# ME 781: Project Report
# Group 12


**: Members :**

**Amey Gohil** (170100019) | **Prakhar Nama** (203370006)
**Deepak Soman** (174100018) | **Subhadeep Chaudhuri** (19i190010)


## **Project Topic:** ML-based Matchmaking

Autumn, 2020
IIT Bombay

# _____ Table of Contents _____

# Problem Definition Description

## End-User Requirement

In the startup that we propose, the **customer** is the end-user. The customer would register with us with the expectation of ***finding an ideal match for her/ him*** for being in a relationship with.

## Market Survey For Potential Competitors

At present, there exist platforms (for dating and long-term engagements) that provide service based on users who register on the respective platforms. Some of the noteworthy competitors in this sector are being listed below:

- **International origin** (all these predominantly focus on matches from a dating perspective):
    - Bumble
    - Tinder
    - Facebook dating
- **Indian origin** (all these platforms predominantly thrive on long-term engagement suggestions):
    - Shaadi.com
    - Jeevansathi.com
    - Bharatmatrimony.com

## What Makes Us Different?

What makes us your should-be choice over the existing available services is the fact that we DO NOT suggest unfiltered matches based on mere location, age and preferences.

We provide ***curated*** choices for you that we believe would truly make you and your partner happy as a couple, based on ***psychological trait analysis*** using state-of-the-art machine learning algorithms and then, based on your personal preferences (age, location, gender, ethnicity, etc.)

## Our Unique Selling Point (USP) and How Do We Protect It?

The procedure that most matchmaking sites use to suggest matches are based only on age, geographical proximity, and some even based on horoscope (sad but true). But living in the age of science and data, we believe in analyzing existing data on the compatibility of couples, and collection/analysis of our own data. Our unique ML algorithm is based on analyzing the 5 major personality traits of individuals (referred to as The Big 5 traits or OCEAN traits), and finding the most compatible match as predicted by our unique algorithm. Then, one may choose a partner as per personal preferences.

But how do we make sure that our USP indeed remains unique?

- ★ We have a first-mover advantage, and therefore will be able to build a mature database with time.
- ★ Copyrighting the deployment algorithm would make sure nobody fishes away our methodology.
- ★ Patenting our methodology is another option.
- ★ Lastly, you can open up new camps, but you can't buy the brains behind it. ;)

## Barriers In The Market

The barriers faced by any emerging start-up can be categorized in the following subcategories:

- ❑ **Market Entry barrier:** This deals with the barriers that a start-up would face when it first attempts to enter its potential market. The barriers faced in each front are being discussed as follows -
    - *Capital Requirements*: Any startup requires certain funds to set sail. So, capital requirements come at the forefront of barriers to get a startup going.
    - *Strong brand identity*: Existing competitor firms might have established a strong identity in the prevailing market, which impedes the smooth entry of any new venture.
    - *Switching costs*: These are one-time costs the buyer faces when switching from an existing supplier's product to a new entrant's.
    - *Patents and government policies*: Governments can limit or prevent entry to industries with various regulatory controls (for eg, limits to access to raw materials)

- ❑ **Market Share barrier:** Once a start-up gains footing in the market, the next struggle that a company faces is to maintain its market position and not give in to future competitors. The barriers in this aspect are as follows -
    - *Technical knowledge base*
    - *Orthodox customer base*

# Technology Landscape Assessment

## Overview

Dating and matrimonial relationships are an important part of adult life. The choice of a partner is guided by motivations for long-term happiness, stability, and reproductive success. The cost of failure in relationships is large emotionally and financially and common reasons for relationship failure have been listed below.

Table 1.List of Major Reasons for Divorce by Individuals and Couples [1].

| Reason for divorce | Individuals (N =52) | Couples (*N = 36) | Couple Agreement |
|---|---|---|---|
| Lack of commitment | 75.0 | 94.4 | 70.6 |
| Infidelity or extramarital affairs | 59.6 | 88.8 | 31.3 |
| Too much conflict and arguing | 57.7 | 72.2 | 53.8 |
| Getting married too young | 45.1 | 61.1 | 27.3 |
| Financial problems | 36.7 | 55.6 | 50.0 |
| Substance abuse | 34.6 | 50.0 | 33.3 |
| Domestic violence | 23.5 | 27.8 | 40.0 |
| Health problems | 18.2 | 27.8 | 25.0 |
| Lack of support from family | 17.3 | 27.8 | 20.0 |
| Religious differences | 13.3 | 33.3 | 0.0 |
| Little or no premarital education | 13.3 | 22.2 | 25.0 |

Lack of commitment is shown as a major cause. The aggregate of all above listed reasons can be summed up by a single index for relationship satisfaction. Such an index will be useful for simplified statistical analysis. Relationship based decision-making has spurned many technological services. These services can be categorized into two main classes:

(a) Ease of accessibility services e.g Tinder, cupid.
(b) Guidance services e.g Bharatmatrimoy,astrology.com.

The ease of accessibility services aims to bring together a large mix of gender and age groups of people and ensure that they have the means to assess each other and communicate between themselves. The guidance-based services aim to help the user make the choice of partner selection. These are based on the detailed matching of preferences, or on the basis of non-scientific methodologies such as astrology. Our project aims to take a more scientific outlook at the workings of a relationship, by studying relationship satisfaction with respect to personality traits.

## Big five personality trait theory

A common human intuition is that people behave in a characteristic way and thus people can be grouped into categories based on their behavior. In psychological trait theory, the Big Five personality traits, is a way of grouping individuals based on personality traits. This theory was developed in the 1980s by brute statistical investigations. Further cross-cultural studies showed that these classifications are universal. The Big Five personality theory states that human personality can be subdivided into five basic traits. The proportion of these five traits in an individual predicts their typical behavior in various scenarios.

The five broad personality traits are described below:

- Extraversion
- Agreeableness
- Openness
- Conscientiousness
- Neuroticism

### Openness

This trait explores attributes such as imagination and insight. People who score high in this trait have a broad range of interests. They are curious and like to meet new people and learn new things. Those that score low are more traditional and struggle with abstract thinking.

### Conscientiousness

Conscientiousness includes high levels of thoughtfulness, good impulse control, and goal-directed behaviors. Highly conscientious people tend to be organized and mindful of details. They plan ahead, think about how their behavior affects others, and are mindful of deadlines.

