OKCupid Project

Codecademy

Data Science Career Path

2021

Project Goals:

"Utilize the skills learned through Codecademy Career Path applying the machine learning techniques to the data set, creating a classification algorithm from the supervised learning models to predict the OKCupid users' zodiac signs"

Exploratory Data Analysis.

DATASET CHARACTERISTICS:

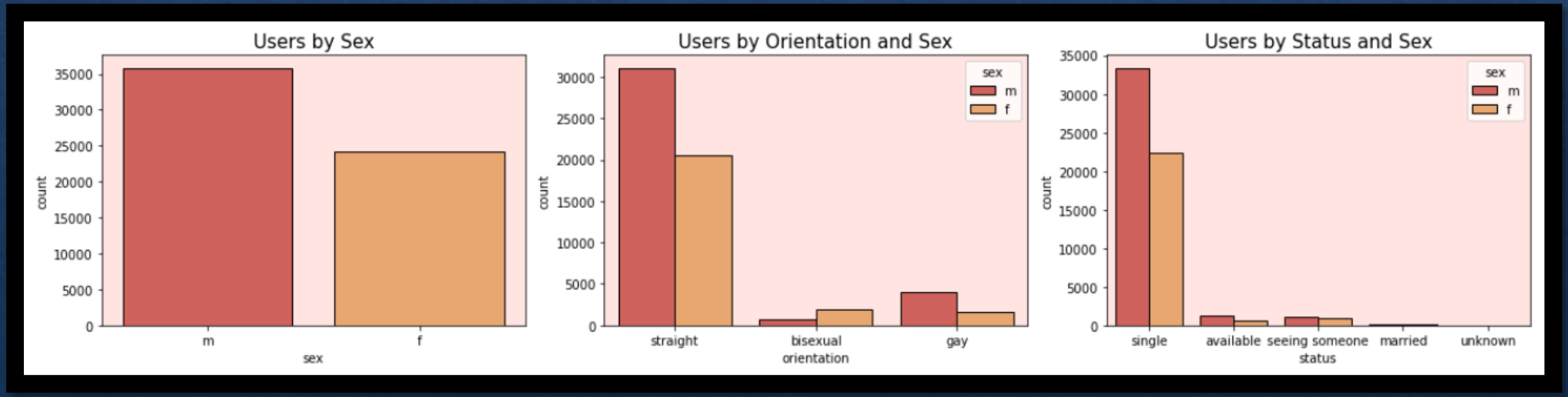
- The dataset presents 59.946 instances in 31 categories of different users. Some of these attributes are age, sex, education, job, sex orientation, zodiac signs, etc.
- Except by the variables Age, Height, and Income, all the rest features are categorical variables.
- The dataset presents several missing values.

SEX USERS, SEXUAL ORIENTATION & STATUS

First plot: it is possible to see that there are more males than females in the data.

Second plot: There are more straight people than bisexual or gay. But, in the same times, it is possible to notice that there are more bisexual females and more male gays.

Third plot: There are more people single than the other states in the dataset.



DIETARY INFORMATION & DRINKS, DRUGS AND TOBACCO CONSUMPTION.

