Dating App Analysis
DS2500 - Team Project Final Report

Team Members:

- Rithika Ravichandran (ravichandran.ri@northeastern.edu)
- Sonali Chainani (chainani.s@northeastern.edu)
- Yonglin Yang (yang.yongl@northeastern.edu)
- Amir Ablassanov (ablassanov.a@northeastern.edu)

Section: 3
Date: December 9, 2025

Introduction

Goal of the Project

The goal of this project is to analyze how demographic and preference-related factors relate to user engagement outcomes on dating apps, such as matches, conversation length, and ghosting.

Problem Statement & Background

Online dating has become central to modern relationships, with 42% of U.S. adults saying dating apps make it easier to find a long-term partner (Forbes, 2025). While these platforms expand access to potential partners beyond immediate social circles, they also present challenges. Many users experience ghosting, mismatches, or conversations that fizzle out, creating stress, disappointment, and insecurity. Because dating platforms have become a major way people form connections, especially among younger adults, understanding what predicts positive outcomes is increasingly important.

This project looks at how demographic and preference-related factors, such as age, gender, education, and age-filter settings, relate to engagement on dating apps. Using an anonymized dataset from a dating platform, our goal is to uncover patterns that explain which groups experience higher engagement and which behaviors or characteristics tend to correlate with better outcomes. These findings not only give users more realistic expectations, but they may also be useful for designers of dating platforms who want to reduce negative experiences like ghosting and create features that support healthier digital interactions.

Why This Matters

This issue matters because online dating now affects millions of people's romantic lives, with significant emotional consequences. A 2023 Forbes Health survey found that 76% of dating adults have experienced ghosting, with many reporting upset or insecurity afterward. Studies also show higher levels of stress, depression, and psychological distress among dating app users compared to non-users. When negative experiences become common, the social and emotional costs can affect self-esteem and mental health. Understanding what influences engagement could help users navigate online dating more confidently and guide designers toward features that reduce harmful experiences like ghosting, improve emotional outcomes, and enhance user well-being.

Dataset

Dataset Overview

Our analysis uses a European Dating App Behavior Dataset, which is an anonymized dataset sourced from GitHub. It contains real conversation data extracted from Tinder user interactions and includes 395 user profiles with 16 features stored in a CSV format. The data has no specific time period and is composed of behavioral metrics such as the number of conversations, average conversation length, number of matches, and ghosting frequency after first messages, as well as several other metrics. It also includes demographics such as age, gender, education level, country, and more.

Data Collection Methodology

The data was originally collected by extracting and aggregating anonymized user data from Tinder conversations. The most likely collection method involved users downloading their personal data from Tinder, which the platform allows. These individual data exports were then aggregated across multiple users and transformed into summary statistics (such as average conversation length, total matches, ghosting counts) rather than containing raw message content or identifiable information. The processed and anonymized dataset was subsequently published on Kaggle, and then someone used that data to create another dataset, which is what we retrieved from GitHub. However, the exact details of the collection process, such as surveys, sensors, or web scraping, are not documented in the dataset information and are therefore unattainable for us.

Key Variables

Variable Name	**Description**	**Type**	**Example Values**
number_of_matches	Total number of matches a user has had	Numeric	66, 112, 3408
number_of_conversations	Total conversations a user has participated in	Numeric	739, 82, 173
average_conversation_length	Average number of messages exchanged per conversation	Numeric	8.56, 12.60, 11.69
number_of_ghostings_after_1st_message	Number of times a user stopped responding after first message	Numeric	66, 1, 23
ghosting_rate	An engineered feature! The proportion of conversations that resulted in ghosting after first message	Numeric	0.089, 0.012, 0.133
age	User's age	Numeric	20-51 years
gender	User's gender	Categorical	Men, Women
education	User's accomplished educational level	Categorical	"Has no high school or college education", college education levels
minimum_of_age_filter	Lower bound of user's age preference range	Numeric	18-46
maximum_of_age_filter	Upper bound of user's age preference range	Numeric	19-95
country	European country where user is located	Categorical	Germany, Sweden, Switzerland
interests_in_gender	User's gender preference for matching	Categorical	"Interest in Men", "Interest in Women", "Interest in Men and Women"
age_filter_range	An engineered feature! The difference between max and min	Numeric	Calculated as (max - min)

	age filters		

Ethical Considerations

While personal identifiers were removed during anonymization, privacy concerns regarding informed consent still remain since it's unclear whether users have explicitly agreed to share their data for research purposes, and for it to be viewed publicly.

This dataset contains several notable biases. Geographically, it's limited to European users, restricting generalizability. The sample represents only active users willing to share their data, and the small size of 395 users likely underrepresents certain demographics. Additionally, key variables are absent, including message content quality, profile attractiveness, and actual dating outcomes, which limits the depth of possible behavioral analysis.

