ML Match

Entry 1:

Jessica Isunza, Read data from .csv

05/18/24

We started following the steps we had learned earlier in our Artificial Intelligence on videogames class, first step was to read the csv file which was the easiest step, we decided to keep the file as it was even if we knew we wouldn't need some columns since we wanted to do the data cleaning through code.

Entry 2:

Jessica Isunza, Display information

05/18/24

To be able to see what the information in the csv file meant we started by displaying it on tables and charts, that way we found that there was numerical and non-numerical data, this later led to some problems since we had to adjust and convert some values.

Entry 3:

Jessica Isunza, Delete missing data

05/18/24

When we displayed the information in tables we discovered that some of the participants did not filled the form completely so we had to check for null values in every column and delete the entire participant row, since we could not have null values and there were fields that were very important to include and having all those missing values would have ruined the algorithm and the final results. We had some problems with this because sometimes we

would delete the entire column by accident leaving us with almost no values, we fixed this by accessing some columns separately in order to prevent affecting other columns.

Entry 4:

Jessica Isunza, Divide data (training/testing)

05/18/24

As we learned from a previous project, we needed to separate data for testing and for training, we separated the features (X) and the target variable (Y) to evaluate our model's performance. After that, we created a preprocessing pipeline to handle missing values and scale numerical data, we filled NaN values with scale features from 0 to 1. We applied this pipeline to the training and testing sets to clean and normalize our data and we converted the processed data back into DataFrames, to make sure it was working we printed the first rows.

Entry 5:

Jessica Isunza & Lucía Castañeda, Choose data and clean unnecessary data

05/20/24

Since the beginning of the project we decided that we would not need the columns "prob" which referred to Rater's rating on whether they believed that interest would be reciprocated, rated 0-10 and "met": Whether the two individuals had met prior to the speed date, coded as 0 or 2, with 0 indicating they had not met before, since our goal was to predict the match probability a new individual would have to be rated as liked or disliked based on our previous data.

Entry 6:

Lucía Castañeda, Read data from drive

05/20/24

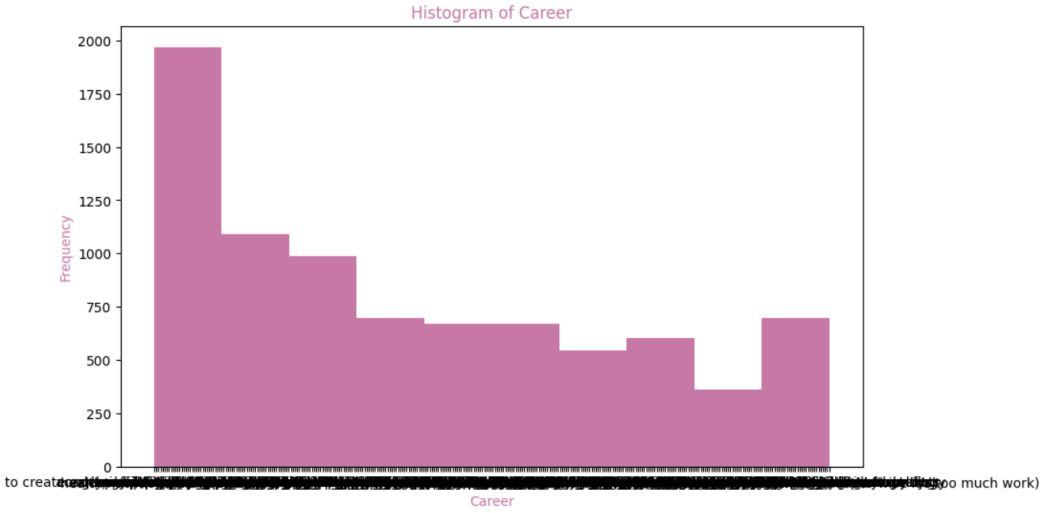
We decided that it was best to call the csv file from drive than adding the file every time we opened collab or every time the servers were down due to inactivity, we did not had problems in this part since it was similar to the first time we read the data, after that we displayed the information to see if it was working correctly.