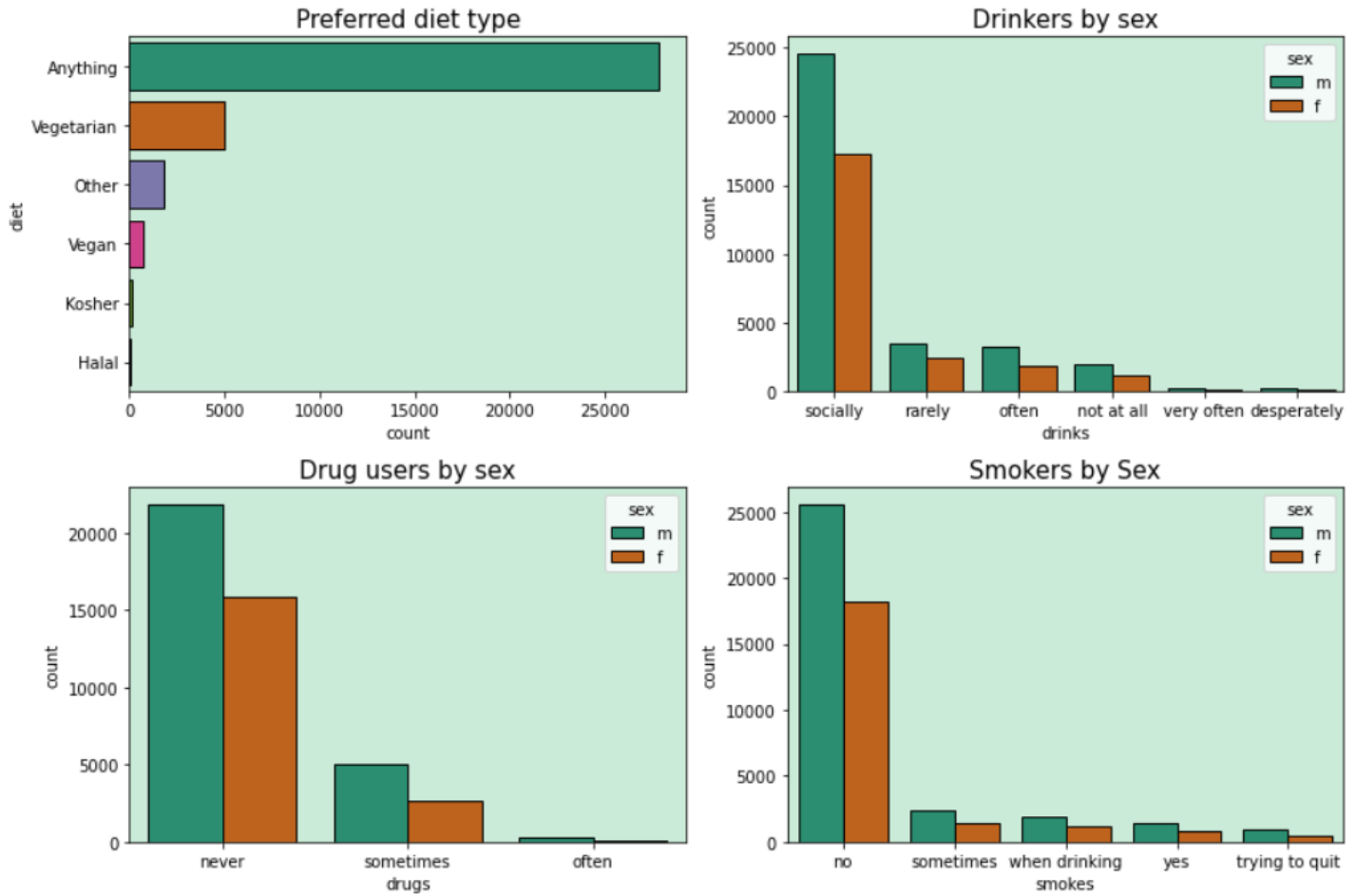
Dietary information: Most user eat "Anything", followed by "Vegetarian".

Consumption of alcoholic beverages: most of the users are "Social drinkers".

Consumption of drugs: most users "never" use drugs.