Methods

Data Preprocessing/Preparation

Our data cleaning, preprocessing, and preparation focused on ensuring data quality and consistency. Our transformations included standardizing column names to lowercase with underscores replacing spaces and hyphens to ensure consistency across all data processing steps. Outliers beyond 3 standard deviations (|Z-score| ≥ 3) numerical features were removed to prevent skewed analysis and model performance. We examined missing values across all features and confirmed minimal missing data, requiring no imputation in this dataset. Lastly, continuous variables were standardized using StandardScaler to achieve zero mean and unit variance, ensuring fair comparison across features with different scales.

Exploratory Data Analysis

Our initial exploration revealed several key patterns in dating app user behavior. We noticed that age distribution analysis showed a concentration of users between 20-35 years old, with fewer users in younger and older age groups. Gender distribution was relatively balanced but showed behavioral differences in matching patterns. We also discovered that users with higher education levels exhibited significantly higher ghosting rates after initial messages, suggesting education may influence communication expectations or selectivity. Preliminary analysis revealed a substantial gender gap in match rates, with women receiving more matches on average than men, confirming broader social dynamics within dating platforms. Lastly, users with wider age filter ranges (greater difference between minimum and maximum preferred age) tended to have higher match counts, indicating that openness in preferences correlates with matching success.

Modeling & Evaluation

We employed a Random Forest Classifier for our binary classification task of predicting "high-match" users because it handles non-linear relationships well, provides inherent feature importance rankings, is robust to overfitting compared to single decision trees, and performs well with mixed feature types (continuous and categorical after encoding). We implemented stratified 80/20 train-test splitting to maintain class distribution consistency. To demonstrate methodological rigor, we adopted a dual-feature approach. The safe feature set included only demographic and preference data available at registration (age, age filter settings), while the leakage feature set added behavioral metrics (conversation counts, ghosting rates) to illustrate data leakage effects. We assessed model performance using accuracy, confusion matrices, and feature importance rankings, then compared the two feature sets to demonstrate leakage impact. We deliberately avoided using protected attributes like gender as predictive features, focusing instead on user-controlled preferences and behaviors to ensure ethical modeling practices.

Results & Conclusions

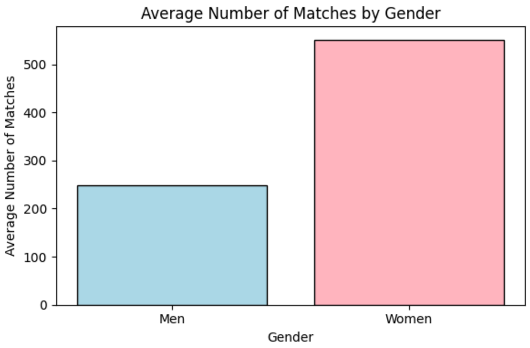
Key Findings

Finding 1: Women receive significantly more matches than men

Users identifying as women received more matches on average, even though men were the majority of the sample. This suggests that gender differences shape engagement outcomes.

Figure 1. Average Number of Matches by Gender

A bar chart comparing mean matches for each gender showing women receive more matches.

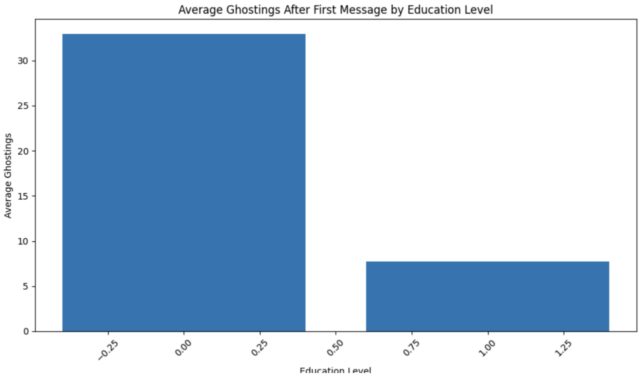


Finding 2: Users with higher education experience less ghosting

When comparing ghosting after the first message, users without higher education were ghosted more frequently. This indicates that education level may influence conversation continuation and early interaction quality.

Figure 2. Ghosting After First Message by Education Level:

A bar chart demonstrating that ghosting frequency decreases as education level increases.

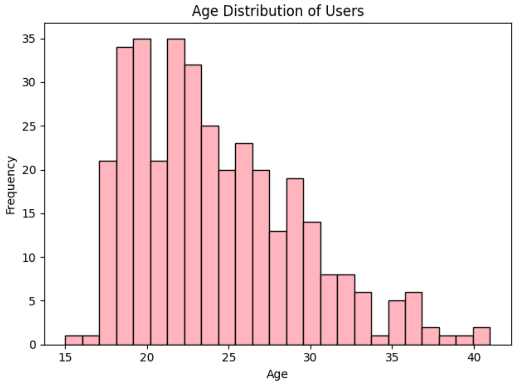


Finding 3: Age plays a smaller role in predicting engagement

Although younger users tend to be more active, age alone did not strongly predict match outcomes. This suggests demographic characteristics do not contribute equally to engagement.

