# FINAL PROJECT – SPEED DATING

CS513-A: KNOWLEDGE DISCOVERY & DATA MINING

# PROJECT GROUP 7

**Atharv Subhekar**

20015840

**Kunal Mandalya**

20012146

**Nidhi Sura**

20011965

**Vismay Rathod**

20013002

# CONTENTS

# PROBLEM STATEMENT

- How do people choose partners in speed dating?
- Do specific qualities in partners stand out more than others?
- Are there certain gender or age-based preferences/biases?
- Can we predict a match with a certain degree of confidence?

# DESCRIPTION OF THE DATASET

Speed Dating events from 2002-2004; 123 features; 8000+ samples

**Demographics**

- Gender
- Age
- Age gap
- Nationality/Race
- Profession

**Qualities**

- Attractiveness
- Sincerity
- Intelligence
- Humor
- Ambition

**Interests**

- Sports
- Dining
- Art
- Hiking
- Concerts

**Preferences**

- Qualities
- Interests
- Same race
- Same religion
- Concerts

**Others**

- Expected matches
- Happiness with speed dating people
- Met before the day
- Wave

## TARGET: 'DECISION'

Why not match?

# APPROACH

**Step 1**

Data cleaning, replacement of missing values

**Step 2**

Data analysis & visualization

**Step 3**

Data pre-processing (encoding & normalization)

**Step 4**

Feature selection methodologies

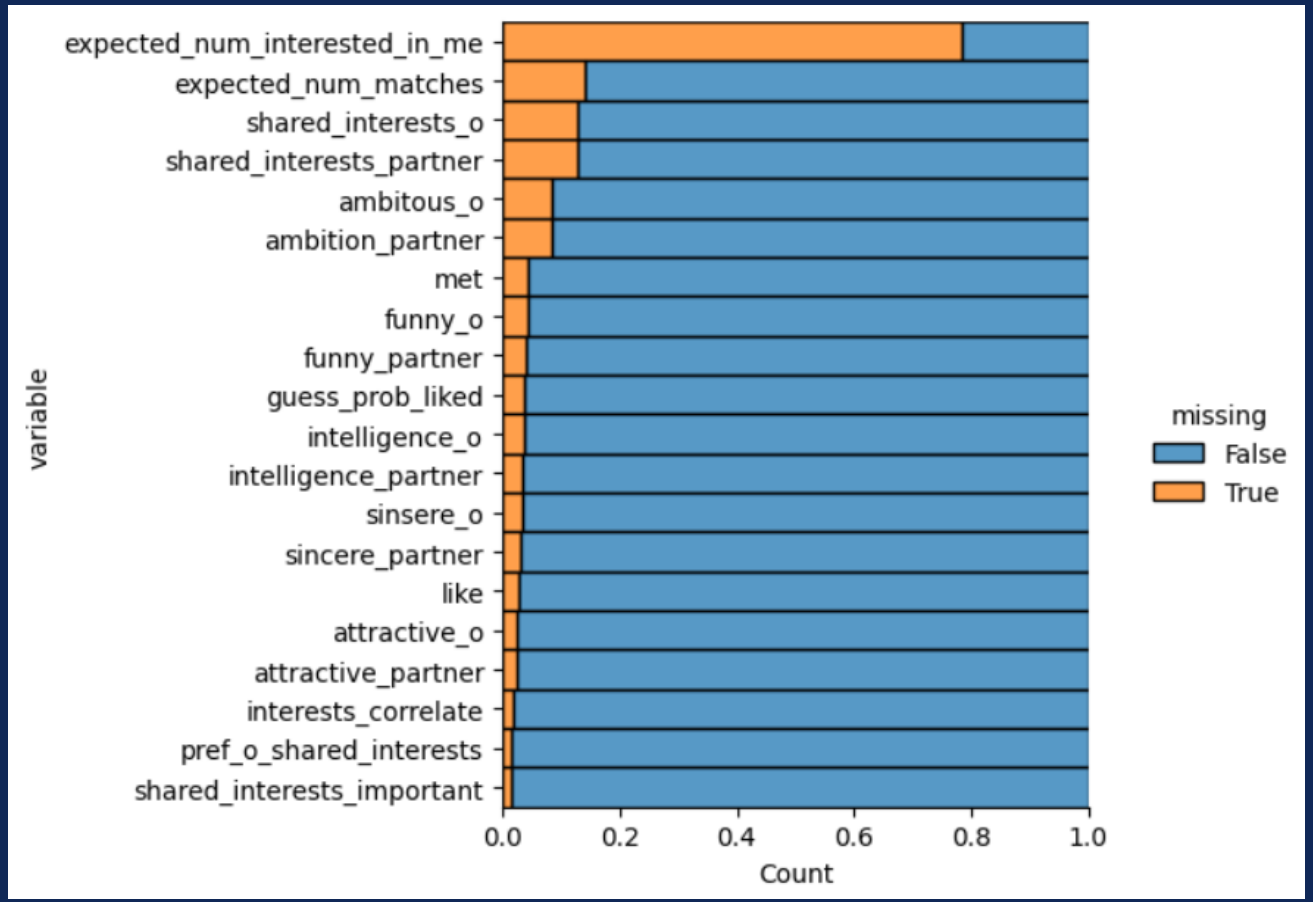
**Step 5**

Target prediction using machine learning models

# STEP 1

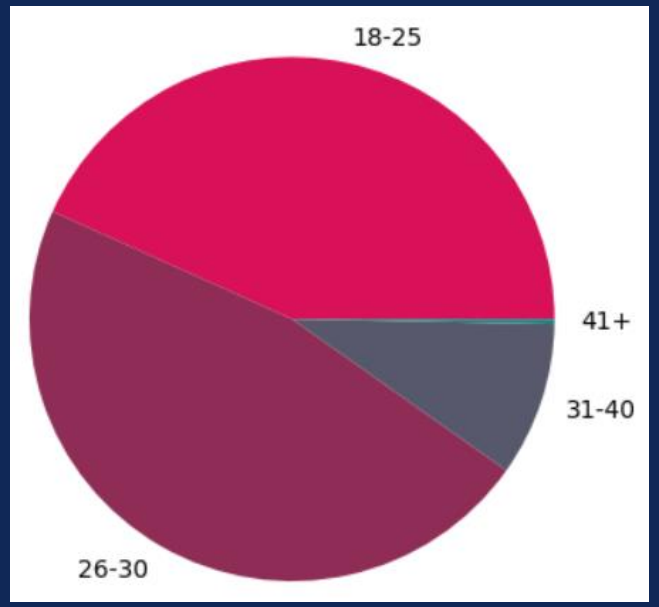
## INITIAL DATA CLEANING, REPLACEMENT OF MISSING VALUES