### Extraversion

Extraversion is characterized by being excited, sociable, talkative, assertive. People high in extraversion are outgoing. People who are low in extraversion are more reserved and have less energy to expend in social settings. Such people prefer solitude in comparison to social settings.

### Agreeableness

Agreeableness includes factors such as trust, generosity, kindness, affection. People who are high in this trait are cooperative while those low in this trait are competitive and manipulative.

### Neuroticism

This trait is characterized by negative emotions such as depression, moodiness, and emotional instability. People who score high in this neuroticism experience mood swings, anxiety, and depression. People who score low tend to be more stable and mentally resilient.

## Influence of Personality traits on relationship satisfaction

Multiple research studies have been able to link observed social patterns with personality traits, for example, the academic success of a student is closely related to their conscientiousness score while knowledge is related to openness score [2]. Similarly, the observed wage differences between men and women can be explained by differences in agreeableness traits which reflect their ability to negotiate for themselves [3], and the wage differences within gender groups can be explained by differences in extraversion. On similar lines, the health of an individual, the manner of conflict resolution [4], infidelity [5] and tendencies of substance abuse [6] have been predicted using the big five personality traits.

It was thus natural to use the big five personality trait theory and explore the link between relationship satisfaction and personality traits. Many studies have been conducted using statistical methods [7,8], these studies have highlighted the negative role of neuroticism on marital satisfaction. These studies showed couples scoring high on neuroticism were less happy. While other studies [9] were able to show by a meta-analysis of available research that conscientiousness played a positive role in relationships. These studies show that a complex pattern exists between personality traits and relationship satisfaction.

Therefore a more robust mathematical model based on machine learning is suitable to be used on such multivariate dyadic problems. The biggest challenge for machine learning would be collection of dyadic data of couples by answering lengthy psychometric questionnaires. Such psychometric questionnaires are readily available in the open literature for use [10].

# Project Planning

The project planning was done to make sure we were always on track, and the project got completed without any hiccups.

## Task Breakdown

The requirements of the project were broken down into the following tasks and subtasks

| Phase 1 | Phase 2 | Phase 3 |
|---|---|---|
| • Project Introduction<br>• Brainstorming<br>• Literature Survey<br>• Market Survey | • Data Exploration<br>• Data Analysis<br>• Conceptual Design<br>• Algorithm Testing<br>• Algorithm Deployment | • Product/Service Marketing<br>• Product Demonstration<br>• Report Preparation |

## Roles and Responsibilities

For efficient handling and monitoring of each task, one of the team members was assigned the responsibility of overlooking the completion of the respective task, while everybody contributed in proportion. The RASIC chart summarizes the division of labor among the requisite tasks.

| Roles | | Deepak | Prakhar | Subhadeep | Amey |
|---|---|---|---|---|---|
| Phase 1 | Project Introduction | S | S | S | R |
| | Brainstorming | R | S | S | S |
| | Literature Survey | R | S | S | A, I |
| | Market Survey | A | S | S | R |
| Phase 2 | Data Exploration | R | A, I | S | S |
| | Data Analysis | S | R | A, I | S |
| | Conceptual Design | A, I | S | R | S |
| | Algorithm Testing | S | S | R | A, I |
| | Algorithm Deployment | S | A, I | S | R |
| Phase 3 | Product/Service Marketing | S | R | S | A, I |
| | Product Demonstration | A, I | S | R | S |
| | Report Preparation | S | R | A, I | S |

## Project Timeline

To handle and oversee properly whether the milestones are reached on time, the chronology of the tasks and their completion status of the project was tracked continuously. The GANTT chart illustrates the proposed project schedule and the efficiency of the members.

A glimpse of the GANTT chart is being given below:

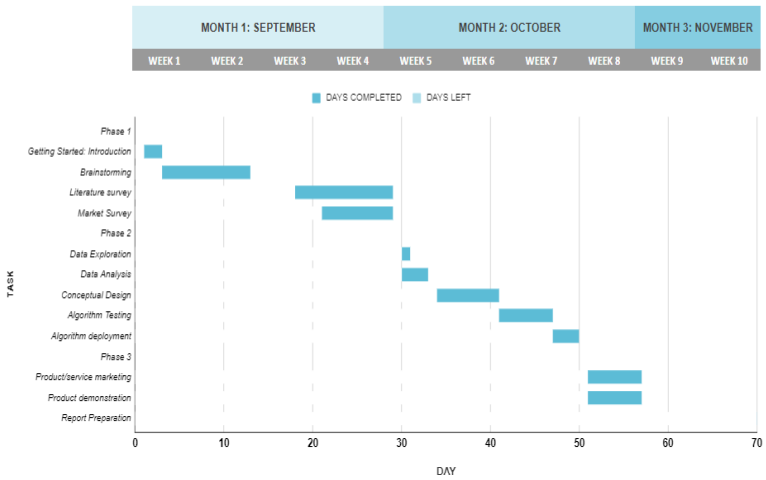
**PROJECT: ME 781 Engineering Data Mining and Applications**

**MEMBERS:** Subhadeep Chaudhuri | Deepak Soman | Prakhar Nama | Amey Gohil

| | |
|---|---|
| PROJECT START DATE: | September 21, 2020 |
| PROJECT END DATE: | December 17, 2020 |
| CONTINGENCY (BUFFER) PERIOD: | 10 days |