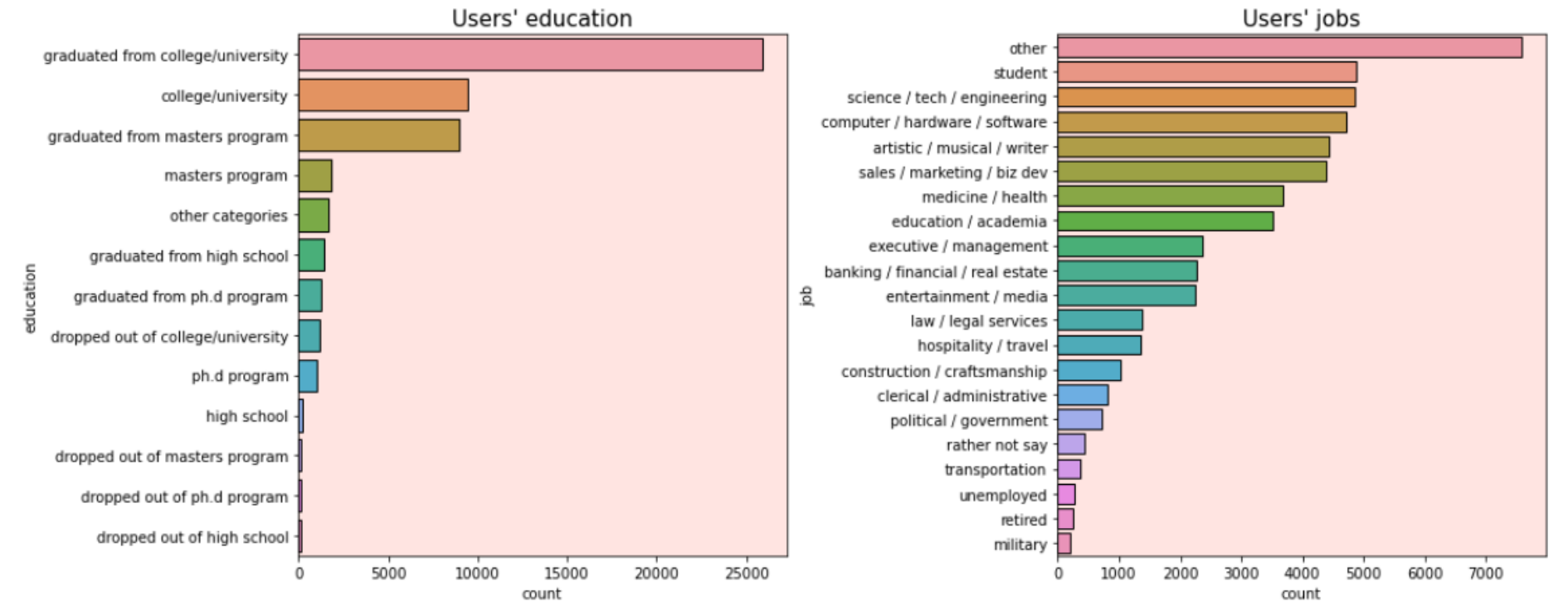
Smokers: Many users chose "no" for smoking.

In addition, we detected that, in all categories, the proportion of males are higher than the females.