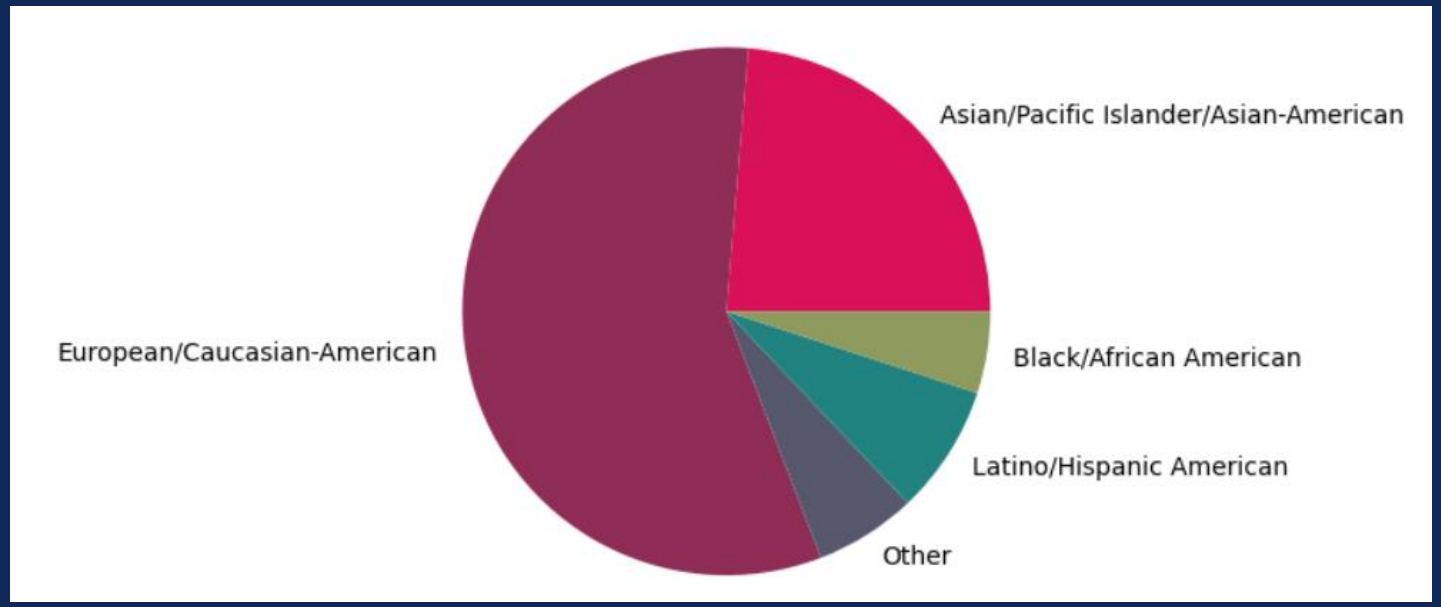
- Delete column with large number of missing values
- Replace missing values in type ‘float’ and ‘int’ columns with median
- Replace missing values in type ‘object’ columns with most frequent value
- Drop irrelevant column ‘has_null’
- Drop repeated columns beginning with ‘d_’, since they only convert numeric value to range
- Remove ‘b’ from column values containing it




