE - synthetic minority oversampling technique is a method which generates synthetic/artificial samples for the underrepresented class. The k parameter of the SMOTE algorithm has been set to 7, which is in line with common practice.

4.3.4 Data Normalization & One hot Encoding

To produce better models and allowing the model to more easily learn and understand the problem, the data is scaled. Additionally, one hot encoding is used to transform the categorical features into a form which improves the resolution of the data, which is important for the SVM and MLP models in particular, allowing the models to perform better (Ravi, 2019).

4.3.5 Description of the experimental procedure

The processed and cleaned dataset is split into two chunks, a training chunk containing 70% of the data and a test chunk containing 30% of the data. Instead of including a validation set to set and tune hyperparameters, this study uses K-Fold Cross Validation. The primary reason for making use of this technique is because it allows a more efficient use of the dataset due to the fact that all instances are used for both training and validation purposes which ultimately increases the generalizability of the results and reduces the chance of a sampling bias.

Parameter tuning - Supervised Learning

To find the best hyperparameters for the models used within this study, grid search is used for most SL methods. The primary reason for choosing grid search over randomized search is the fact that the dataset is a relatively small, thus all possible parameter settings can be tested without requiring much computational power and time.

Due to the high complexity of the MLP model, using grid search to find the optimal hyperparameters results in a computationally expensive process which can take up to several days. Therefore, grid search is uncommon when applying deep learning models. Resulting in the testing of several hyperparameter combinations. The hyperparameter tuning results and optimal settings of SL are presented in appendices 5, 6, 7 and 8.

Parameter tuning - Unsupervised Learning

To find the optimal hyperparameter settings for the K-Means Algorithm, the weighted sum statistic (WSS) and silhouette coefficient are used. Although the WSS steadily keeps on dropping when increasing the amount of clusters beyond eight, doing so objects the very philosophy behind the inclusion of the clustering through USL: namely, to cluster types of people which will simplify the predictive models while still preserving as much predictive power as possible. Thus, the silhouette coefficient values serve as the directive. As shown in appendix 4, a value of K = 8 performs best taking the aforementioned into account.

The hyperparameters of the Hierarchical Clustering and GMM are not tuned due to the fact that the hyperparameters such as the amount of clusters need to have exactly the same value as the baseline model. Having less or more clusters in comparison to the baseline model will cause the models to be incomparable as the information is distributed dissimilar.

4.4 Evaluation metrics

To extensively evaluate the performance of the models and the results in comparison to other studies, this study includes multiple evaluation metrics.

The first metric used in this study is the accuracy metric. Accuracy focuses on measuring how many observations are correctly classified, either positive or negative. Most prior studies include accuracy, ensuring that a good basis of comparison between the different studies can be made concerning the model performances. Additionally, the accuracy metric is a valid and useful metric when the data is properly balanced.

Secondly, the F1 score metric will also be used as an evaluation metric. The F1 metric combines precision and recall into a single metric by taking the harmonic mean of the two. The primary reason for using the F1 score as an evaluation metric is the fact that the threshold of the model can be optimized by changing the F beta score (Neptune AI, 2022).

Finally, the evaluation is bolstered by using the Receiver Operating Characteristic Area Under the Curve (ROC AUC). The ROC AUC is a metric that is most useful when the dataset is relatively balanced. Additionally, the positive- as well as negative target variable of this study are equally important as both yield valuable information for training the models. When this is the case, the ROC AUC metric proves to perform well (Neptune AI, 2022).

4.5 Software & Algorithms

All practicalities of this study are executed in Rstudio, using the programming language R version 4.2.2 (RStudio Release Notes, n.d.). The packages used are mostly BaseR, Ggplot2, Caret and Dplyr. Some other packages used in a few cases are: Smote family, Plyr, pROC and all model specific packages.

Chapter 5. Results

This chapter discusses the results of each of the sub research questions.

5.1 Performance Supervised Learning

Section 5.1 addresses sub research question 3: *How does the performance of a set of selected machine learning algorithms compare to the performance of a baseline logistic regression model for the speed dating data?*

This section discusses the results of each of the machine learning techniques used on the speed dating dataset. After training each model on the trainingset, the newly created models are tested on the testset and a performance comparison is made between the predictions and the reality. The trainset performances can be found in appendix 9. For several models, grid search is applied to find the best hyperparameters which yield the best performance, the results of the gridsearch are reported per model in appendices 5, 6, 7 and 8. To analyze the overall performance of the models, several evaluation metrics are used and presented in Table 2. The reported accuracy metric includes the accuracy mean per fold and the standard deviation per fold. Additionally, the corresponding confusion matrices per model can be found in appendix 10.

Model	Accuracy	F1-Score	AUC
Logistic Regression	0.802	0.806	0.809
Support Vector Machines	0.915	0.921	0.923
Random Forest	0.928	0.927	0.928
XGboost	0.939	0.934	0.934
Multi Layer Perceptron	0.912	0.909	0.912

Table 2 shows the supervised learning model test set performances of Logistic regression, Support Vector Machines, Random Forest, XGboost and Multi Layer Perceptron. Performance metrics provided are: Accuracy, F1-Score and AUC.

As shown in Table 2 all included SL models show a significant performance increase of at least 10% across all evaluation metrics in contrast to the baseline logistic regression model's performance of approximately 80%. The best performing model is the XGboost model, yielding a 13.7% higher accuracy performance, a 12.8% higher F1 performance and a 12.5% higher AUC performance in comparison to the baseline model. The low difference in evaluation metrics affirms the robustness of the models.

The more complex SL models appear to perform better at prediction for the speed dating dataset, which is likely caused by their ability to capture more complex relationships and interactions within the data. This hypothesis is also generally confirmed by the results of the grid search, which shows that increasing the complexity of the hyperparameter combinations did in fact increase the performance significantly. The side note is that at a certain point, the model becomes too complex and the performance on the testset starts dropping, as can be seen with XGboost.

The low values of the standard deviation per cross validation fold indicate that the models are robust and they perform well without any major predictive anomalies occurring. Additionally, the low difference between the performances on the training and testset indicate that the models yield proper generalizable predictive results.

5.2 Performance Unsupervised Learning

Section 5.2 addresses sub research question 2: *To what extent can clusters and types of people be discovered from the speed dating data and are they a good indicator for matchmaking?*

This section discusses the results of the best performing SL method of section 5.1 combined with the clusters created by the K-means clustering, Hierarchical clustering and GMM. Although the clustering approach is applied for every SL model and feature combination, only the best performing SL, the XGboost model, is included in this section to avoid an improper extensiveness of the results section.

As described in the methodology and experimental setup section, the clustering is based on three combinations of replaced initial features which are potentially representative of the type a person is, the feature combination table can be found in appendix 11, a visualization of the clustering is shown in figure 3. The performances for each of the clustering and feature model combinations are presented in appendix 12. Feature model 2 is chosen as best performing combination model due to various reasons, which are further elaborated on in section 6.1.2.

