# HW1: Speed Dating - What Makes One Tick?

## Submission deadline: November 29, 23:59 CET

**Assignment.** In this assignment, we explore the Speed Dating dataset. You are asked to build several models, each related to different problems/questions. The questions are related to predictive analytics and interpreting the decisions made by the predictive models. It is up to your group to decide which type(s) of model you choose for a particular question, note that some “black-box” models (e.g. random forests, neural networks, ...) may be harder to compare/interpret than more “glass-box” models (e.g. trees, rule based classification, ...). On the other hand, creative solutions that are able to explain the decision logic of a “black-box” model might deserve some extra points.

**The dataset.** You are given a dataset about Speed Dating. This dataset was compiled by Columbia Business School professors Ray Fisman and Sheena Iyengar. Data was gathered from participants in experimental speed dating events from 2002-2004.

During the events, the attendees would have a four minute “first date” with every other participant of the opposite sex. At the end of their four minutes, participants were asked if they would like to see their date again. They were also asked to rate their date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests. The dataset also includes questionnaire data gathered from participants at different points in the process. These fields include: demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate, and lifestyle information.

The dataset can be found at `canvas.tue.nl` under the name `speed_dating_assignment.csv`.

**Important.** Your goal is to build a predictive model that would generalize well for future cases. Think what information would be available at time of casting a prediction. It is your argued choice what features to use.

**Tooling.** You can use your own preferred programming language or machine learning environment for the assignments, e.g. Weka, RapidMiner, scikit-learn (Python) or R.

**Submission.** Submit your code and your report (with explicit answers to the questions 1.1, 1.2, 1.3, 2.1 and 2.2 introduced below) via `canvas.tue.nl`, before **November 29**, 23:59 CET. A grade will be given for the group as a whole. Please include a peer-review paragraph stating the degree to which everyone in the group has contributed to the completion of the assignment. If someone contributed much more or much less than the others, and you think this should be reflected in grading, please state this explicitly as well. The report must be submitted as a **PDF**. If you work with Microsoft Word, please export your report to PDF before submitting.

# 1 Data Exploration and Predictive Analytics

## 1.1 Visualization (20 points)

We start this assignment by exploring the dataset a bit. The task is to construct a matrix and to visualize it. The matrix should have two dimensions. Rows will be the ages for the subject (age), and columns will be the ages for the subject's partner (age as taken from the relevant dataset case). The values should be an interesting statistic that can tell you something about the decision of a subject in a speed date given the age of the subject and the age of the partner. Note that in the dataset, each case – represents a speed date focused on the subject. If you want access to attributes about the partner, you will need to cross-reference it, using pid == iid.

Include the visualization of the matrix, and comment on the relation(s) that you manage to discover.

## 1.2 Separate model per gender (20 points)

Build a model on 80% of the females, and report some performance measures on the predictions of the other 20% of the females, do the same for males. So both models do not use the gender feature. The models should be trained to predict the class attribute dec. Think of an applicable performance metric, and compare to a *reasonable* baseline. Is accuracy the best performance metric to use here?

Having reported performance measures, now examine the models themselves, print the parameters and logic of the models or find a way to visualize the models. Try to speculate about differences among the two models.

You should now do a small experiment as to see the stability of the model. Choose a different 80% of the females and build a third model. Have you ended up with the same model as you got the first time (above) for females? Using the new 20% test cases, compare your performance measures and internals. Reflect whether the two female models are closer to one another than to the male model.

### 1.3 Feature engineering (20 points)

Often, data-scientists, take insight from data exploration to introduce new features to help the models. Consider for example the exploration you have done in assignment 1.1 at the top. Can you think of a new (calculated) feature (or few) that will possibly improve the model. Introduce the new feature(s), build two new models (for females and for males). Repeat the previous steps for reporting performance, and examine the models. Discuss any observations you make w.r.t. the various models that you have (the two new ones compared to each other, a new one compared to the one using only the raw features, etc.).

## 2 Discrimination Awareness

### 2.1 Unisex model (20 points)

Let's assume, it is not legitimate to build separate models per gender. Therefore, you need to build a single model. More than that, the model still must not take into account the gender of neither of the partners. Build a model on 80% of the speed dates, evaluate on the other 20%.

### 2.2 Comparing the unisex model to the per-gender models (20 points)

Find cases that diverge between the unisex model, and the per-gender model (i.e. different prediction of a specific male subject and specific female partner between the males model and the unisex model). Find a way to explain this divergence. You can start your analysis either by inspecting the samples on which the models disagree, or by starting from summaries on the separate models (using similar ideas / visualizations from above or different) and reflecting about cases for which the prediction therefore diverges.