Heatmap of top 20 columns with the most missing values

# STEP 2
## DATA ANALYSIS & VISUALIZATION

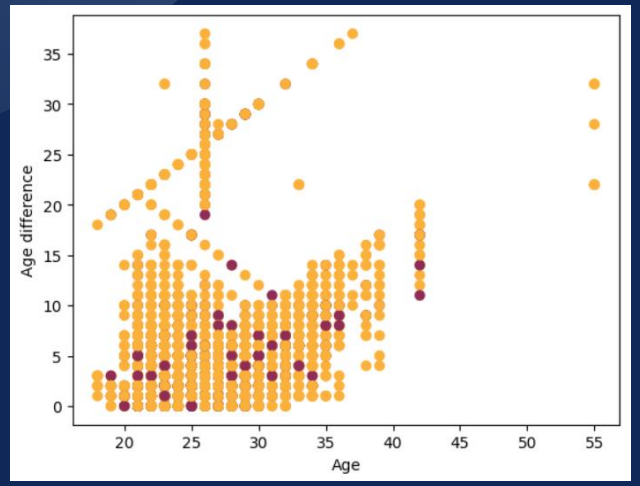


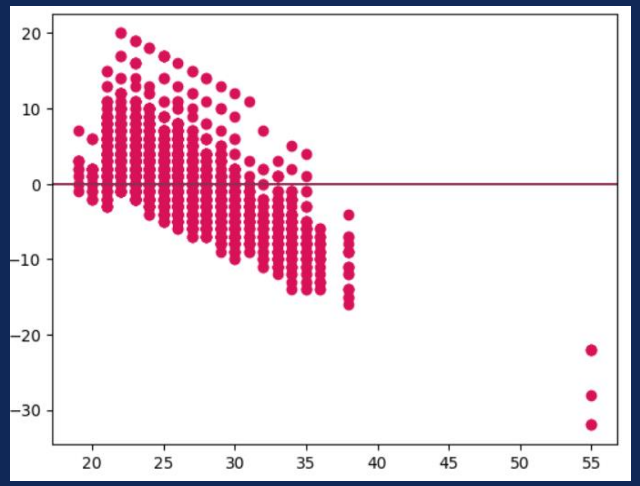

Demographics

# STEP 2
## DATA ANALYSIS & VISUALIZATION

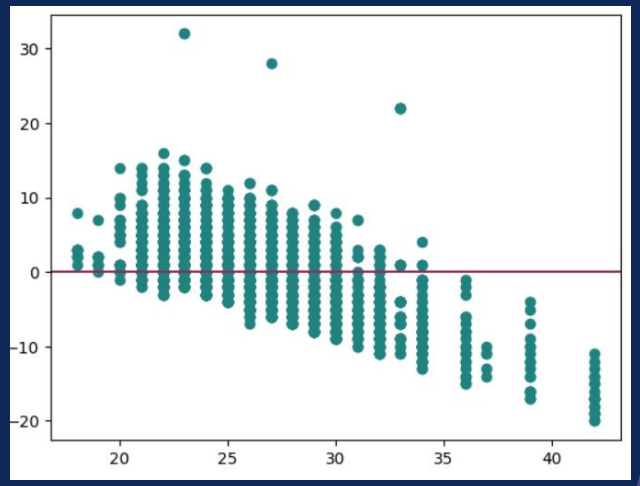
**Age vs. age difference**
Overall




**Age vs. age difference**
Women

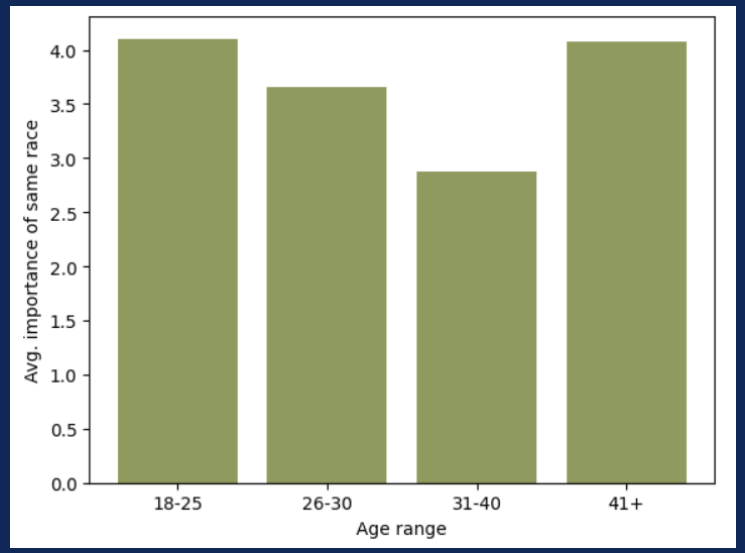



**Age vs. age difference**
Men

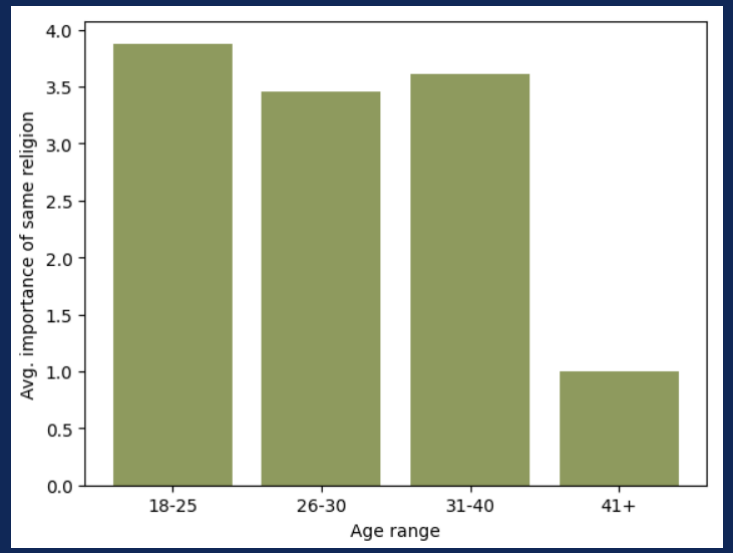


# STEP 2
## DATA ANALYSIS & VISUALIZATION

**Age vs. importance of same race**

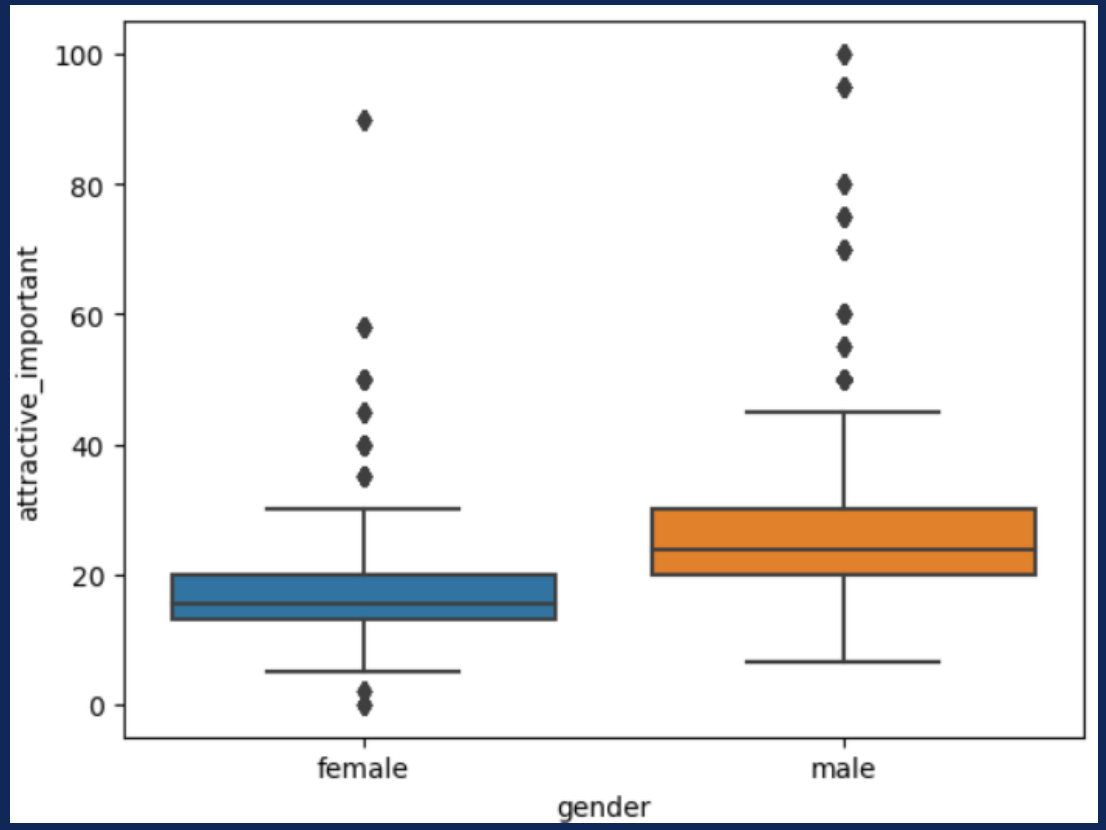


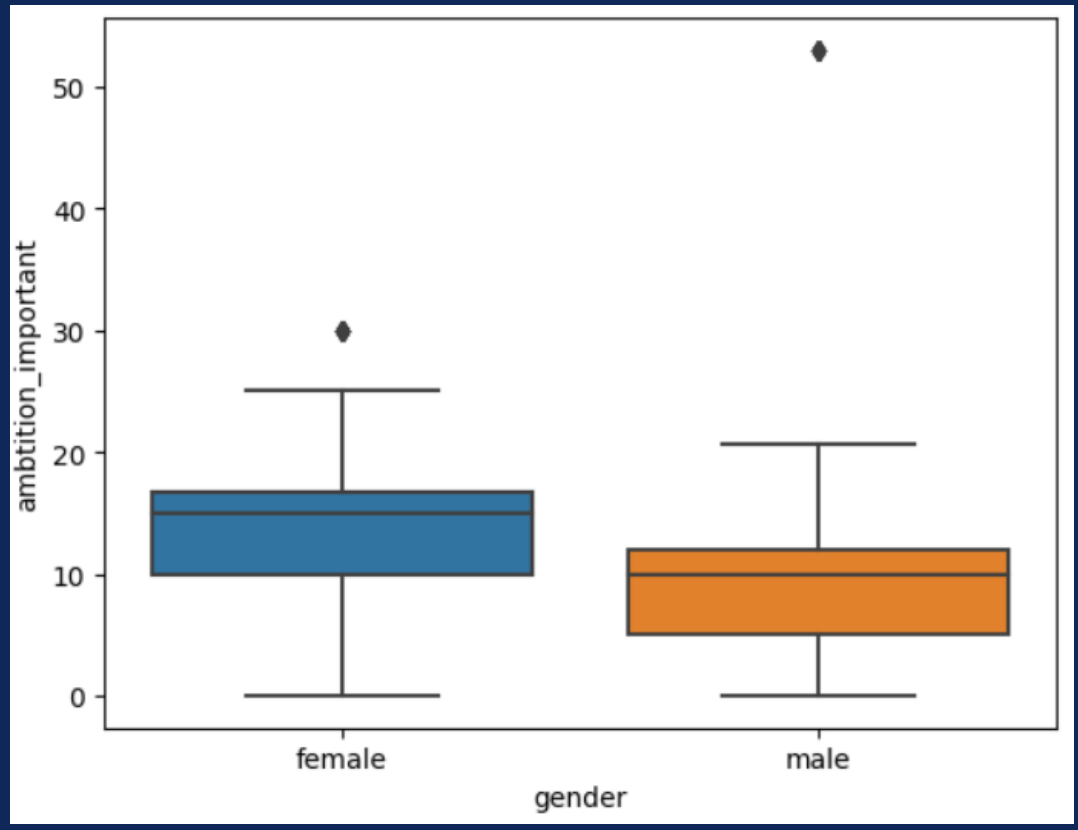

**Age vs. importance of same religion**



# STEP 2

## DATA ANALYSIS & VISUALIZATION

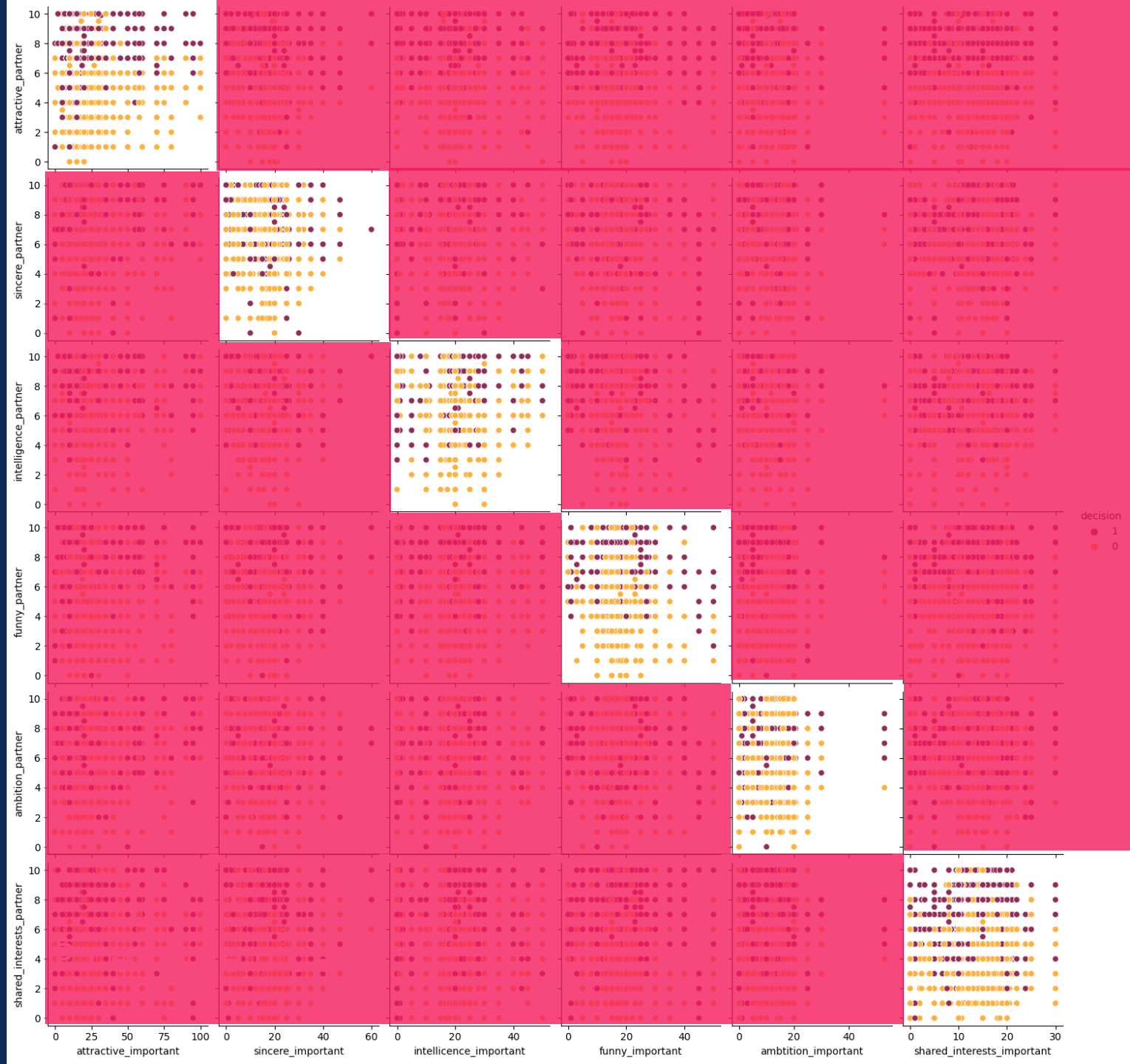


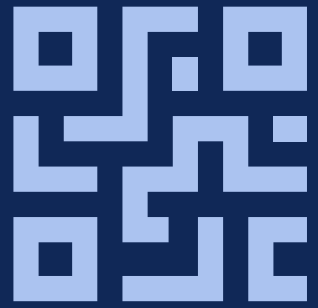

Preferences

# STEP 2
## DATA ANALYSIS & VISUALIZATION

Pair plot of importance of qualities vs. partner's possession of said qualities



# STEP 3:
## DATA PRE-PROCESSING (ENCODING & NORMALIZATION)



Ordinal encoder

Min-max scaler

# STEP 4
## FEATURE SELECTION METHODOLOGIES

- Pearson
- Chi-2
- RFE
- Logistic Regression
- Random Forest

| | Feature | Pearson | Chi-2 | RFE | Logistics | Random Forest | Total |
|---|---|---|---|---|---|---|---|
| 1 | shared_interests_partner | True | True | True | True | True | 5 |
| 2 | guess_prob_liked | True | True | True | True | True | 5 |
| 3 | like | True | True | True | True | True | 5 |