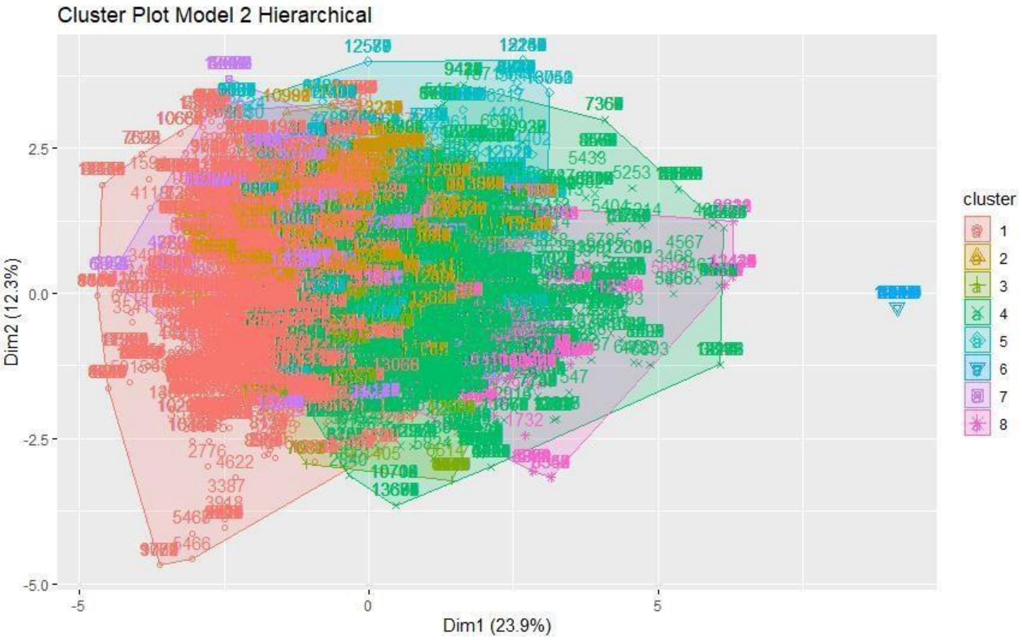


Figure 3 shows the clustering visualization of feature model 2 combined with the XGBoost and Hierarchical model.

Table 3 shows the performances of all three clustering techniques combined with feature model 2. The K-means model is the worst performing model in comparison to its superior Hierarchical Clustering Model and GMM. Concerning the latter two, the evaluation metrics performances are nearly identical, the GMM performs slightly better on accuracy metric (0.726) in comparison to the Hierarchical Model (0.725), whereas the Hierarchical Model performs marginally better on the F1 metric (0.729) and on the AUC metric (0.727), whereas the GMM scores (0.728) on the F1 metric and (0.726) on the AUC metric.

Unsupervised Model	**Feature Model**	**Accuracy**	**F1-Score**	**AUC**
K-Means	2	0.710	0.727	0.709
Hierarchical Clustering	**2**	**0.725**	**0.729**	**0.727**
Gaussian Mixture Model	2	0.726	0.728	0.726

Table 3 shows the unsupervised learning test set performances of the K-means, Hierarchical clustering and Gaussian Mixture Models. Performance metrics provided are: Accuray, F1-Score and AUC.

Although the GMM has a slightly better performance on the accuracy metric in contrast to the Hierarchical Model, the standard deviation per fold metric of the GMM is considerably higher at 0.022 in comparison to the Hierarchical Model standard deviation per fold of 0.016. This lower standard deviation per fold performance indicates that the hierarchical model is more robust, and is therefore considered to be the better model of the two.

Final Model	**Accuracy**	**F1-Score**	**AUC**	**Features Reduced**
Initial Model - XGB	0.939	0.934	0.934	-
Hybrid Model - XGB + Hierarchical	0.725	0.729	0.727	20

Table 4 shows the test set performance comparison between the initial model (XGB) and the hybrid model (XGB + Hierarchical Clustering). Provided are the Accuracy, F1-score and AUC performance metrics.

To make a clear overall comparison between the initial XGBoost model and the hybrid XGboost model, the performances and specifications of the models are shown in table 4.

In total, 20 features are reduced to 1 cluster feature. The total accuracy dropped by 0.214 to 0.725 the F1 score dropped 0.205 to 0.729. The AUC dropped 0.207 to 0.727. The standard deviation per fold increased by 0.009 to 0.016 which indicates a drop in the robustness of the model.

5.3 Features and their predictive power

Section 5.3 addresses sub research question 1: *Which features have most predictive power for the outcome of a speed date?*

This section discusses which features of the final two models have most predictive power. The final two models are the initial model and the hybrid model as described in section 5.2. The features shown and discussed in this chapter are further elaborated on in the feature legend which can be found in appendix 3.