| FULL PROJECT STATUS: | |
|---|---|
| % COMPLETED : | 100 |
| PROGRESS BAR : | |

| TASK NAME | START DATE | END DATE | DURATION (IN DAYS) | MEMBER RESPONSIBLE | PERCENT COMPLETED | PROGRESS BAR |
|---|---|---|---|---|---|---|
| **Phase 1** | **22-Sep-2020** | **20-Oct-2020** | **28** | | | |
| Getting Started: Introduction | 22-Sep-2020 | 24-Sep-2020 | 2 | Prakhar | 100 | |
| Brainstorming | 24-Sep-2020 | 4-Oct-2020 | 10 | Deepak | 100 | |
| Literature survey | 9-Oct-2020 | 20-Oct-2020 | 11 | Deepak | 100 | |
| Market Survey | 12-Oct-2020 | 20-Oct-2020 | 8 | Prakhar | 100 | |
| **Phase 2** | **21-Oct-2020** | **10-Nov-2020** | **20** | | | |
| Data Exploration | 21-Oct-2020 | 22-Oct-2020 | 1 | Deepak | 100 | |
| Data Analysis | 21-Oct-2020 | 24-Oct-2020 | 3 | Amey | 100 | |
| Conceptual Design | 25-Oct-2020 | 1-Nov-2020 | 7 | Subhadeep | 100 | |
| Algorithm Testing | 1-Nov-2020 | 7-Nov-2020 | 6 | Subhadeep | 100 | |
| Algorithm deployment | 7-Nov-2020 | 10-Nov-2020 | 3 | Prakhar | 100 | |
| **Phase 3** | **11-Nov-2020** | **13-Dec-2020** | **32** | | | |
| Product/service marketing | 11-Nov-2020 | 17-Nov-2020 | 6 | Amey | 100 | |
| Product demonstration | 11-Nov-2020 | 17-Nov-2020 | 6 | Prakhar | 100 | |
| Report Preparation | 30-Nov-2020 | 13-Dec-2020 | 13 | Amey | 100 | |
| **PROJECT END** | | | | | | |




## Project Monitoring

We monitored the effectiveness of each member by rotating the lead in each week so that no one was overburdened, yet no one could escape his responsibilities. The lead was assigned the task of overlooking every meeting, and to keep track of the project progress in that week.

**Scheduled meetings ( ~ 45mins) and respective agenda:**

- ❏ Every Wednesday - Weekly Target Plan
- ❏ Every Friday - Implementation Progress
- ❏ Every Sunday - Discussion and resolving issues

# Conceptual design document

## ML Model selection

- **Model 1:** Multiple linear regression models with 10 predictor variables and 2 computed measures from the existing variables.
- **Model 2:** Multi-layer Perceptron architecture with 1 hidden layer, 10 input features and one output neuron for the predicted score.

The selected models have comparable accuracies in the predictions and exhibited similar MSE.

**Note:** Accuracy of Model 2 might get better with the availability of a large dataset for model training and validation. This can be a possible direction for future work.

## ML Models verified

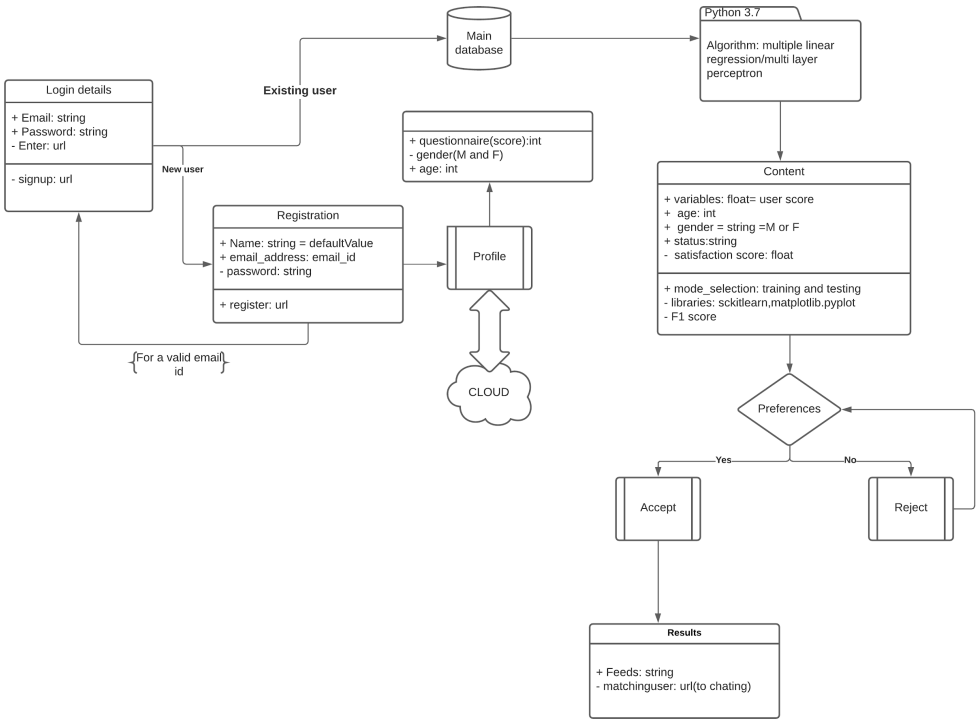
- Polynomial regression model for degree 2 and 3.

## ML Dataset selection

- Dataset of marital satisfaction vs personality traits (of both male and female), along with details on age, social status considered for our formulation.