| 4 | expected_num_matches | True | True | True | True | True | 5 |
| 5 | attractive_partner | True | True | True | True | True | 5 |
| 6 | funny_partner | True | True | True | True | True | 5 |
| 7 | ambition_partner | True | True | True | True | False | 4 |
| 8 | sincere_partner | True | True | True | True | False | 4 |
| 9 | attractive_o | True | False | True | True | False | 3 |
| 10 | funny | True | False | True | True | False | 3 |
| 11 | intelligence | True | False | True | True | False | 3 |
| 12 | d_age | True | False | True | True | False | 3 |
| 13 | attractive | True | False | True | True | False | 3 |
| 14 | attractive_important | True | False | False | True | False | 2 |
| 15 | sports | True | False | True | False | False | 2 |

# STEP 5:

## PREDICTION USING MACHINE LEARNING MODELS

**Decision Tree**

**AdaBoost Classifier**

**Random Forest**
With Bagging
With Randomized Search CV

**K-nearest neighbor**
with Grid Search CV

**Logistic Regression**
with Grid Search CV

**Support Vector Machine**
with Grid Search CV

**Naive Bayes Classifier**

**Multilayer Perceptron**

# DECISION TREE CLASSIFIER

Accuracy – 74%

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.78 | 0.78 | 0.78 |
| 1 | 0.68 | 0.68 | 0.68 |

# ADABOOST CLASSIFIER

Accuracy – 74%

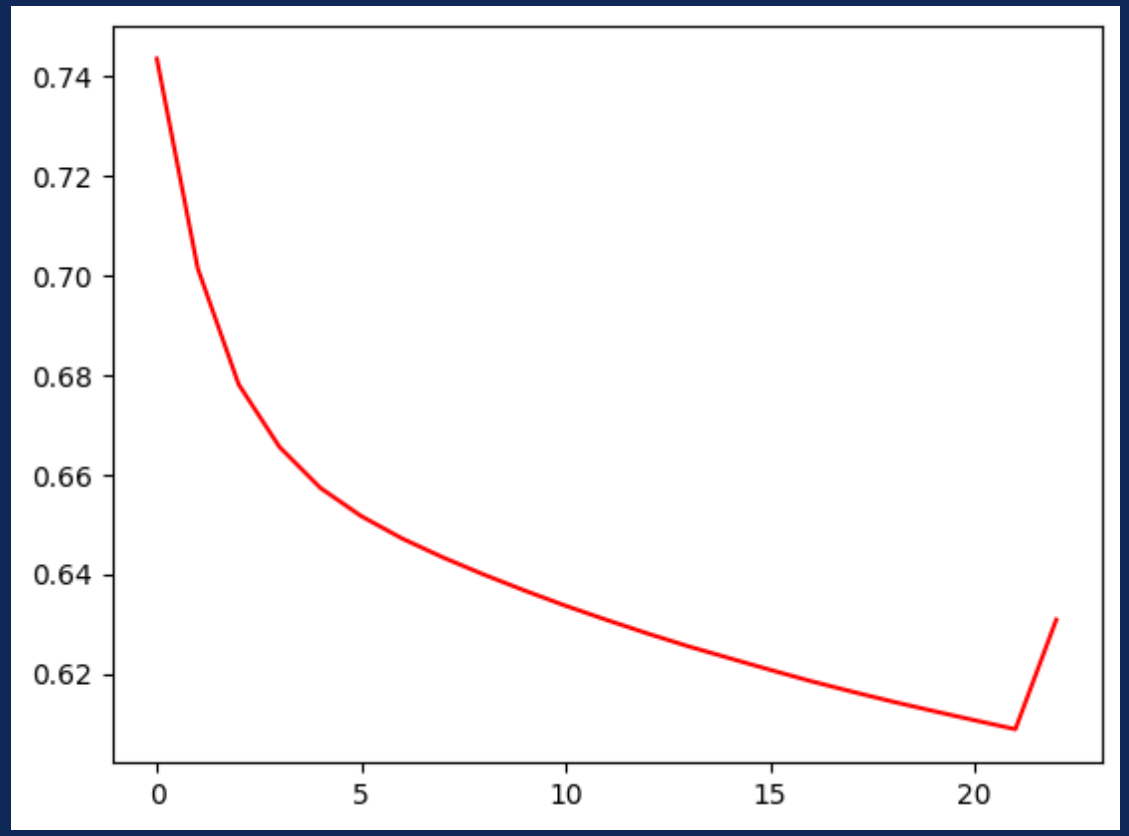
| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.81 | 0.83 | 0.82 |
| 1 | 0.76 | 0.73 | 0.75 |