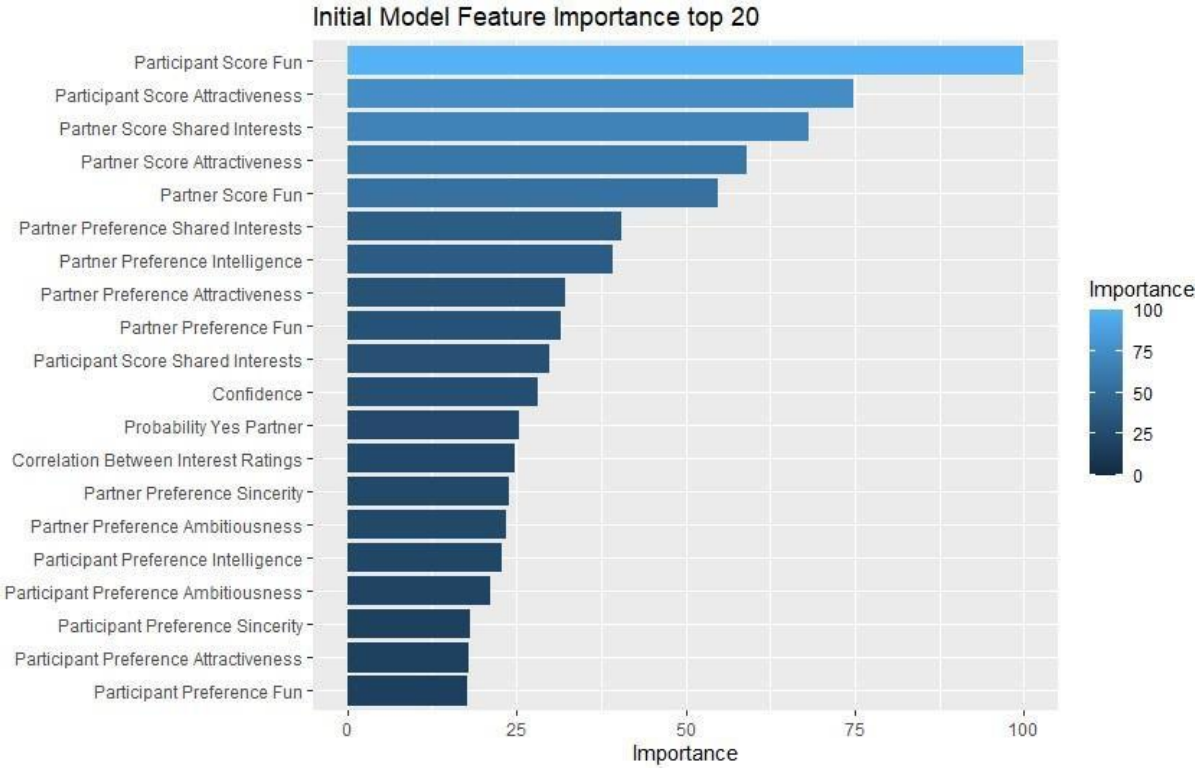


Figure 4 shows the visualization of the feature importance for the initial model.

The predictive power and the importance of the top 20 features of the initial model is visualized and shown in figure 4, the variable importance per group of the top 20 for the complete model is shown in appendix 13.

For the initial model, the partner preference related features have most predictive power at 23.84% of which shared interests, intelligence and attractiveness are most important.

The rating of partner features comes in second place at 23.46%, of which shared interests, attractiveness and fun are the most important. Concerning the rating of participant, fun, attractiveness and shared interests have the most predictive power.

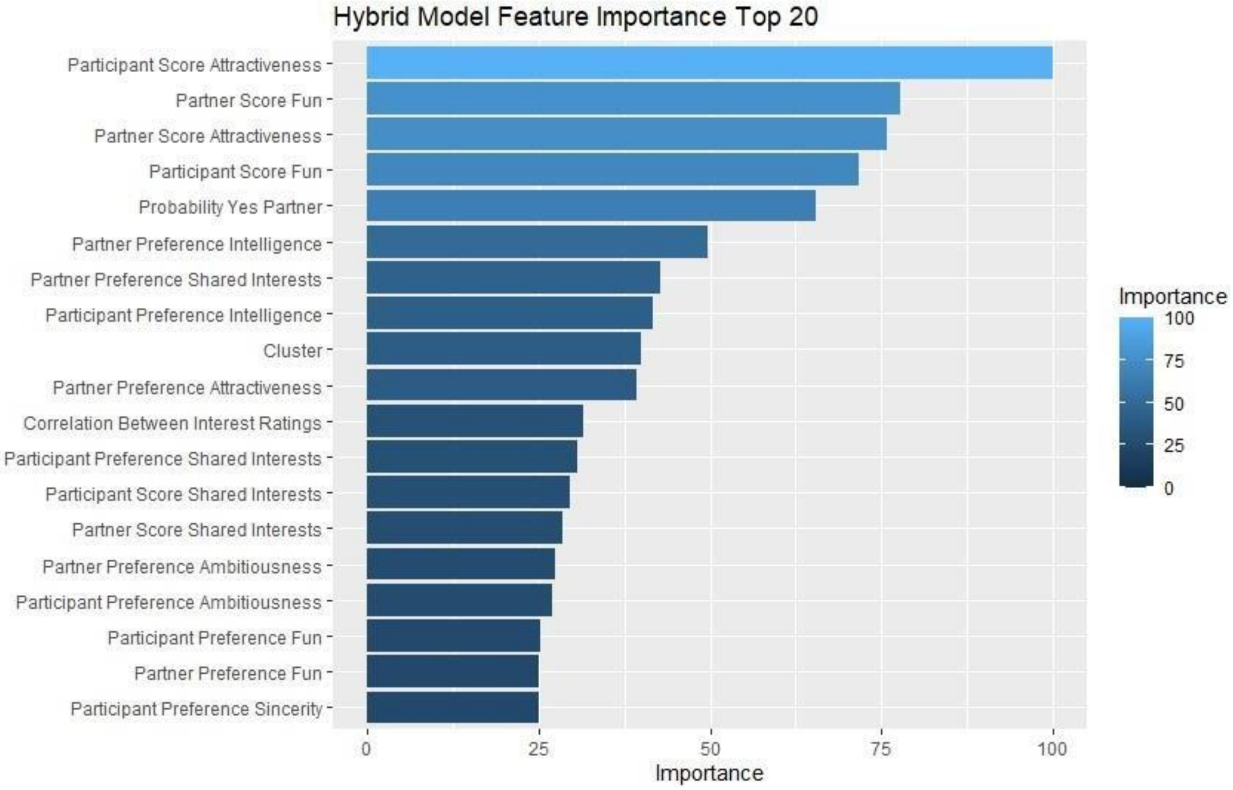


Figure 5 shows the visualization of the feature importance for the hybrid model.

Figure 5 shows a visualization of the top 20 most important features of the hybrid model. Although the majority of the most important features is the same for both models as shown in the figures, the replacement of several features by the clustering feature caused a slight shift in non replaced feature importance values as well. The newly formed clustering feature consisting of the features discussed in section 5.2 has a mediocre feature importance, scoring higher than most a priori features but lower than several a posteriori features such as the participant and partner scores concerning several personal characteristics such as intelligence, attractiveness and the extent to which the partner and participant have common interests. As seen in Figure 4, the confidence feature of the initial model solely holds mediocre predictive power as its importance value is 28, which is around 2.8%. The cluster feature, consisting of the confidence feature and several other features, has an importance score of 40 as is shown in figure 5.

Appendix 13 shows the feature importance per group of the 20 most important features. The feature importance results and in particular the effect of the inclusion of the clustering technique is further discussed in section 6.1.1.

5.4 Performance Gender Comparison

Section 5.4 addresses sub research question 4: *How does the performance of machine learning models for the speed dating data compare for females and males?*

In this section, the results per gender for the model performance of the initial XGboost model, which is discussed in section 5.1, and the performance of the hybrid model, which is discussed in section 5.2, are further looked in to in order to answer sub research question 4. Table 5 shows the performances of the models per gender. The reported accuracy metric includes the accuracy mean per fold and the standard deviation per fold.

Model	Gender	Accuracy	F1-Score	AUC
Initial - XGB	Male	0.939	0.937	0.940
Initial - XGB	Female	0.938	0.937	0.938
Hybrid - XGB + Hierarchical	Male	0.726	0.729	0.727
Hybrid - XGB + Hierarchical	Female	0.725	0.728	0.726

Table 5 shows the test set model performances of the initial and hybrid models for both males and females. Provided are the accuracy, F1-score and AUC metrics.

The performance metrics of the initial model for males and females are nearly identical and show no significant differences. The accuracy score is slightly higher for males at 0.939, the F1 score is identical at 0.937 and the AUC score marginally higher at 0.940.