Figure 3. **Age Distribution of Users**

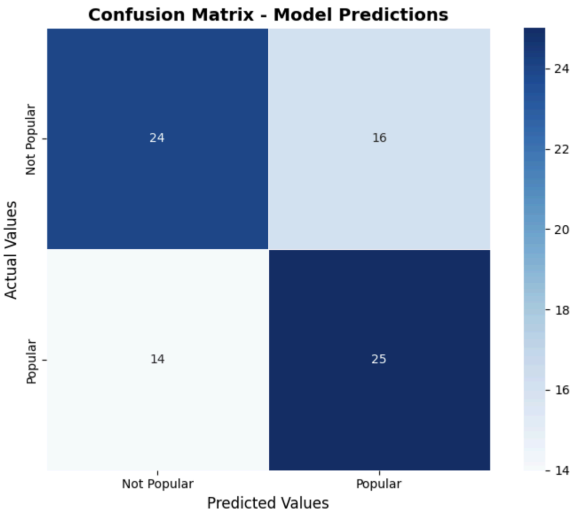
A histogram showing the dataset is mostly younger adults, but without a strong relationship to match success.

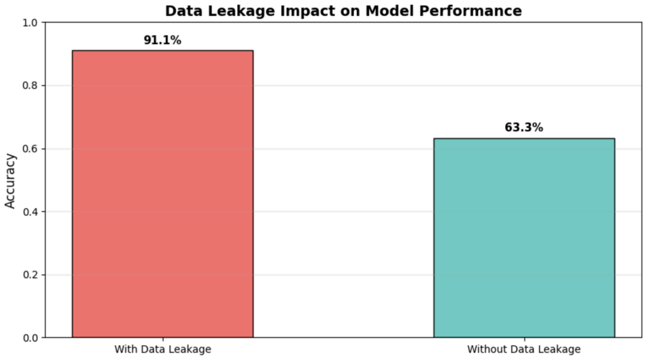


Model Performance

We trained a Random Forest classifier to predict whether a user would become a high-match user using only pre-interaction information (age and age-filter settings). The realistic model reached **67.6% accuracy**, showing moderate predictive power using controllable features. Models that included future metrics (like number of messages or ghosting) reached over 90% accuracy, demonstrating clear data leakage.

Metric	Value	Interpretation
Accuracy	67.6%	Predicts high-match users somewhat better than chance using only information available at signup





Conclusions

Returning to our main question of what patterns influence dating app outcomes, we found that engagement is far from random. Gender and education showed clear relationships with success, while age played a smaller role than expected. Women received more matches despite being outnumbered on the platform, and users with higher education experienced less ghosting after initial messages.

Our machine learning model achieved approximately 63% accuracy in predicting popularity using only pre-interaction features like age and age-filter settings. Higher accuracy only emerged when incorporating post-interaction data, revealing the impact of data leakage and highlighting the inherent limits of early prediction.

These results demonstrate that user characteristics and preferences fundamentally shape engagement patterns. Understanding these factors could help set realistic user expectations and inform app design decisions that promote healthier interactions.

Future Work & Limitations

Limitations of Current Analysis

This dataset only includes European users, limiting generalizability to other regions or platforms. The anonymized data lacks detail on key variables, message content, personality, appearance, and geographic distance, which likely affects outcomes. Our measures of ghosting and engagement are proxies rather than direct observations and may oversimplify complex interactions. Finally, our predictive model used only pre-match features to avoid data leakage, which maintains realism but reduces accuracy and prevents causal inference.

Future Research Directions

Future work could look at users across different countries or platforms to see whether similar patterns appear in other dating environments. Access to richer data, such as message timing or profile information, could help explain engagement more accurately and reveal emotional factors that numbers alone can't capture. It would also be valuable to explore algorithmic fairness and whether certain groups consistently experience lower engagement. In the long term, models could be developed to identify healthy interaction patterns or guide app designs that reduce negative experiences like ghosting.

References

Phares, Emily & Harmon, Meaghan. (2025). Dating Statistics and Facts in 2025. Forbes Health. https://www.forbes.com/health/dating/dating-statistics/

OnePoll / Forbes Health. (2025). Ghosting and Dating Survey Data. Forbes Health. https://www.forbes.com/health/dating/dating-statistics/

BetterHelp. (2023). Dating Stress Survey. BetterHelp Online Therapy. https://www.betterhelp.com

Sevi, B., et al. (2020). Swipe-based dating applications and psychological distress. BMC Psychology. https://bmcpsychology.biomedcentral.com

Dating_Apps_datasets. (2025). GitHub. https://github.com/WiktoriaGolebiewska/Dating-Apps/blob/main/Dating_Apps_datasets/df_europe_clean.csv