# MULTI LAYER PERCEPTRON

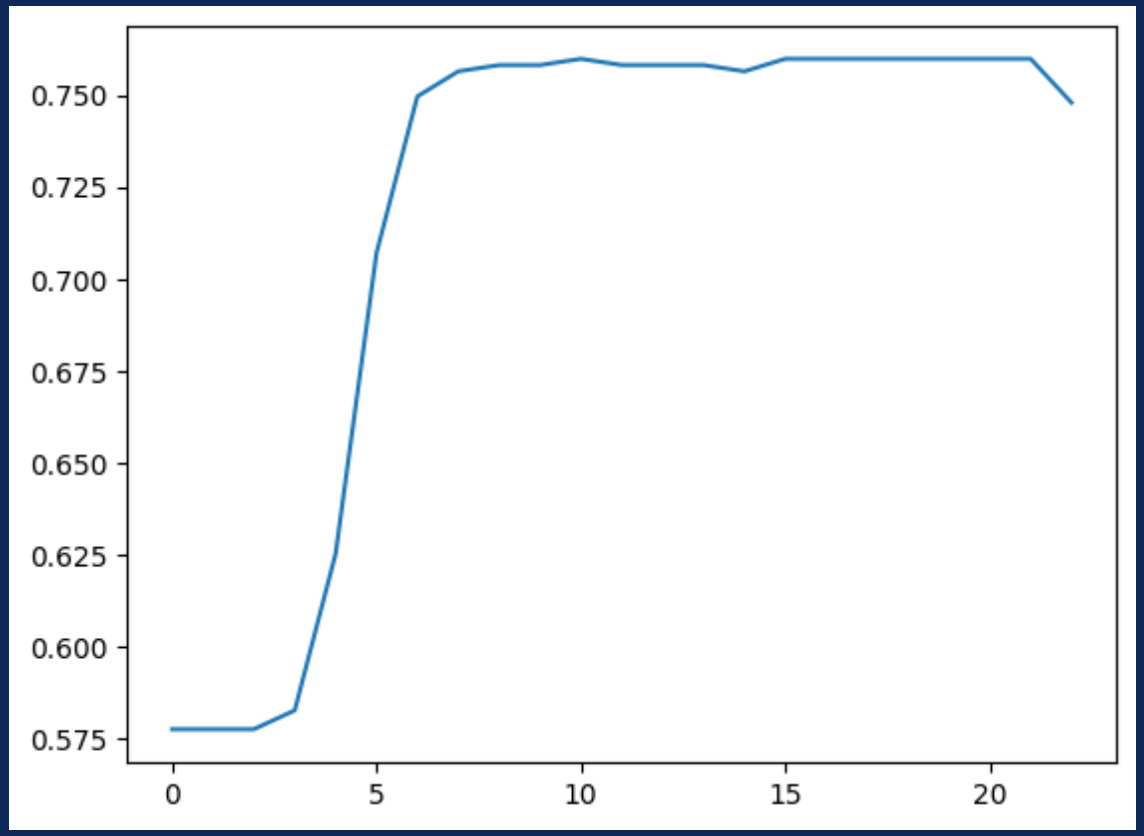
Accuracy – 74%

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.80 | 0.83 | 0.81 |
| 1 | 0.77 | 0.73 | 0.75 |

# MULTI LAYER PERCEPTRON




Loss Curve




Validation Score

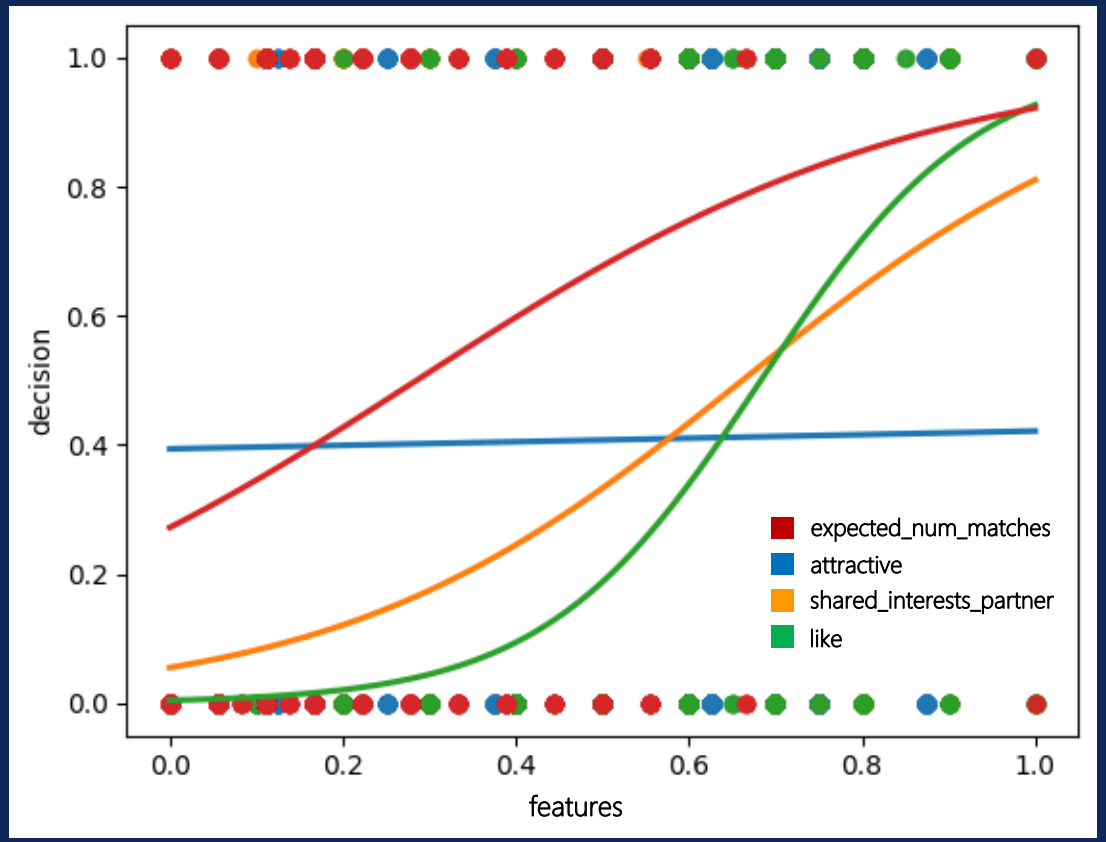
# LOGISTIC REGRESSION + GRIDSEARCHCV

GridSearchCV Best Parameters

| Feature | Tuned Value |
|---|---|
| 'class_weight' | 'balanced' |
| 'fit_intercept' | True |
| 'penalty' | None |
| 'solver' | 'lbfgs' |

Classification Report

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.84 | 0.76 | 0.8 |
| 1 | 0.7 | 0.8 | 0.75 |



# SUPPORT VECTOR MACHINE + GRIDSEARCHCV

**GridSearchCV Best Parameters**

| Feature | Tuned Value |
|---|---|
| ‘kernel’ | ‘linear’ |
| ‘degree’ | 1 |
| ‘class_weight’ | 'balanced' |
| ‘probability’ | True |

**Classification Report**

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.84 | 0.75 | 0.79 |
| 1 | 0.7 | 0.8 | 0.74 |