Concerning the hybrid model, the performances are nearly identical as well, whereas the males scoring slightly higher on the accuracy, F1 and AUC metrics in comparison to females at 0.726, 0,729 and 0.727 respectively. The standard deviations per fold for both sexes remained the same at 0.007 for the initial model and 0.016 for the hybrid model.

Confusion Matrix Initial M			Confusion Matrix Initial F		
	Reference			Reference	
Predicted	No	Yes	Predicted	No	Yes
No	993	74	No	953	79
Yes	53	994	Yes	48	975
Confusion Matrix Hybrid M			**Confusion Matrix Hybrid F**		
	Reference			Reference	
Predicted	No	Yes	Predicted	No	Yes
No	776	291	No	754	282
Yes	288	759	Yes	281	738

Table 6 shows the Confusion Matrices per model for both males and females.

Regarding the confusion matrix comparison between the initial model and the hybrid for both males and females, there are no notable anomalies. Although the correctly classified negative (true negatives) and positive classes (true positives) show a significant drop, the drop is relatively comparable and the increase of false positives and false negatives is relatively comparable as well. The confusion matrix comparison is shown in Table 6. Altogether it can be stated that the hybrid model has its predictive shortcomings in comparison to the initial model, the error patterns are qualitatively similar.

Chapter 6. Discussion

This chapter evaluates the results and puts them into perspective with existing literature. Additionally, the societal and scientific relevance of this study, as well as the limitations and future directions are debated. To maintain consistency and to improve readability, this discussion section is organized according to the sub research questions.

6.1 Discussion per research question

6.1.1 Features and their predictive power - Sub Research Question 1

The feature importances of the initial model as well as the hybrid model show many resemblances. For both models, the a priori partner and participant preference related features yield most predictive power, contrary to what is concluded in earlier studies. The results of this thesis for instance contradict and refute the hypothesis of Eastwick et al. (2017), as they state that ideal partner preference-matching effects were extremely small with typically no different from zero. The a priori feature importances of this study are however in line with the feature related results of the study by Los (2017) as well as both the a priori and a posteriori feature importances of the study by Yunfan (2019).

6.1.2 Performance Unsupervised Learning - Sub Research Question 2

With regard to the clustering results, there are several interesting, yet debatable findings. First of all, the initial model has an AUC performance of approximately 0.94, whereas the final hybrid model has a performance of 0.73. This shows a drastic drop in performance. On the other hand, the high dimensionality and complexity of the initial dataset is drastically reduced. This reduced complexity allows the model to require less time and computational power to run, as well as increasing the interpretability.

As discussed in section 3.3.2, three feature combination models are created of which the second model is chosen to construct the final hybrid model. The primary reason for choosing this model over the other two combinations is the fact that it allows the simplicity of the model and predictive performance to be in harmony. Twenty of the features are replaced by the cluster. This drastically simplifies the model, while still maintaining a reasonable predictive performance.

Whether the usage of the hybrid model and including the clustering technique is expedient, depends on the intentions and goals of the study. A study which aims to achieve the highest predictive performance should not include this technique. Whereas a study at which the focus lies on the simplicity and usability of the model might prefer the usage of the additional technique if the drop in performance is considered to be reasonable and worthwhile.

I consider the drop in predictive performance in exchange for more simplicity, interpretability and reduction of computational time too high for the three feature replacement combinations. Adding the additional step in the data science workflow does allow for interesting combinations and results. Although the currently used feature combinations cause a severe drop in performance, it might be the case that other feature combinations vastly reduce the complexity and dimensionality of the model while hardly causing any drop in performance. These combinations could be tested in future research as well as the usage of other clustering techniques such as fuzzy clustering.

Additionally, the information and patterns which lie within the current types of people clusters created by the model can be further explored and studied by other scientific departments such as the psychology department.

6.1.3 Performance Supervised Learning- Sub Research Question 3

In order to make a clear performance comparison with the existing literature, this section uses the AUC metric as comparative metric due to the fact that it is included in the prior literature. As discussed in the results section, the XGBoost model yields the best performance in this study. For the other studies which lack the usage of the XGBoost model, the best performing model is the RF model, this result is not surprising as XGBoost is considered to be an upgrade of the RF model. The study by Los (2017) and Eastwick et al. (2022) both have a RF performance of 0.63 and 0,71 respectively. The study by Yunfan (2019) does include a XGBoost model which has a performance of approximately 0.70.

The model performances of the other studies however differ quite a lot in comparison to the performance of my initial XGboost model which yields an AUC of roughly 0.94. This enormous difference in performance is caused by the fact that this study includes more features, both a priori and a posteriori, in comparison to the other studies, whereas the study by Los (2017) for instance only includes a priori features. The hypothesis of Los: *"Generally, the more we know about each person, the better we can predict their actions."* is confirmed by this study.

As explained in section 5.1, increasingly complex models tend to have better predictive power, however, when the model becomes too complex, as is the case for several hyperparameter combinations of XGboost, the model starts to overfit and the testing performance drops. Although I initially expected the MLP model to outperform the XGBoost model, this was not the case with the used hyperparameters. The final MLP model took several hours to run and due to limited time and computational power, I did not get the opportunity to try out higher amounts of hidden units per layer. As the performance of the final MLP model is relatively close to the performance of the XGBoost model and there is room for improvement concerning its hyperparameters, I believe it is plausible to hypothesize that the MLP model is likely to outperform the XGBoost model when it has the right hyperparameters. In order to test this hypothesis, future studies could consider using Python instead of R due to the fact that it generally speeds up the process for deep learning problems.

6.1.4 Performance Gender Comparison - Sub Research Question 4

Regarding the performance of the initial and hybrid model for each of the sexes there are no notable differences as all of the evaluation metrics are nearly identical. As stated in section 2.2, males and females have a different mentality concerning dating, love and relational interactions, which is largely explained by the influence of hormones, gender norms and societal mentality towards desired sexrelated behavior. As a result, one would expect a larger difference in predictive power for the models, which evidently is not the case and is partially contradicted by this study. To further examine this contradiction and look into differences between males and females, future research could be conducted by researching feature importance of both sexes to discover gender related behavioral patterns.

6.2 Societal & Scientific Relevance

As discussed in section 1.1, the societal relevance of this study lies in the fact that this study aims to create a thorough basis of knowledge to understand various aspects of speed dating. Speed dating has become an important aspect of our lives due to various reasons such as the aforementioned relational disposability and several technological advancements. This study for instance explains which personal traits are considered to be important and highly desirable.