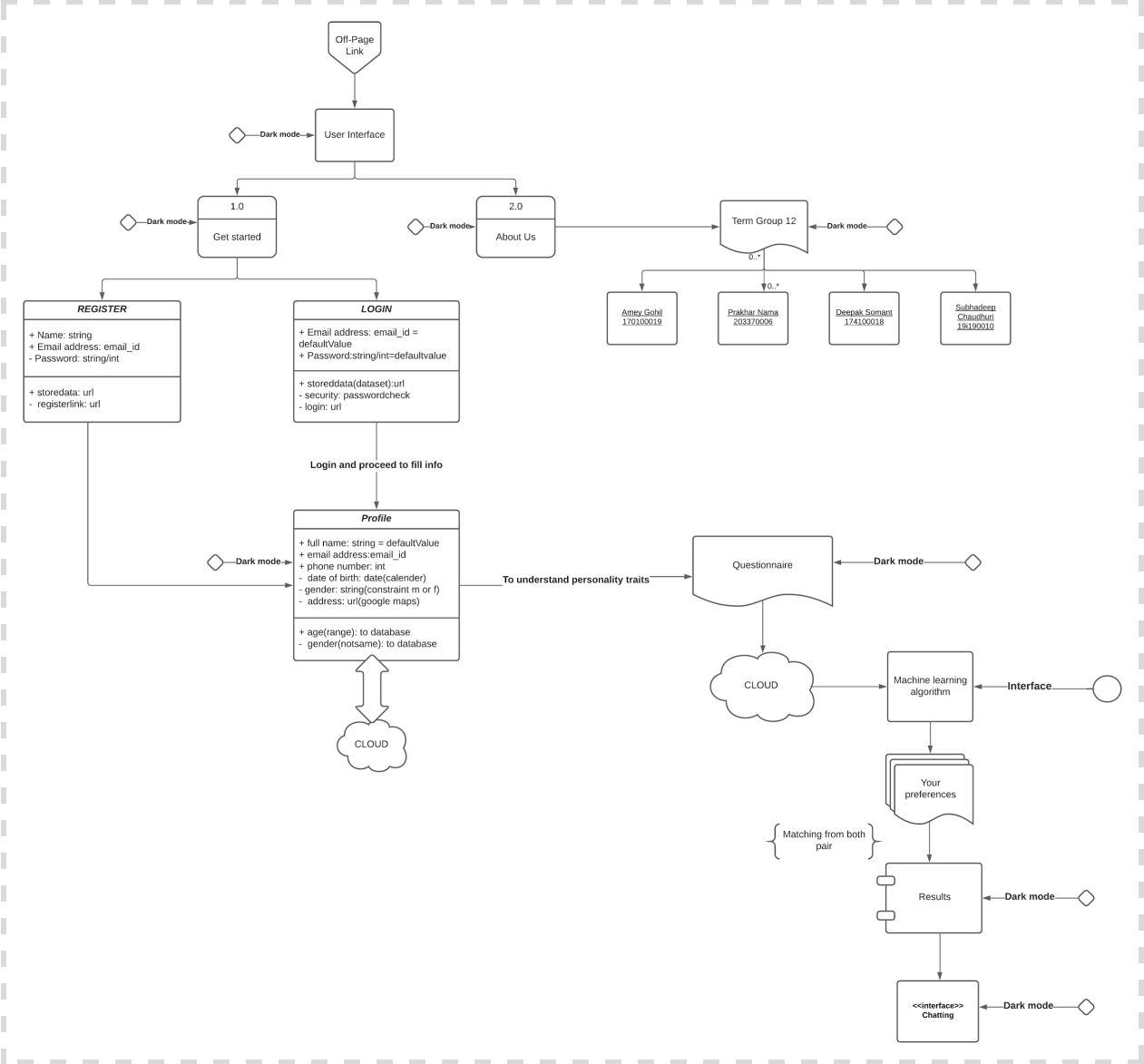
To know the details related to the dataset, click the link here

## High Level activity object and class diagram




**To have a look at the detailed chart, click here**

## High-level sequence diagram with the type of user input and output

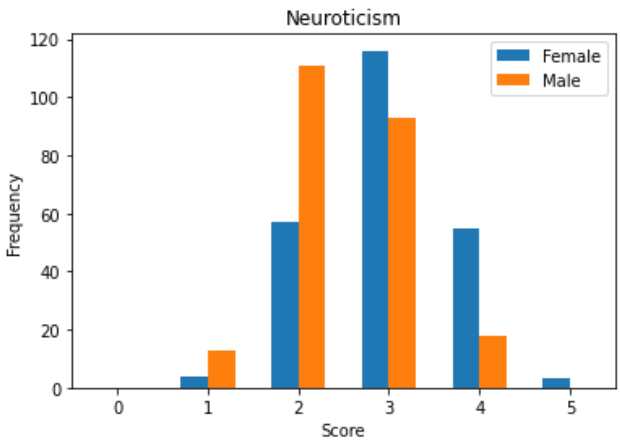



**To have a look at the detailed chart, click [here]**

## ML Model Output

- ❑ Predicted relationship satisfaction score between the current user and other registered users in the database.
- ❑ Suggest an ordered (in descending order of expected satisfaction scores) list of registered individuals on our platform to the current user.

# Descriptive Statistics Performed




## Neuroticism

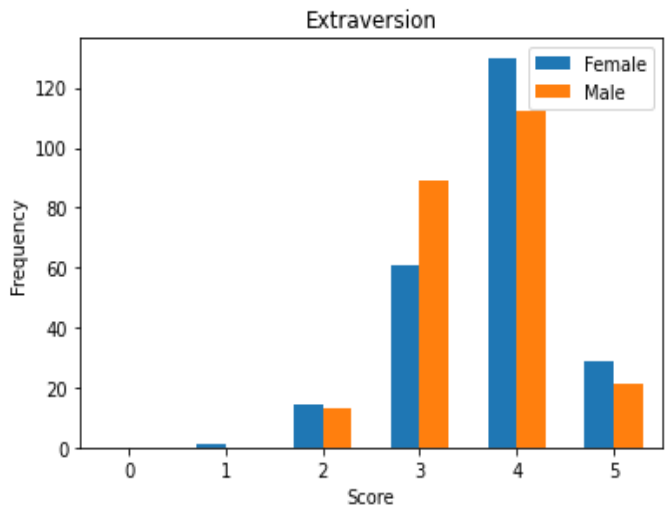
| | Female | Male |
|---|---|---|
| Mean | 2.9 | 2.5 |
| Sample Sd | 0.771 | 0.716 |

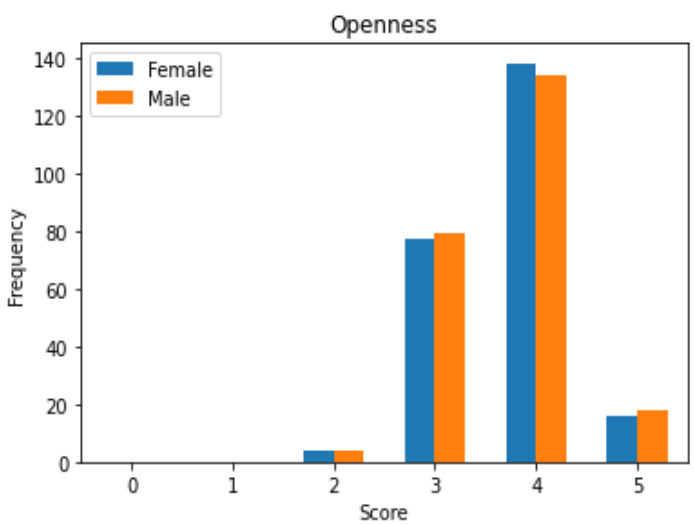
Mean sample neuroticism score of females is higher than the male.

## Extraversion

| | Female | Male |
|---|---|---|
| Mean | 3.7 | 3.6 |
| Sample Std | 0.766 | 0.727 |

Mean sample extraversion score of females is almost the same as in the case of males.