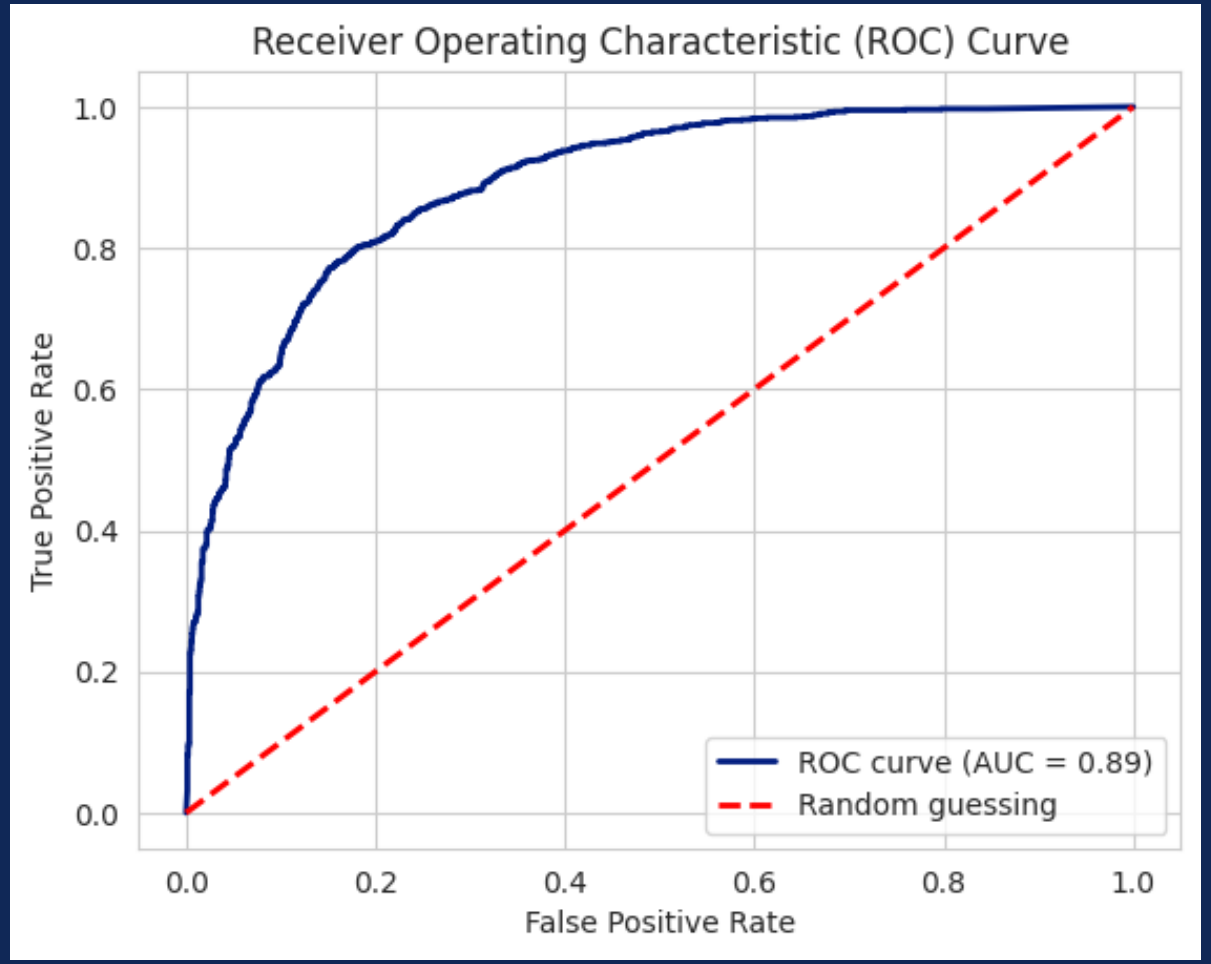
# K-NEAREST NEIGHBORS+ GRIDSEARCHCV

GridSearchCV Best Parameters

| Feature | Tuned Value |
|---|---|
| 'metric' | 'manhattan' |
| 'n_neighbors' | 19 |
| 'weights' | distance |

Classification Report

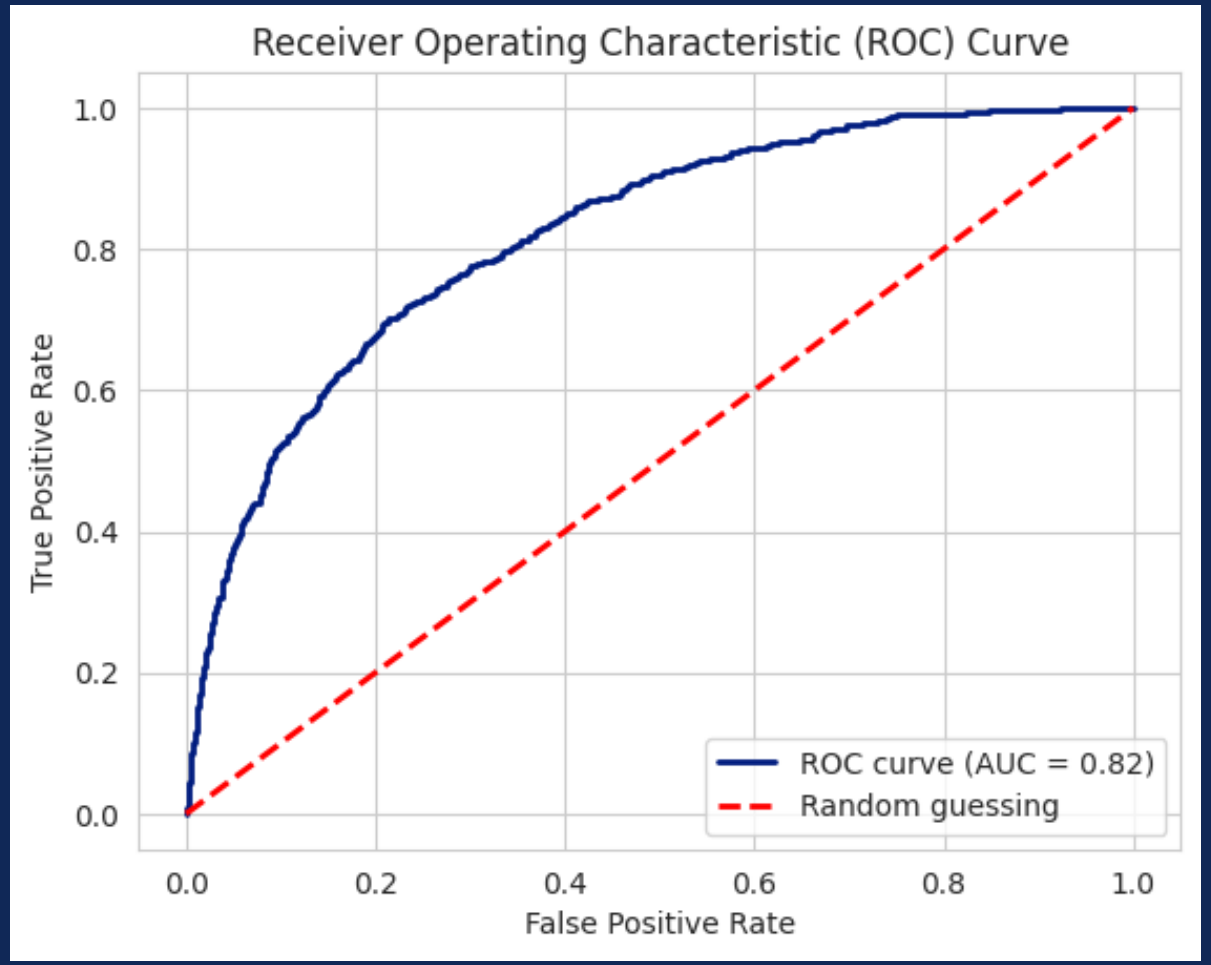
| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.84 | 0.85 | 0.84 |
| 1 | 0.78 | 0.77 | 0.78 |



# NAIVE BAYES

Classification Report

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.82 | 0.69 | 0.75 |
| 1 | 0.64 | 0.78 | 0.70 |