This study has two interesting additions to the speed dating and relational interaction scientific field. The first addition is using USL combined with SL. This unique combination allows complex models which contain relatively high dimensional data to be simplified. The consequence of this simplification, however, is that the predictive performance of the model drops. The model becomes less computationally expensive, converges more easily and is more straightforward to interpret. This is especially of importance for cross sectional research such as the psychological scientific research field. The following addition is that results in regard of gender related predictive performance for each model lays a groundwork for refuting statements made by previous studies. These previous studies noted that there are huge differences regarding men and woman in terms of dating and having a relational mindset. The results of this research suggest that these differences may be considerably less than people might think. To explore this even further, feature importances for both sexes can be investigated during future research. Also, the significantly higher predictive performance of the initial model relative to the models in other studies indicate that the inclusion of a posteriori features are important in predicting the outcome of speed dating. Therefore, it should also be incorporated into the model if it were possible. There is a limitation to the generalizability of the results. The dataset is limited to heterosexually and cis oriented individuals, the results are not fully generalizable to the entire population and should certainly be taken into when interpreting this study or using it as a foundation for future work.

Chapter 7. Conclusion

Recapitulatory, this study has contributed to several aspects of the relational interaction and speed dating research field. By using SL data science techniques, various patterns and critical features within the data are discovered. For the hybrid model as well as the initial model, it turns out that the a priori partner and participant preference related features are most significant. Regarding the a posteriori features, a participant which is considered to be fun, attractive and has similar interests as their speed dating partner is more likely to score a match. A remarkable pattern which is exposed is the fact that the predictive performances of both models do not significantly differ between the sexes.

The initial XGBoost model proves to be the best performing model, yielding a predictive performance of almost 0.94 on the AUC metric. Additionally, to deal with the high dimensionality of the data, to increase interpretability for future research and limit computational expenses, this best performing predictive SL model has been modified. After testing several USL learning technique combinations with the XGboost model, the Hierarchical clustering and XGboost algorithm combination proves to be the best final hybrid model. Although the performance does show a significant AUC drop from 0.94 to 0.73, it is an interesting methodological combination. The usage of the additional clustering technique in future related research might be beneficial dependent on the goal of the user, whereat a tradeoff between simplicity, functionality and predictive power has to be considered. After all, we need creative ways to simplify and deal with the ever-increasing complexity of our data driven world.

Chapter 8. References

Bibliography

Anders, S. M., Steiger, J., & Goldey, K. L. (2015). Effects of gendered behavior on testosterone in women and men. *Pnas Vol. 112, No. 45*, 6.

Bridges, R. S. (2014). Neuroendocrine Regulation of Maternal Behavior. *PMC Pubmed Central*, 15.

Brilliant.org. (2022). *Gaussian Mixture Model*. Retrieved from Brilliant : https://brilliant.org/wiki/gaussian-mixture-model/

Data Aspirant. (2020). *Unsupervised Learning Algorithms*. Retrieved from Data Aspirant: https://dataaspirant.com/unsupervised-learning-algorithms/

Diaz, F., Metzler, D., & Amer-Yahia, S. (2010). Relevance and ranking in online dating systems. *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval.* Geneva: DBLP.

Eastwick, P. W., Joel, S., Molden, D. C., Blozis, S., Carswell, K. L., & Finkel, E. J. (2022). Predicting romantic interest during early relationship development: A preregistered investegation using machine learning. *European Journal of Personality*, 37.

Eduminatti. (2020, December 12). *Social media sets unrealistic standards in society*. Retrieved from Eduminatti Official : https://eduminattiofficial.medium.com/social-media-set-unrealisticstandards-in-society-eb5a6dc599c5

Fisman, R., & Lyenger, S. (2002). *Gender differences in mate selection: Evidence from a speed dating Experiment.* New York, USA.

Gillath, O. L. (2016). *Generalization disposability: Residential mobility and the willingness to dissolve social ties.*

IBM. (n.d.). *Learn - Unsupervised learning*. Retrieved from IBM Cloud: https://www.ibm.com/cloud/learn/unsupervised-learning

Institute for Clinical Evaluative Sciences. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes.

Los, A. (2017). *Modelling an individual's selection of a partner in a speed-dating experiment using a priori knowledge.* Stockholm: KTH Royal Institute Of Technology .

Neptune AI. (2022, July 21). *Blog - F1 Score-Accuracy-ROC-AUC-PR-AUC*. Retrieved from Neptune AI: https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc

Nvidia. (2022). *Datascience-XGBoost*. Retrieved from Nvidia.com: https://www.nvidia.com/enus/glossary/data-science/xgboost/

Nvidia. (2022). *K-Means Clustering Algorithm*. Retrieved from Nvidia: https://www.nvidia.com/enus/glossary/data-science/k-means/

Nvidia. (2022). *Unsupervised Learning*. Retrieved from Nvidia: https://www.ibm.com/cloud/learn/unsupervised-learning

Paraschakis, D., & Bengt, N. J. (2020). Matchmaking under fairness constraints: A speed dating case study. *International Workshop on Algorithmic Bias in Search and Recommendation @ ECIR'20.* Malmö.

Ravi, R. (2019). *One Hot encoding*. Retrieved from TowardsDataScience: https://towardsdatascience.com/one-hot-encoding-is-making-your-tree-based-ensemblesworse-heres-why-d64b282b5769

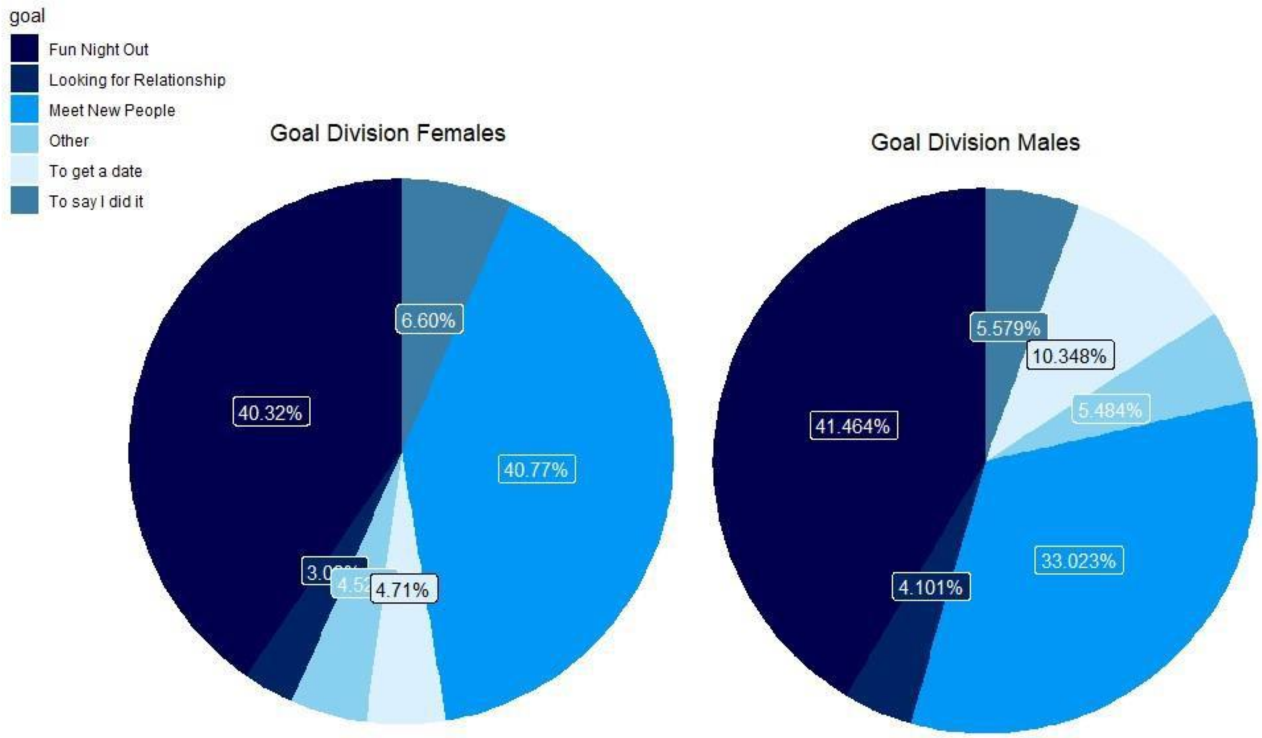
Scikit-Learn. (2022). *Gaussian Model*. Retrieved from Scikit-Learn: https://scikitlearn.org/stable/modules/mixture.html#:~:text=A%20Gaussian%20mixture%20 model%20is, Gaussian%20distributions%20with%20unknown%20parameters.

