**Group project 15**

Submitted by

Dinesh Chandra Gupta Malepati (11615792)

Sai Kumar Pathipati (11600326)

Nikhil Rachagani (11596241)

Sai Manikanta Praneeth Perala (11614202)

**Analyzing the OkCupid Profiles Dataset to Predict Relationship Status**

**Backstory and Introduction**

During the recent Valentine’s week, a question intrigued me. Seeing many couples roaming around, I could see most variety crowd in relationships. So, what can be a deciding factor to be in a relationship and what characteristics play important role to effect the relationship status of a person.

An increasingly common method for people to find possible romantic partners is through online dating services. Given the popularity of these platforms, it is crucial to comprehend user behavior and preferences in order to match people more effectively. In order to forecast users' relationship status based on their profiles, we will examine the OkCupid Profiles Dataset in this project.

**Dataset**

<https://www.kaggle.com/datasets/yashsrivastava51213/okcupid-profiles-dataset>

On Kaggle, we found the OkCupid Profiles Dataset, which is what we'll be utilizing. More over 68,000 OkCupid user profiles are included in this dataset, along with 31 attributes such user demographics, hobbies, and lifestyle. To protect user privacy, the information was collected from OkCupid in 2016 and anonymized. Predicting the user's relationship status—which may be "single," "seeing someone," or "married"—is the goal of our endeavor. Our investigation will center on this target variable.

Age: the user's age

status: the user's profile's current state (e.g., single, seeing someone, married, etc.).

gender: the user's gender.

orientation: The sexual preference of the user (e.g., straight, gay, bisexual, etc.)

drinks: How frequently the individual consumes booze.

drugs: the drug use of the user.

Height: the user's height, measured in inches.

job: The user's position or line of work.

location: the location of the user.

pets: The preference of the user for pets.

Smoke: the user's smoking behavior.

body type: The body type of the user (e.g., thin, overweight, average, etc.).

diet: The nutritional choices of the user.

education: The user's degree of education.

Race: the race or ethnicity of the user.

Religion: religious convictions of the user.

Zodiac sign: the user's zodiac sign, or sign.

Speaks: languages that the user is fluent in

income: The user's current income.

Out of all the entries we can use up to 1500 entries for our project. We have many variables deciding the factors out of which the most influential variables for determining the relationship status such as income, location, job, height, weight and language are analyzed for our project.

**Procedure**

The data must first be cleansed and processed to verify its correctness and completeness before any analysis can be performed. Descriptive statistics are then used to compute and summarize crucial statistical variables like means, standard deviations, and correlations in order to gain a better understanding of the data. The next step is to do exploratory data analysis to find patterns in the data and show correlations between variables. To create a prediction model that can show how changes in one variable impact other variables, regression analysis is utilized. The validity and usefulness are assessed based on the regression analysis's findings. Finally, conclusions are formed in light of the research topic and the data.

Data Cleaning > Descriptive statistics> Exploratory Data Analysis > Regression Analysis> Outcomes> Conclusion

**Expected Outcomes and Conclusion**

The goal of this project is to create a model that, based on a user's profile's qualities, can reliably predict that user's relationship status. The strategy would allow online dating sites to provide their members more accurate matches and suggestions. The project's effects would improve the website's user experience, user engagement, and retention rates. In order to predict users' relationship status based on their profile traits, we will examine the OkCupid Profiles Dataset in this project. We will combine machine learning algorithms with data analytics techniques to build a model that accurately predicts the outcome variable. The project's results will help online dating businesses improve their user experiences, increase user retention, and make better suggestions to their customers.