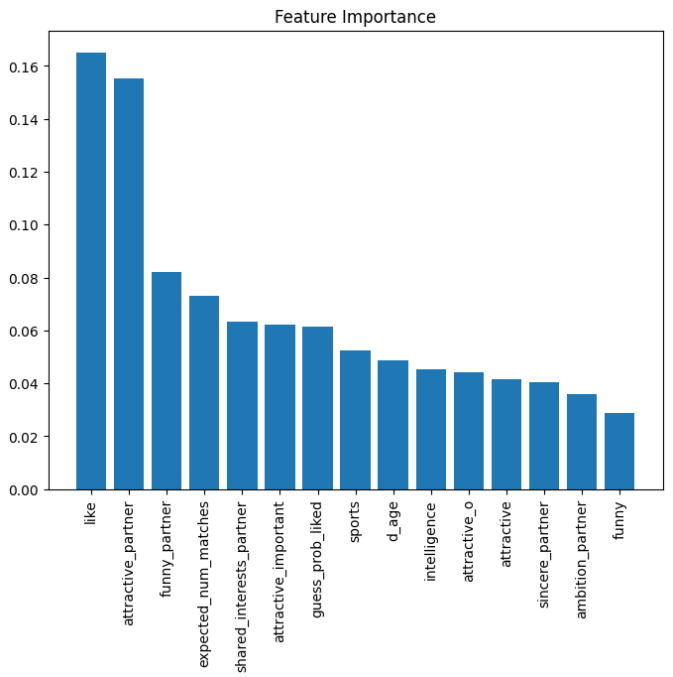


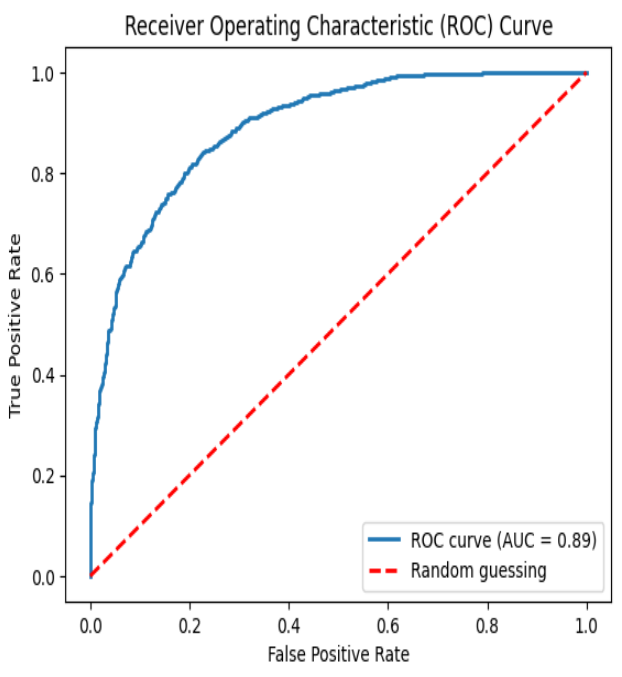
# RANDOM FOREST WITH RANDOMIZEDSEARCHCV

Classification Report

| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.83 | 0.83 | 0.83 |
| 1 | 0.76 | 0.76 | 0.76 |

| Accuracy | 80% |
|---|---|

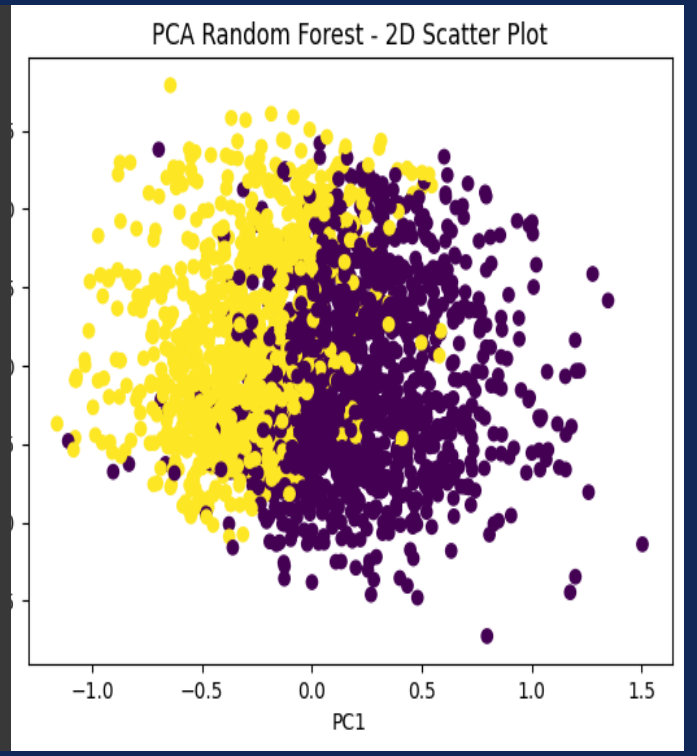


# PCA USING RANDOM FOREST

| Accuracy | 70% |
|---|---|

| | PC1 | PC2 |
|---|---|---|
| shared_interests_partner | -0.405030 | 0.116345 |
| guess_prob_liked | -0.357104 | -0.039994 |
| like | -0.406305 | 0.120415 |
| expected_num_matches | -0.081315 | -0.064535 |
| attractive_partner | -0.358364 | 0.158319 |
| funny_partner | -0.411648 | 0.130678 |
| ambition_partner | -0.274412 | 0.081027 |
| sincere_partner | -0.296954 | 0.069262 |
| attractive_o | -0.068593 | -0.119715 |
| funny | -0.067543 | -0.106705 |
| intelligence | -0.147328 | -0.198739 |
| d_age | -0.000298 | 0.026350 |
| attractive | -0.100449 | -0.184507 |
| attractive_important | -0.032373 | -0.099597 |
| sports | -0.182368 | -0.896004 |