Seidman. (2019, July 3). *Do People Have A Type*. Retrieved from Psychology Today: https://www.psychologytoday.com/us/blog/close-encounters/201907/do-people-reallyhave-dating-type

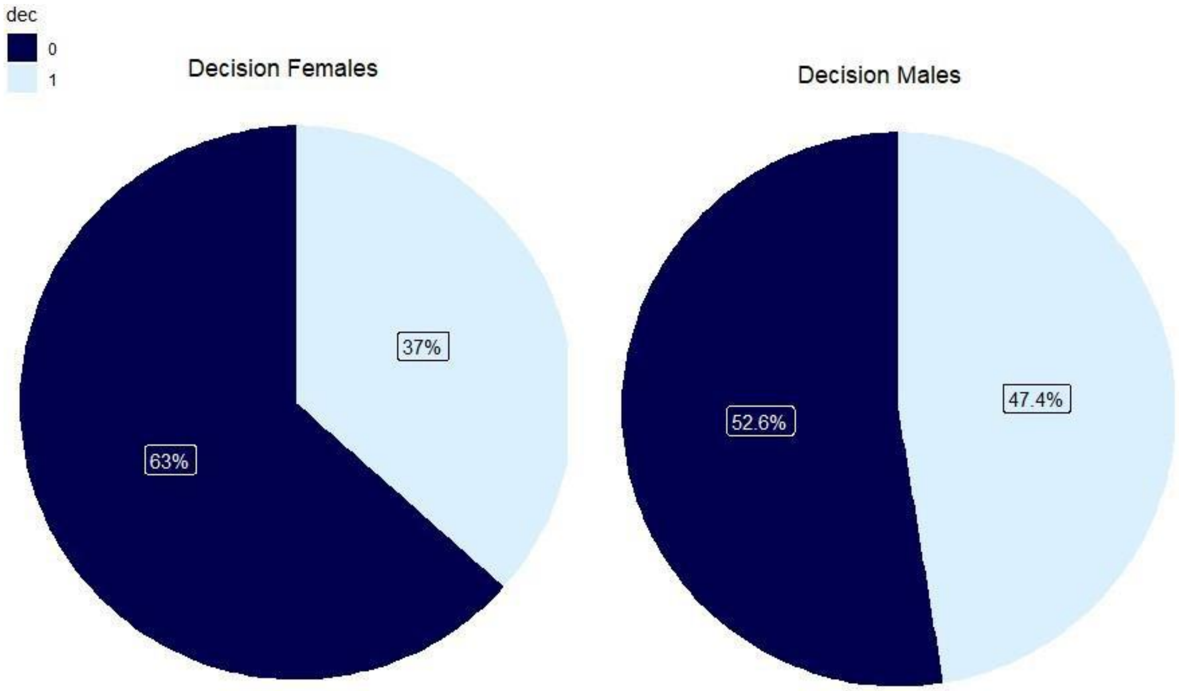
Yunfan, S. (2019). *The secret of love in speed dating.* Los Angeles: University of California.

Chapter 9. Appendices

Appendix 1 - Chapter 4.2 - Goals & Intentions



Appendix 2 - Chapter 4.2.2 - Decision & Match



Appendix 3 - General - Feature Legend

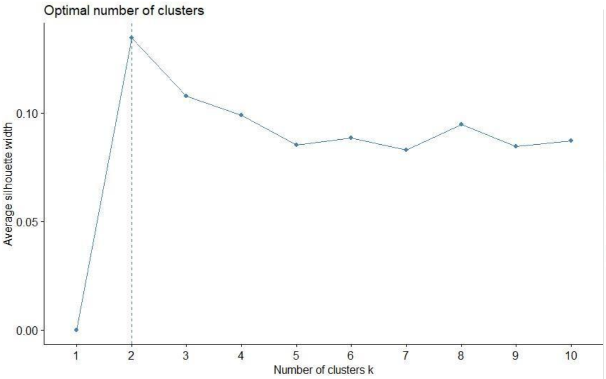
Code	**Meaning**	**Values Range**
	Transformed Feature	
	No missing values - no imputation	
	Missing values - median imputed	
Overall		
gender	Gender	Female (0), Male (1)
Outcome		
match	If both parties voted "yes" on the dec and dec_o variables then match = 1. Otherwise match = 0	**Match (1), no match (0)**
dec	Decision participant	Yes (1), No (0)
dec_o	Decision partner	Yes (1), No (0)
A priori		
age_o	Age partner	Numeric
age	Age participant	Numeric
race	Race participant	Black/African American=1, European/CaucasianAmerican=2, Latino/Hispanic American=3, Asian/Pacific Islander/Asian-American=4, Native American = 5, Other =6
Race_o	Race partner	Black/African American=1, European/CaucasianAmerican=2, Latino/Hispanic American=3, Asian/Pacific Islander/Asian-American=4, Native American = 5, Other =6
Samerace	Same race	Yes (1), No (0)
Field_cd	Field of study Coded	Yes (1), No (0)
Career_c	Career Coded	1 = Lawyer 2= Academic/Research 3 = Psychologist 4 = Doctor/Medicine 5 = Engineer 6 = Creative Arts/Entertainment 7= Banking Finance/Marketing/Business/CEO/Entrepreneur/Admin 8 = Real Estate 9 = International/Humanitarian Affairs 10= Undecided 11 = Social Work 12 = Speech Pathology 13 = Politics 14 = Pro sports/Athletics15 = Other 16 = Journalism 17 = Architecture
Pf_o_int	Preference partner for intelligence	Score 100 total
Pf_o_att	Preference partner for attractiveness	Score 100 total
Pf_o_amb	Preference partner for ambitiousness	Score 100 total
Pf_o_sin	Preference partner for sincerity	Score 100 total
Pf_o_fun	Preference partner for fun	Score 100 total
Pf_o_sha	Preference partner for shard interests	Score 100 total
Shar1_1	Preference participant for shared interests	Score 100 total
Amb1_1	Preference participant for ambitiousness	Score 100 total
Fun1_1	Preference participant for Fun	Score 100 total