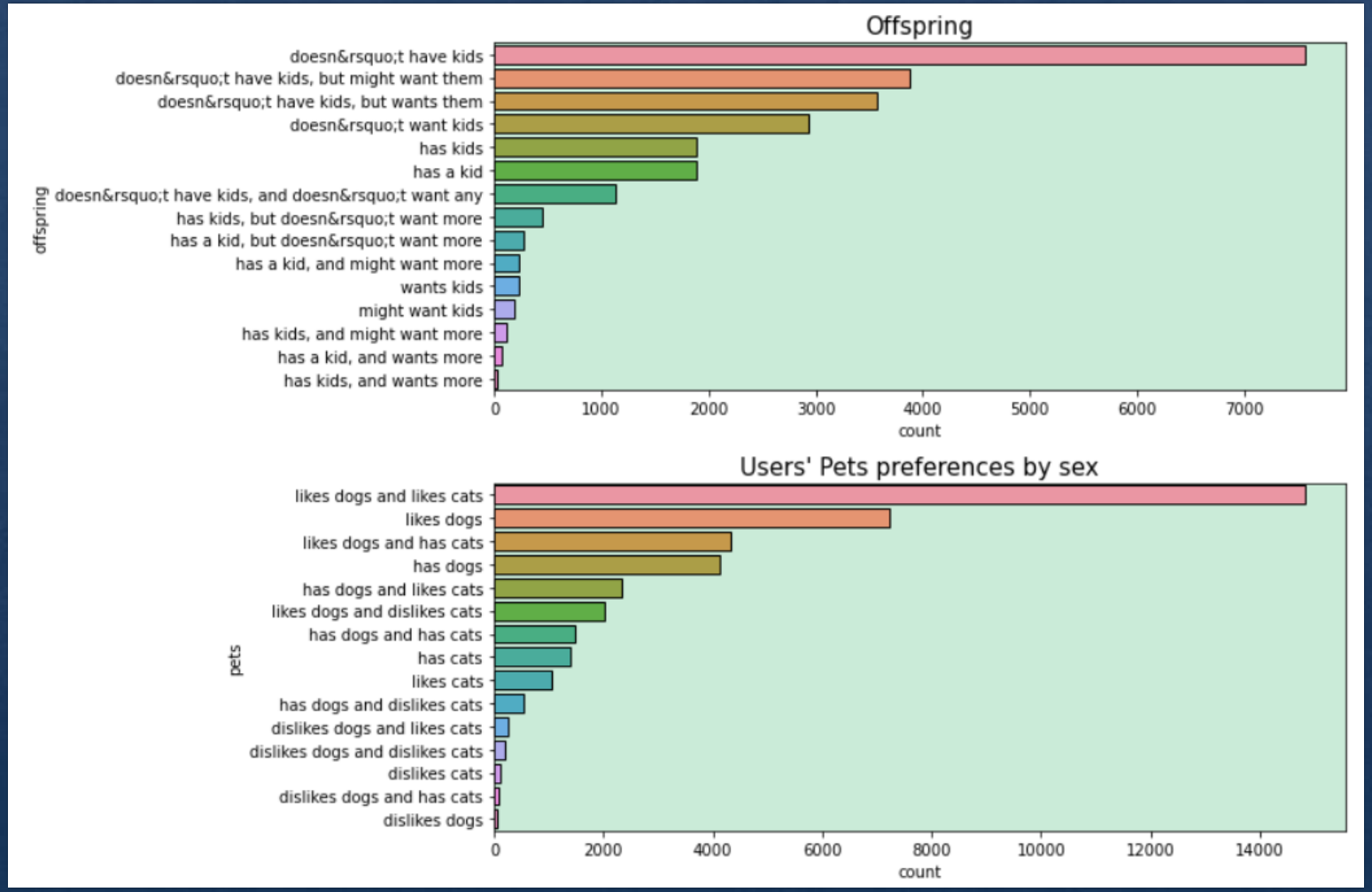
EDUCATION & JOBS

- **Users's Education:** Many of users are **graduate from college/university** followed by people who are still in college/university and people who is graduated from master programs.
- **Users' jobs:** Most users don't fit into the categories provided, but there are a fair share of students, tech, computer, artists and business folks.