## Openness

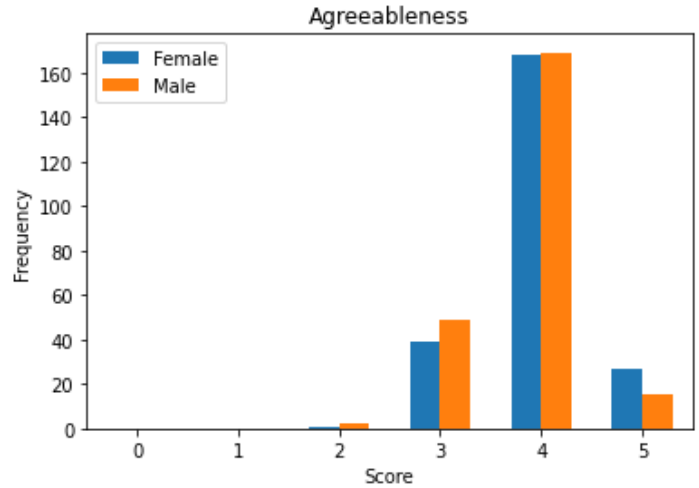
| | Female | Male |
|---|---|---|
| Mean | 3.7 | 3.6 |
| Sample Std | 0.614 | 0.628 |

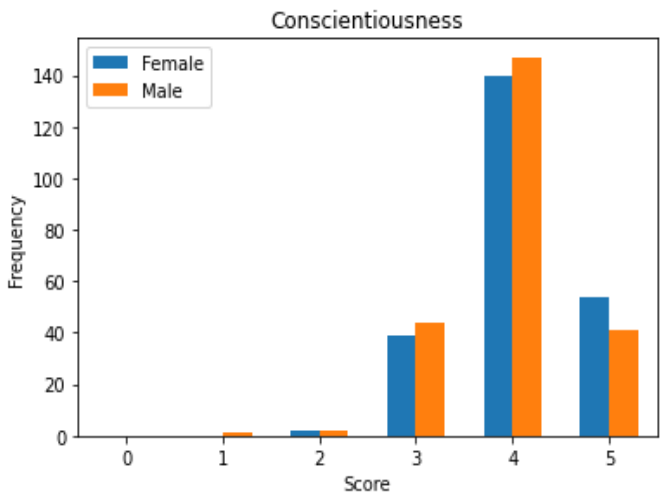
Mean sample openness score of females is almost the same as in the case of males.

## Agreeableness

| | Female | Male |
|---|---|---|
| Mean | 3.9 | 3.8 |
| Sample Std | 0.542 | 0.529 |

Mean sample agreeableness score of females is almost the same as in the case of males.




## Conscientiousness

| | Female | Male |
|---|---|---|
| Mean | 4.0 | 3.9 |
| Sample Std | 0.653 | 0.657 |

Mean sample conscientiousness score of females is almost the same as in the case of males.

# Model Validation

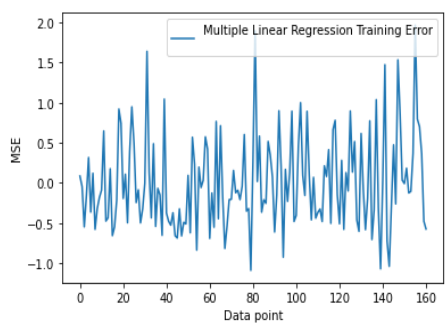
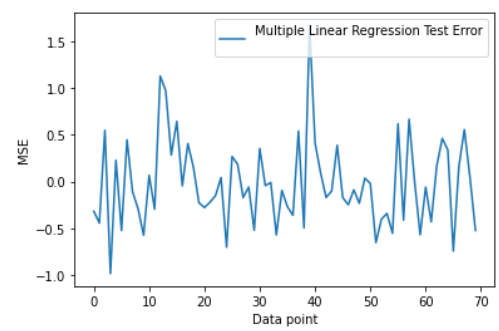
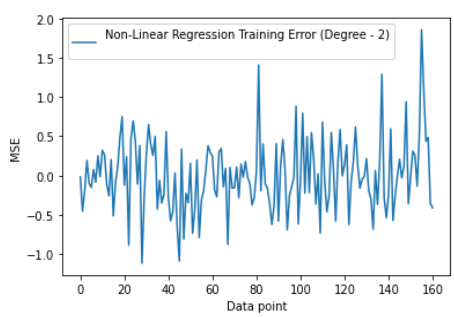
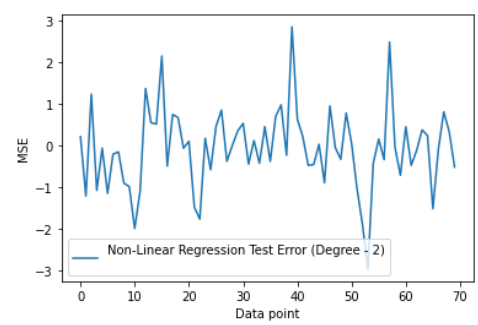
## **Model 1**: Multiple Linear Regression Model

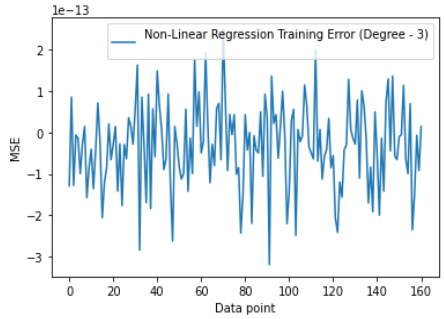
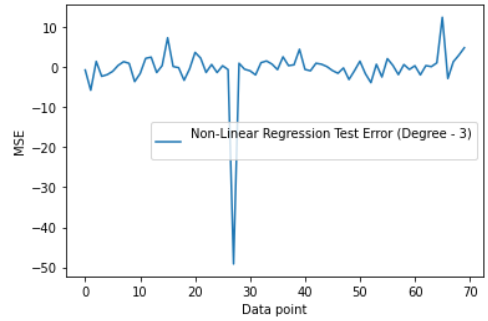
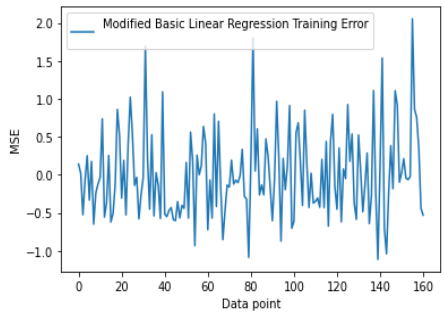
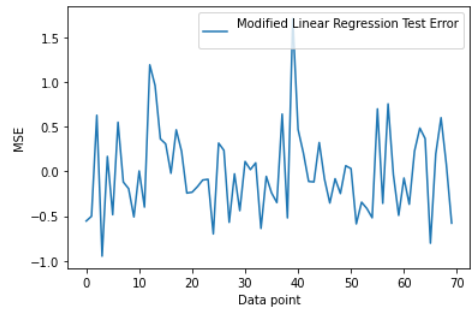
We implemented the Multiple Linear Regression Model and some of its modified versions as follows:

- Multiple Linear Regression
- Non-Linear Multiple Regression with degree 2 and 3
- Multiple Linear Regression with additional variables which are a function of existing variables. For this, we added two variables of similarity and dissimilarity measures using the euclidean norm with the two vectors of character traits of male and female.

We trained the model on 70% of the total data and evaluated the model on the rest.

Results and plots are described in the table below:

| Model | Train | Test |
|---|---|---|
| Multiple Linear Regression | **MSE** - 0.330369<br>  | **MSE** - 0.238187<br>  |
| Non-Linear Multiple Regression (degree 2) | **MSE** - 0.212730<br>  | **MSE** - 0.941158<br>  |

| | | |
|---|---|---|
| Non-Linear Multiple Regression (degree 3) | **MSE** - 1.165637e-26   | **MSE** - 41.049492   |
| Multiple Linear Regression with similarity and dissimilarity variables | **MSE** - 0.318433   | **MSE** - 0.210403   |

**Conclusions:**

- Train MSE obtained in Non-Linear regression decreased significantly on increasing the degree of polynomial, but it also led to a rise in train error. This is caused due to overfitting and hence increasing the degree will not yield correct results.

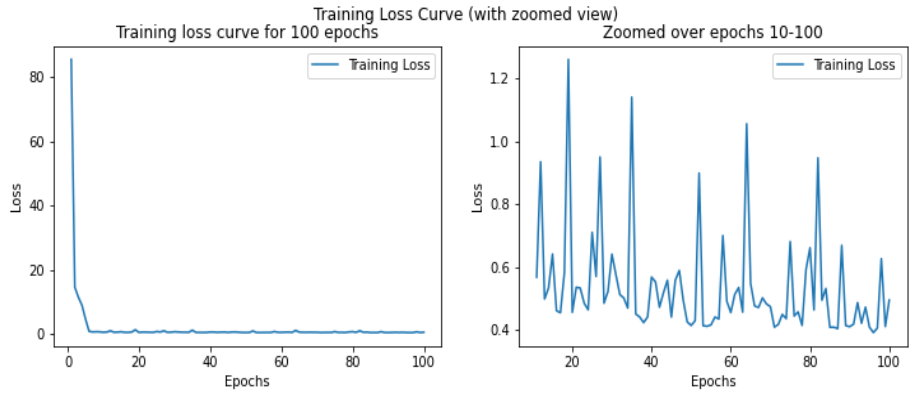
- Proceeding with the multiple linear regression model, we found that test and train error were nearly the same, and in most cases less in the model with additional variables than in the normal model. So we are proceeding to select the model with the additional variables of similarity and dissimilarity measures to be our final Linear Regression model.

## **Model 2**: Multi-Layer Perceptron

We implemented a simple multi-layer perceptron architecture with:

- A single hidden layer of 20 neurons.
- The input layer consists of the number of features considered for the formulation (which in our case is 10)
- The output layer consists of a single neuron for the predicted relationship satisfaction score
- Loss function: Mean Squared Error loss
- Optimizer considered: SGD

We train the model on a dataset consisting of 70% of the total data and evaluate the model on the rest. On training the model over 100 epochs, we get the following loss curve during the training phase:




- The loss curve suggests effective training of the model over the epochs.
- On a certain dataset split and the corresponding evaluation of the model, the MSE came out to be 0.22028, which is pretty good.
- The model was used to predict the score for a previously unseen data point, on which the predicted score and the actual score were very close. The screenshot attached below bears proof of the result.

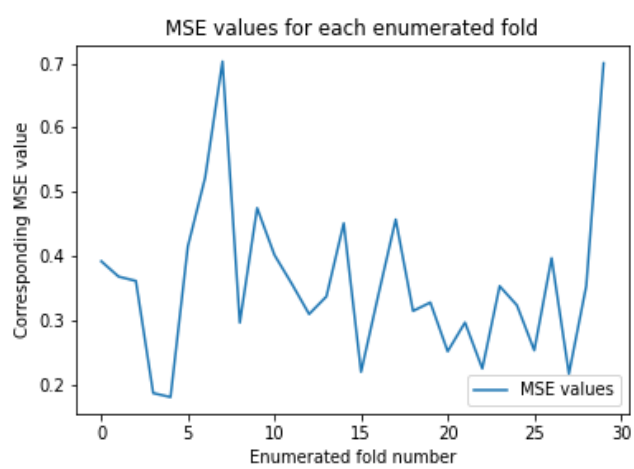
```
X_random = X[-1:]
pred_random = model.predict(X_random)
print("\033[1m Predicted Score: \033[0m", pred_random[0][0])
print("\033[1m Actual score: \033[0m",y[-1:].values[0])
```

```
Predicted Score:  4.378393
Actual score:  4.5
```

- For cross-validating the performance of our model, we also evaluated the model by k-fold cross-validation with 10 folds, and 3 repetitions of the cross-validator.

  The mean MSE from all the evaluations came out to be 0.360 with a standard deviation of 0.124, which is very much comparable to that from Model 1.