Entry 7:

Lucía Castañeda, Delete faulty data

05/20/24

We noticed there were many inconsistencies in the column “career” since some of the participants didn´t answered correctly, at first we tried to correct miss spellings and same professions written differently but we found that some of the participants filled the fields with things like “scientific research for now, but who knows”, “not decided yet” and “not sure”, this made the data cleaning process harder.



After trying to correct the data in the column and displaying it we noticed that there were too many categories to be displayed in any graph so we decided to delete that column since we also noticed that what really matters for an individual to consider choosing another was related to looks and income and not their career.

Entry 8:

Lucía Castañeda, ML classification models

05/20/24

We used three types of binary classification models, our focus was on binary classification due to the outcome being either you were matched or you didn't find a match. The three models used were logistic regression,

decision trees and random forest.

We measured the accuracy of each model, the best one was random forest.

Entry 9:

Lucía Castañeda, Most important data weights

We made a correlation matrix and found out surprising insights on the values and how they were related. Like sincerity and intelligence.

We also found out that attractiveness was a high weight that affected the model, while intelligence and sincerity were one of the lowest.

05/20/24

Entry 10:

Lucía Castañeda, Extra testing

We created extra code for testing made up people or if someone wants to know if they can get a match with the people from the data. We found out that someone with high income but low attractiveness was not able to get a match.

05/20/24

Entry 11:

Jessica Isunza & Lucía Castañeda, Visualization

05/20/24

After we finished we decided to change the colors of our graphs and tables, since we had the default colors and we didn't like them we decided to investigate how to personalize graphs and tables to be pastel colors and pink themed, for these we searched in many stack overflow forums. (example references below)

Conclusions:

We spent a significant amount of time refining the data and determining what would be useful and what wouldn't. One of the biggest challenges was deciding how to handle missing or incorrectly filled data. Ultimately, we chose to remove rows that were practically useless due to this issue.

We were surprised by the relationships we found among some attributes, such as attractiveness and its influence on decision-making, as well as sincerity and intelligence, which are often scored lower. We also discovered that individuals with lower intelligence tend to answer questions incorrectly or leave fields blank, resulting in missing data.

We learned that the most important and challenging aspect of predicting data is cleaning it and establishing correlations to determine the desired outcome and the type of machine learning model to use.

We had fun doing this analysis and inputting people at the end to see if they are capable of finding a match with the given data

Jessica: I learned that choosing the database is not the most important thing, i used to think that if a database was not suitable for the project you had to search for another one, i learned that data needs to be cleaned and some values need to be converted in order to be used

correctly, i also learned the importance of all the process that we need to do before even thinking about the main Machine Learning algorithm, i think that the visualization part of this course was very useful for this project since it was easier to know what to do and what to use after displaying the information as tables and diagrams.

Lucy: I learned a lot about managing data and the importance of not skipping steps while analyzing everything since a simple missing value or a badly formatted one can change the model drastically and affect its accuracy. Even though our accuracy was 76% I believe it is better considering binary classification gives you a 50% naturally. This project helped us learn about data cleaning, managing and analysis as well as machine learning models and how to use them and about dating preferences.

References:

- *How to change the plot line color from blue to black*. (n.d.). Stack Overflow. https://stackoverflow.com/questions/41709257/how-to-change-the-plot-line-color-from-blue-to-black
- *How to change a the colors of specific rows and columns in a table?* (n.d.). Stack Overflow. https://stackoverflow.com/questions/67245827/how-to-change-a-the-colors-of-specific-rows-and-columns-in-a-table
- *How to change plot background color?* (n.d.). Stack Overflow. https://stackoverflow.com/questions/14088687/how-to-change-plot-background-color
- *Changing text color in cells for a table generated with matplotlib*. (n.d.). Stack Overflow.

https://stackoverflow.com/questions/64610994/changing-text-color-in-cells-for-a-table-generated-with-matplotlib