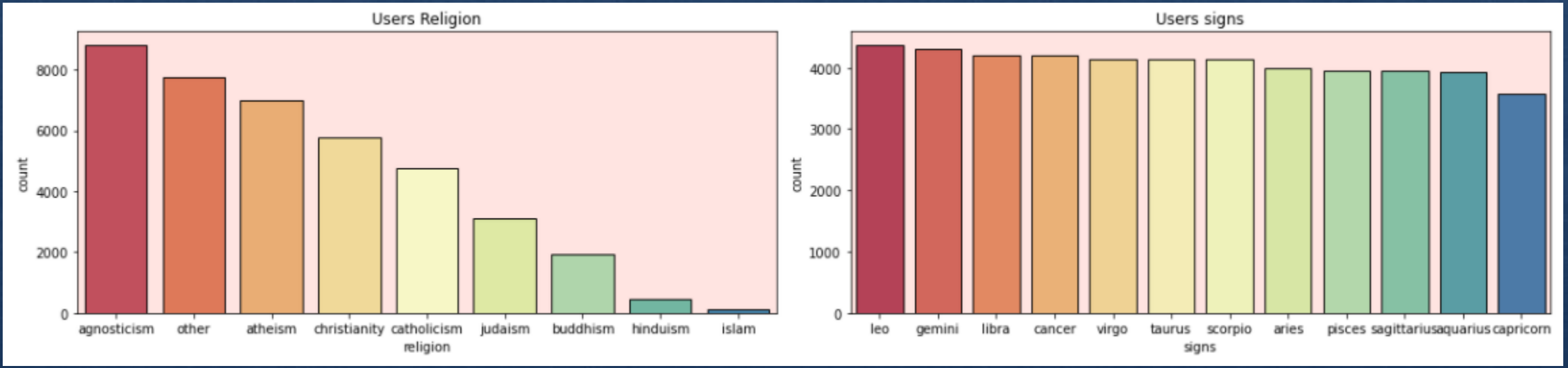


OFFSPRING & PET PREFERENCES.

- Offspring: Most users do not have kids follow people who does not have kid, but they probably want them.
- Pet preferences: the users likes dogs and cats.



USERS' RELIGION AND ZODIAC SIGNS



In the users' religion graph, we notices that the majority was not very religious, identifying themselves as agnostic, other, or atheists.

Looking at users' signs, there are mainly evenly distributed. However, Capricorns users are the rarest in the dataset and Leo users are the most common

MACHINE LEARNING MODELS.

To try find the best way to predict the zodiac sign of the people in the dataset, I will proceed to train two classification models:

- KNeighborsClassifier
- DecisionTreeClassifier Model.

MACHINE LEARNING MODELS.

KNN Model.

Using the default hyperparameters values, this model had a 10% accuracy in the training dataset and 8% accuracy in the testing dataset. These values are not a good sign what means that the model does not capture any logic.

Train Prediction Report

	precision	recall	f1-score	support
aquarius	0.10	0.16	0.12	2886
aries	0.08	0.35	0.14	2908
cancer	0.09	0.32	0.14	3099
capricorn	0.15	0.04	0.06	2612
gemini	0.10	0.10	0.10	3180
leo	0.18	0.03	0.05	3216
libra	0.16	0.02	0.04	3082
pisces	0.10	0.05	0.07	2867
sagittarius	0.14	0.03	0.05	2843
scorpio	0.18	0.03	0.04	3013
taurus	0.24	0.02	0.04	2989
virgo	0.17	0.03	0.05	3031
accuracy			0.10	35726
macro avg	0.14	0.10	0.08	35726
weighted avg	0.14	0.10	0.08	35726

MACHINE LEARNING MODELS.

Train Prediction Report

	precision	recall	f1-score	support
aquarius	0.19	0.09	0.12	2886
aries	0.14	0.12	0.13	2908
cancer	0.14	0.18	0.16	3099
capricorn	0.25	0.06	0.09	2612
gemini	0.20	0.08	0.12	3180
leo	0.11	0.33	0.16	3216
libra	0.11	0.32	0.16	3082
pisces	0.26	0.03	0.06	2867
sagittarius	0.16	0.06	0.09	2843
scorpio	0.14	0.09	0.11	3013
taurus	0.19	0.07	0.10	2989
virgo	0.13	0.10	0.11	3031
accuracy			0.13	35726
macro avg	0.17	0.13	0.12	35726
weighted avg	0.17	0.13	0.12	35726

Decision Tree Model.

Training a model with the default hyperparameters, I got a tree model has a depth of 18 branches.

Besides, in this model, the accuracy is not so different from the accuracy of KNN model. However, it is higher.

The optimization of the hyperparameters of this model has not changed much the results obtained.

CONCLUSIONS & NEXT STEPS.

- The trained models could not capture the logic inside the chosen data.
- The trained models are underfitted.
- Next steps will be to seriously consider whether it is possible to predict user's astrological signs, or if there is a way to do it aggregating more data.
- Another possibility is to add additional models to see if any more predictive power could be squeezed out of the algorithms.