A plot of each individual MSEs from the evaluations on each fold is given below:



# User Interface

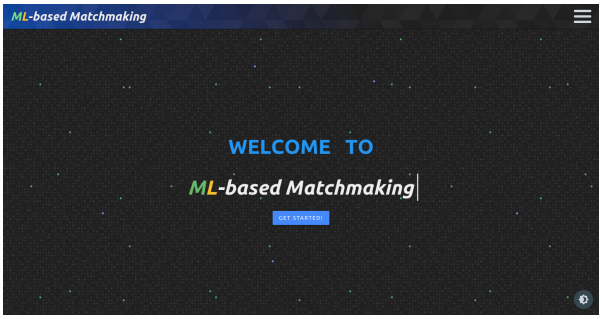
We prepared a user interface for the end-users using the following:

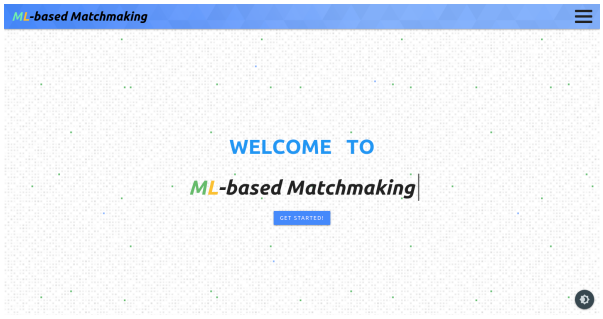
- ❏ Front end: HTML, CSS, JavaScript
- ❏ Back end: PHP and MySQL

Here, we present a glimpse of the User interface.

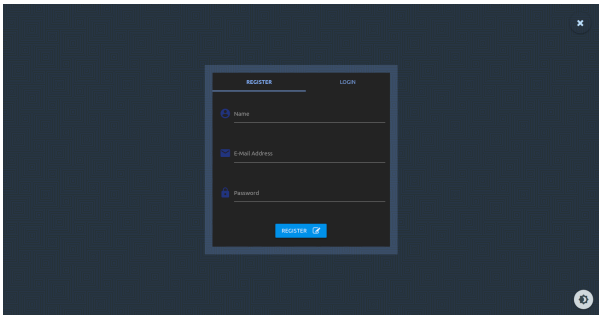
★ What's the best part? It comes with a toggle option between ***Dark Mode*** and ***Light Mode***.