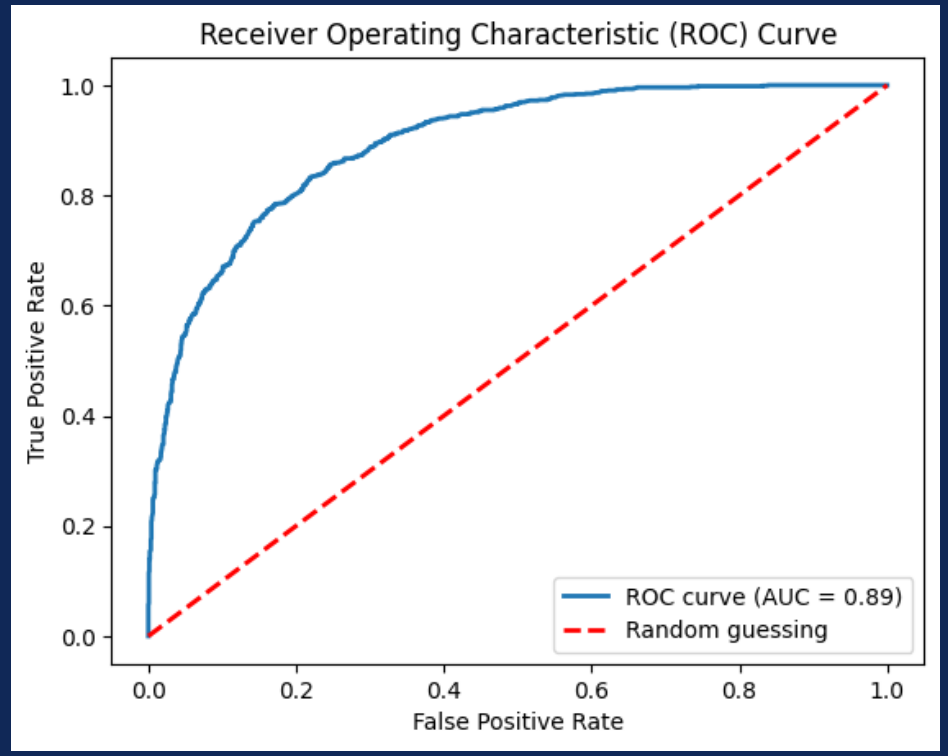


# RANDOM FOREST WITH BAGGING

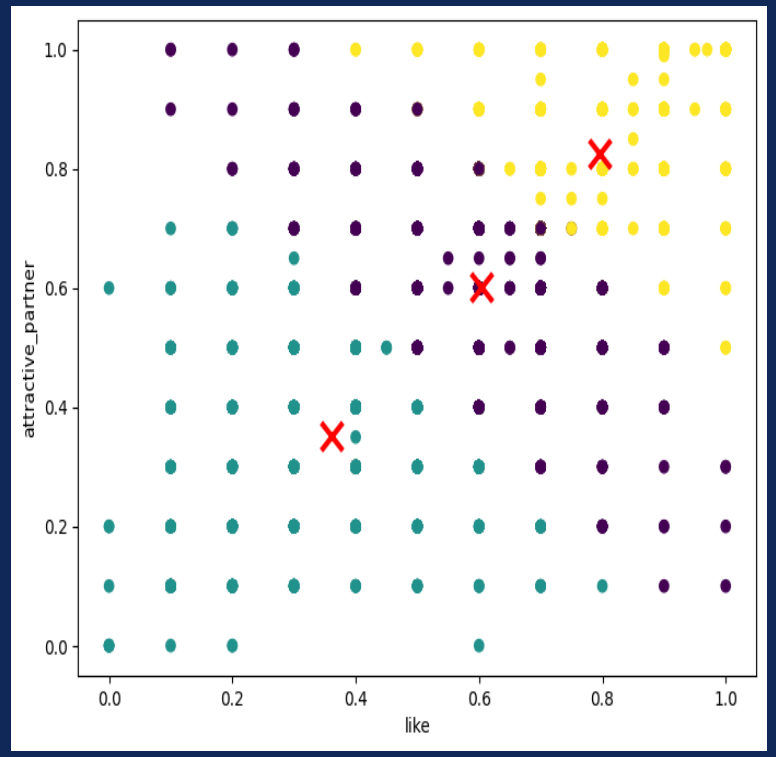
Classification Report

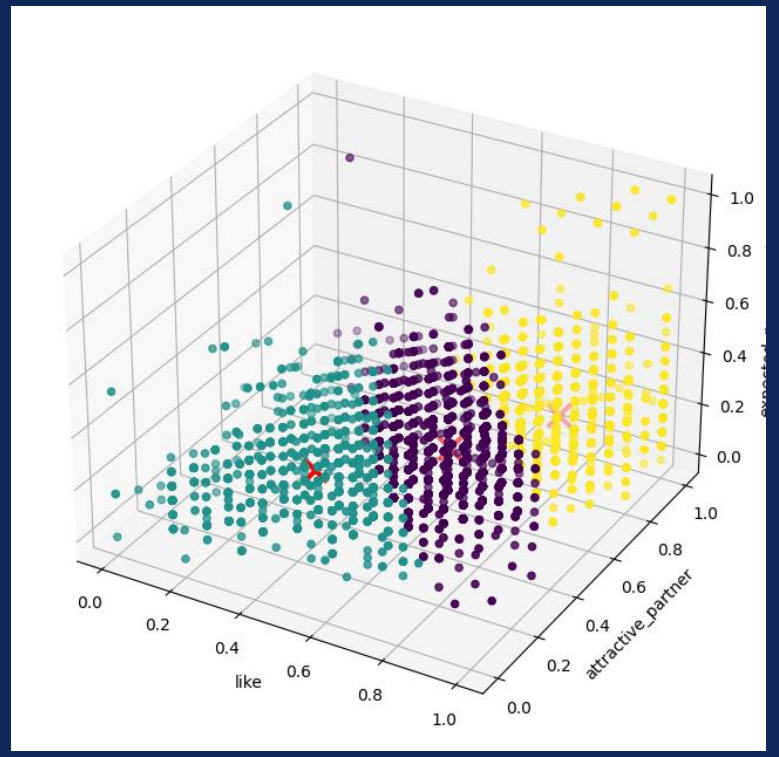
| Class | Precision | Recall | F1 Score |
|---|---|---|---|
| 0 | 0.83 | 0.85 | 0.84 |
| 1 | 0.78 | 0.75 | 0.77 |

| Accuracy | 81% |
|---|---|



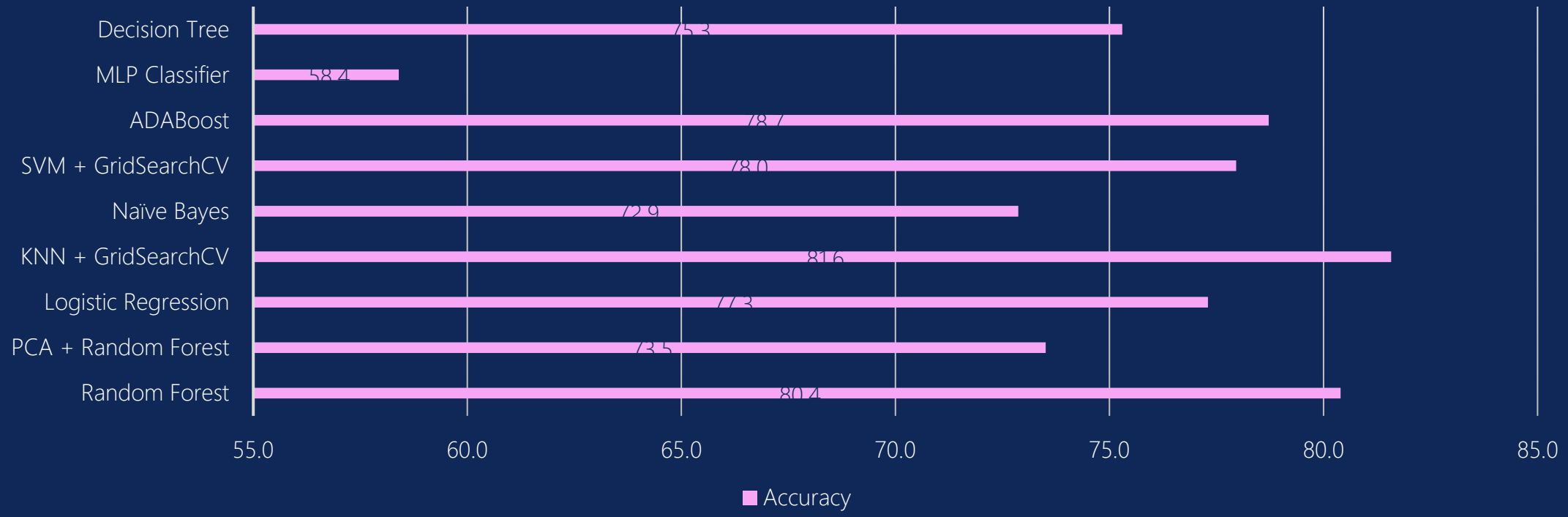
# KMEANS CLUSTERING




| cluster | shared_interests_partner | guess_prob_liked | like | expected_num_matches | attractive_partner | funny_partner |
|---|---|---|---|---|---|---|
| 0 | 0.548488 | 0.511061 | 0.604300 | 0.163273 | 0.602664 | 0.639063 |
| 1 | 0.369828 | 0.396658 | 0.361649 | 0.156358 | 0.351432 | 0.435232 |
| 2 | 0.686517 | 0.618132 | 0.795423 | 0.214799 | 0.824670 | 0.786117 |

# COMPARATIVE PLOT



# THANK YOU