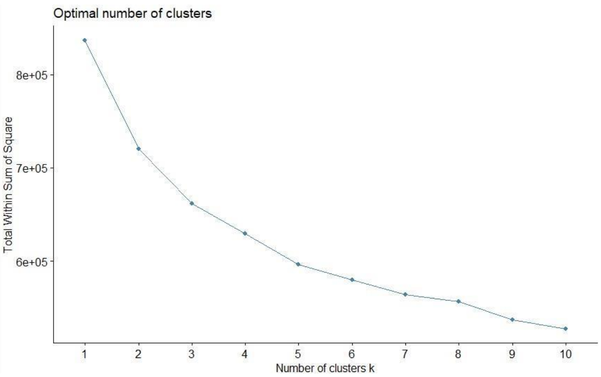
Intel1_1	Preference participant for intelligence	Score 100 total
Attr1_1	Preference participant for attractiveness	Score 100 total
Sinc1_1	Preference participant for sincerity	Score 100 total
Goal	What is your primary goal in participating in this event? - Participant	Seemed like a fun night out=1, To meet new people=2, To get a date=3, Looking for a serious relationship=4, To say I did it=5, Other=6
Date	In general how frequently do you go on dates? - Participant	Several times a week=1, Twice a week=2, Once a week=3, Twice a month=4, Once a month=5, Several times a year=6, Almost never=7
Go_out	How often do you go out (not necessarily on dates)? - Participant	Several times a week=1, Twice a week=2, Once a week=3, Twice a month=4, Once a month=5, Several times a year=6, Almost never=7
sports	Interest in Sports	on a scale of 1-10
Tvsports	Interest in tvsports	on a scale of 1-10
Exercise	Interest in exercise	on a scale of 1-10
Dining	Interest in dining	on a scale of 1-10
Museums	Interest in museums	on a scale of 1-10
Hiking	Interest in hiking	on a scale of 1-10
art	Interest in art	on a scale of 1-10
Gaming	Interest in gaming	on a scale of 1-10
Clubbing	Interest in clubbing	on a scale of 1-10
Reading	Interest in reading	on a scale of 1-10
Tv	Interest in tv	on a scale of 1-10
theater	Interest in theater	on a scale of 1-10
Movies	Interest in movies	on a scale of 1-10
Concerts	Interest in concerts	on a scale of 1-10
Music	Interest in music	on a scale of 1-10
shopping	Interest in shopping	on a scale of 1-10
yoga	Interest in yoga	on a scale of 1-10
Confidence		
Match_es	Out of the 20 people you meet, how many matches do you expect to get?	Not confident (0-33.3%) Neutral (33.3-66.6%) Confident (66.6%-100%)
A posteriori		
Met	Have you met before?	Yes (1), No (0)
Prob_o	How probable do you think your partner will say yes for you?	(1) not likely, (10) extremely likely
Int_corr	Interest Correlation	On a scale of 0-1
Attr	Attractive: What do you look for in opposite sex	Attributes 1 = awful, 10 = great
Sinc	Sincere: What do you look for in opposite sex	Attributes 1 = awful, 10 = great
Intel	intelligence: What do you look for in opposite sex	Attributes 1 = awful, 10 = great
fun	Fun: What do you look for in opposite sex	Attributes 1 = awful, 10 = great
Amb	Ambitiousness: What do you look for in opposite sex	Attributes 1 = awful, 10 = great

Shar	Sharded interests: What do you look for in opposite sex	Attributes 1 = awful, 10 = great
Attr_o	Rating by partner Attractive	Score 100 total
Fun_o	Rating by partner Fun	Score 100 total
Sinc_o	Rating by partner Sincerity	Score 100 total
Intel_o	Rating by partner Intelligence	Score 100 total
Amb_o	Rating by partner Ambitiousness	Score 100 total
Shar_o	Rating by partner Shared Interests	Score 100 total

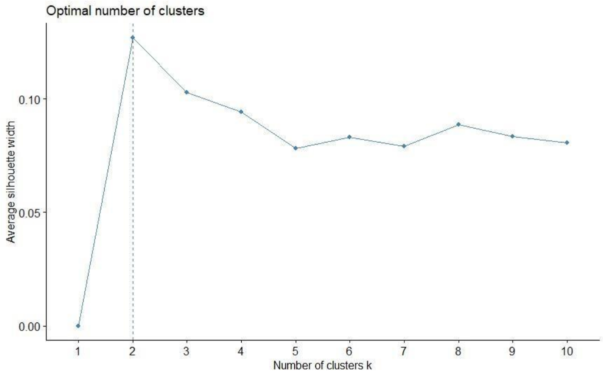
Appendix 4 - Chapter 4.3.6 - Hyperparameter tuning K-Means

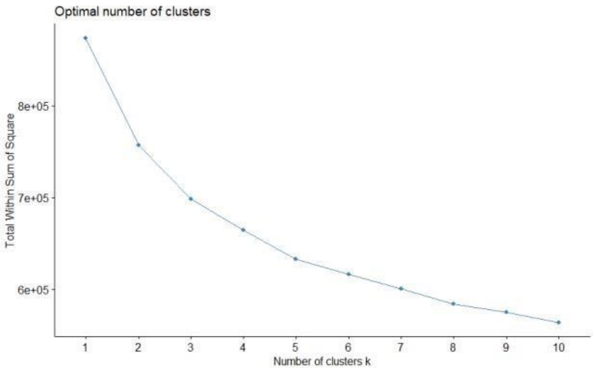
Model 1 WSS, Silhouette Coefficient:



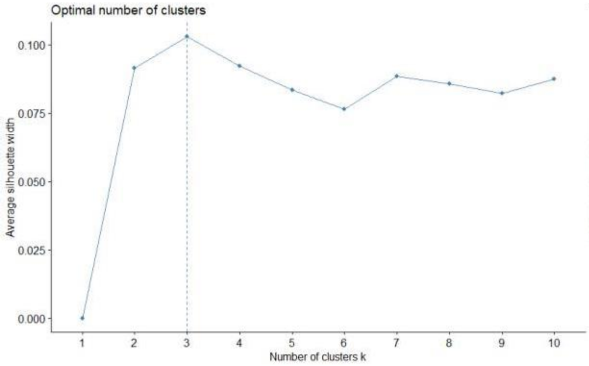


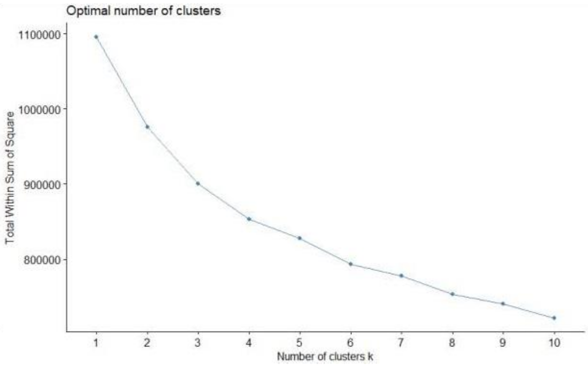
Model 2 WSS, Silhouette Coefficient:





Model 3 WSS, Silhouette Coefficient:





Appendix 5 - Chapter 5.1 - Gridsearch for SVM

Degree	Scale	C	AUC	F1
1	0.001	0.25	0.877	0.795
1	0.001	0.50	0.878	0.798
1	0.001	0.75	0.882	0.800
1	0.001	1.00	0.883	0.802
1	0.010	0.25	0.881	0.801
1	0.010	0.50	0.884	0.801
1	0.010	0.75	0.884	0.802
1	0.010	1.00	0.884	0.801
1	0.100	0.25	0.884	0.802
1	0.100	0.50	0.884	0.802
1	0.100	0.75	0.884	0.802
1	0.100	1.00	0.884	0.802
2	0.001	0.25	0.884	0.802
2	0.001	0.50	0.882	0.801
2	0.001	0.75	0.885	0.804
2	0.001	1.00	0.887	0.807
2	0.010	0.25	0.919	0.808
2	0.010	0.50	0.921	0.840
2	0.010	0.75	0.929	0.853
2	0.010	1.00	0.934	0.858
2	0.100	0.25	0.936	0.862
2	0.100	0.50	0.940	0.876
2	0.100	0.75	0.937	0.873
2	0.100	1.00	0.934	0.869
3	0.001	0.25	0.885	0.867
3	0.001	0.50	0.886	0.805
3	0.001	0.75	0.890	0.810
3	0.001	1.00	0.892	0.813
3	0.010	0.25	0.932	0.870
3	0.010	0.50	0.942	0.877
3	0.010	0.75	0.950	0.887
3	0.010	1.00	0.954	0.891
3	0.100	0.25	0.955	0.912
3	0.100	0.50	0.962	0.915
3	0.100	0.75	0.963	0.914
3	**0.100**	**1.00**	**0.964**	**0.915**

Appendix 6 - Chapter 5.1- Gridsearch for Random Forest

Mtry	AUC	F1
2	0.967	0.905
3	**0.974**	**0.921**

Appendix 7 - Chapter 5.1- Gridsearch for Multi Layer Perceptron

Hidden units layer 1	Hidden units layer 2	Hidden units layer 3	AUC	F1
16	32	64	0.872	0.871
32	64	128	0.904	0.896
64	**128**	**256**	**0.912**	**0.909**

Appendix 8 - Chapter 5.1- Gridsearch for XGboost

Max Depth	N Rounds	AUC	F1
3	50	0.938	0.863
3	100	0.953	0.880
3	150	0.959	0.893
3	200	0.963	0.899
3	250	0.965	0.904
3	300	0.967	0.908
3	350	0.968	0.912
3	400	0.969	0.914
3	450	0.969	0.916
3	500	0.969	0.917
5	50	0.961	0.897
5	100	0.970	0.915
5	150	0.971	0.920
5	200	0.974	0.924
5	250	0.976	0.925
5	300	0.977	0.927
5	350	0.976	0.927
5	400	0.978	0.927
5	450	0.976	0.927
5	500	0.975	0.928
7	50	0.977	0.928
7	100	0.978	0.929
7	150	0.978	0.931
7	200	0.979	0.931
7	250	0.979	0.931
7	**300**	**0.979**	**0.932**
7	350	0.979	0.931
7	400	0.978	0.931
7	450	0.977	0.931
7	500	0.977	0.932

Appendix 9 - Chapter 5.1 - Performances on Training sets

Model	Accuracy	F1-Score	AUC
Logistic Regression	0.809	0.807	0.810
Support Vector Machines	1.000	1.000	1.000
Random Forest	0.928	0.926	0.929
XGboost	1.000	1.000	1.000
Multi Layer Perceptron	1.000	1.000	1.000

Appendix 10 - Chapter 5.1- Confusion Matrices Test Sets

Confusion Matrix LR			Confusion Matrix SVM			Confusion Matrix MLP		
	Reference			Reference			Reference	
Predicted	No	Yes	Predicted	No	Yes	Predicted	No	Yes
No	1672	377	No	1872	94	No	1872	134
Yes	427	1693	Yes	227	1976	Yes	167	1901
Confusion Matrix RF			Confusion Matrix XGB					
	Reference			Reference				
Predicted	No	Yes	Predicted	No	Yes			
No	1904	106	No	1946	100			
Yes	195	1964	Yes	153	1970			

Appendix 11 - Chapter 5.2 - Feature Combination Models

	Phase I	Phase II	Phase III	Total Replaced
Model 1	Interests, Hobbies (17)	-	-	17
Model 2	Interests, Hobbies (17)	Confidence, Date, Go Out (3)	-	20
Model 3	Interests, Hobbies (17)	Confidence, Date, Go Out (3)	Field of study, Future Career (2)	22

Appendix 12 - Chapter 5.2 - Performances Hybrid Models XGB

Unsupervised Model	Feature Model	Accuracy	F1-Score	AUC
K-Means	1	0.726 (0.023)	0.728	0.726
K-Means	2	0.710 (0.026)	0.727	0.709
K-Means	3	0.698 (0.024)	0.719	0.697
Hierarchical Clustering	1	0.739 (0.015)	0.748	0.739
Hierarchical Clustering	**2**	**0.725 (0.016)**	**0.729**	**0.727**
Hierarchical Clustering	3	0.707 (0.019)	0.715	0.707
Gaussian Mixture Model	1	0.737 (0.017)	0.727	0.706
Gaussian Mixture Model	2	0.726 (0.022)	0.728	0.726
Gaussian Mixture Model	3	0.710 (0.019)	0.733	0.737

Appendix 13 - Chapter 5.3 - Feature Importance of top 20 features divided per group.

Initial Model:

Feature Group	Feature Importance
Partner Preference	23.84%
Rating of Partner	23.46%
Rating of Participant	18.22%
Participant Preference	9.78%
Other	7.84%

Hybrid Model:

Feature Group	Feature Importance
Partner Preference	18,57%
Rating of Partner	17,89%
Rating of Participant	21,25
Participant Preference	15,23%