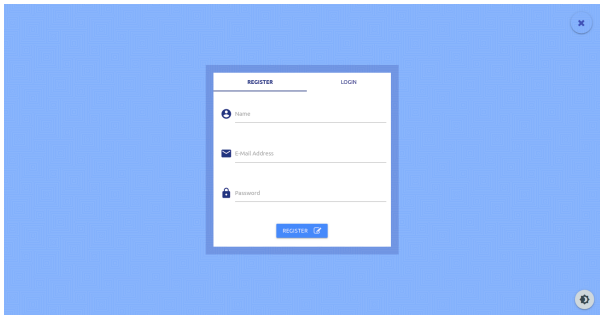
## Home Page

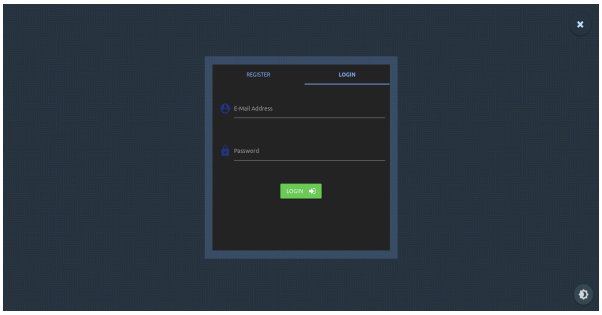


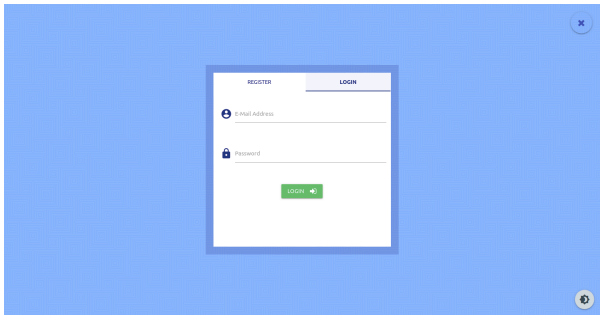




## Registration / Login Page

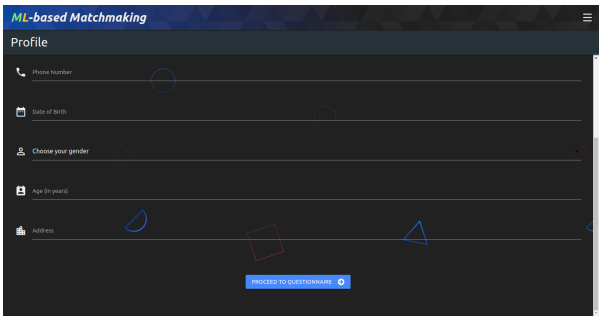


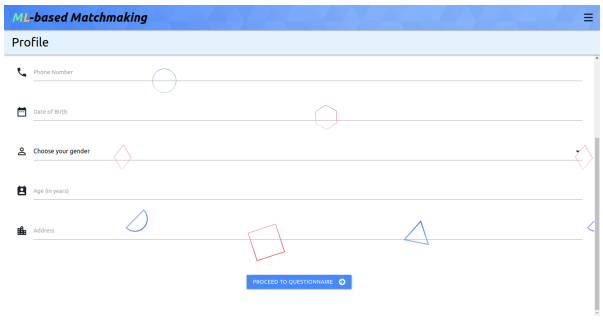



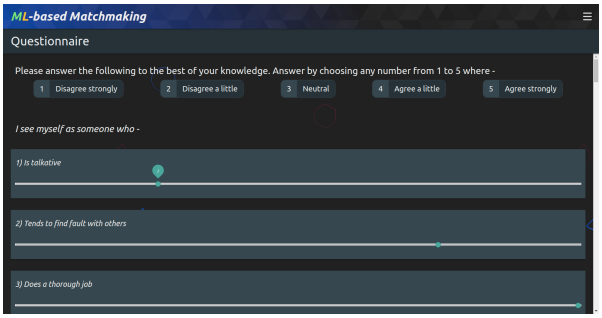



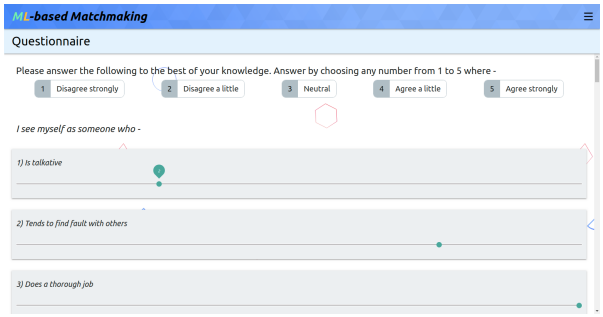
## User Profile Page



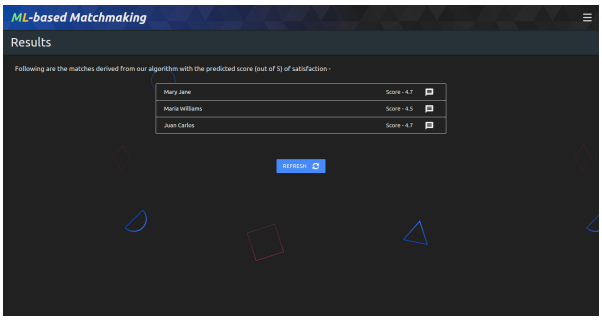



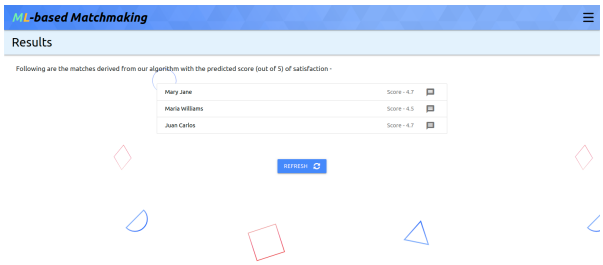

## Questionnaire Fill up Section




## Results With The Best Found Matches

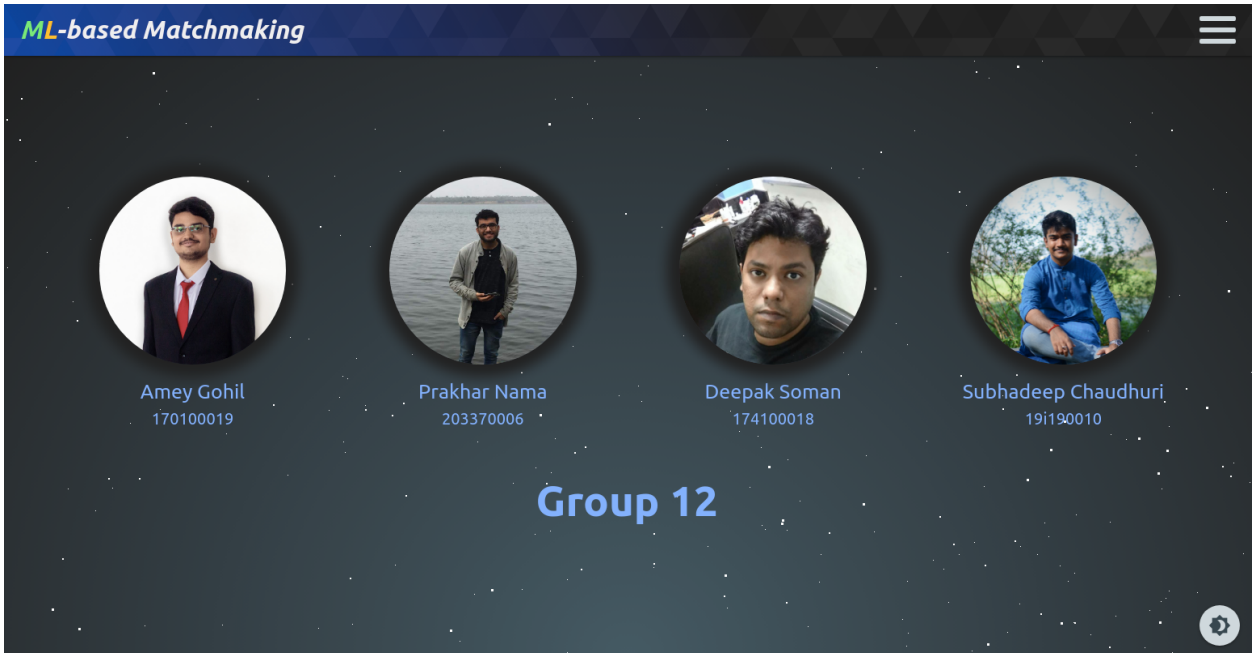



The output on the last page gives you the users registered in our platform who we think would be an ideal match for you.

And off you go! Have fun. 💗

# About Us


ML-based Matchmaking
Amey Gohil
170100019
Prakhar Nama
203370006
Deepak Soman
174100018
Subhadeep Chaudhuri
19i190010
Group 12

# References


1. Shelby B. Scott, “Reasons for Divorce and Recollections of Premarital Intervention: Implications for Improving Relationship Education”,Couple Family Psychol, 2013 June ; 2(2): 131–145.

2. Mikael Jensen, “Personality Traits, Learning, and Academic Achievements”, Journal of Education and Learning; Vol. 4, No. 4; 2015.

3. Leonora Risse, “Personality and pay: do gender gaps in confidence explain gender gaps in wages?”, Oxford Economic Papers, 70(4), 2018, 919–949.

4. John M. Malouff, “The Five-Factor Model of personality and relationship satisfaction of intimate partners: A meta-analysis”, Journal of Research in Personality 44 (2010) 124–127.

5. Hossein Dabiriyan, “Personality traits and conflict resolution styles: A meta-analysis”, Personality and Individual Differences 157 (2020) 109794.

6. Aleksandra Rogowska, “The Relationship of Number of Sexual Partners with Personality Traits, Age, Gender and Sexual Identification”,Psychology & Sexuality.

7. Terracciano, Antonio et al. “Five-Factor Model personality profiles of drug users.” BMC psychiatry vol. 8 22. 11 Apr. 2008, doi:10.1186/1471-244X-8-22.

8. Kathrin Schaffhauser, “Dyadic longitudinal interplay between personality and relationship satisfaction: A focus on neuroticism and self-esteem”, Journal of Research in Personality 53 (2014) 124–133.

9. Kourosh Sayegmiri, “The relationship between personality traits and marital satisfaction: a systematic review and meta-analysis”, BMC Psychology, 2020,8:15.

10. John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), Handbook of personality: Theory and research (Vol. 2, pp. 102–138). New York: Guilford Press.