An Analysis of OKCupid User Data

Pavlo Baranchuk

Outline

About data

Exploration data

Tests

Idea of model

My results

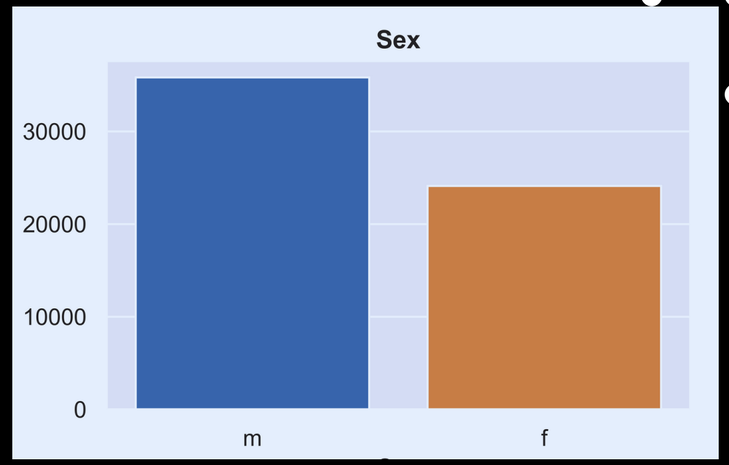
Conslusion

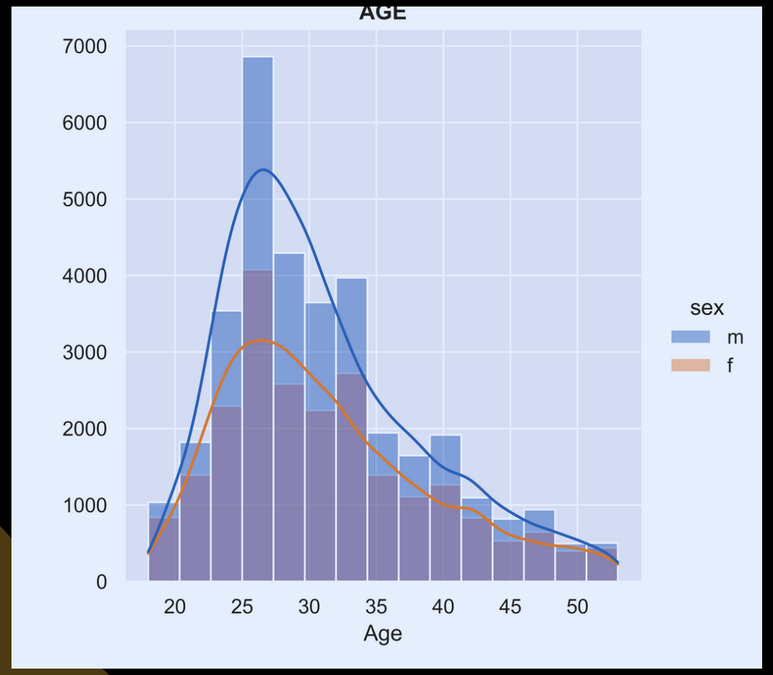
About data

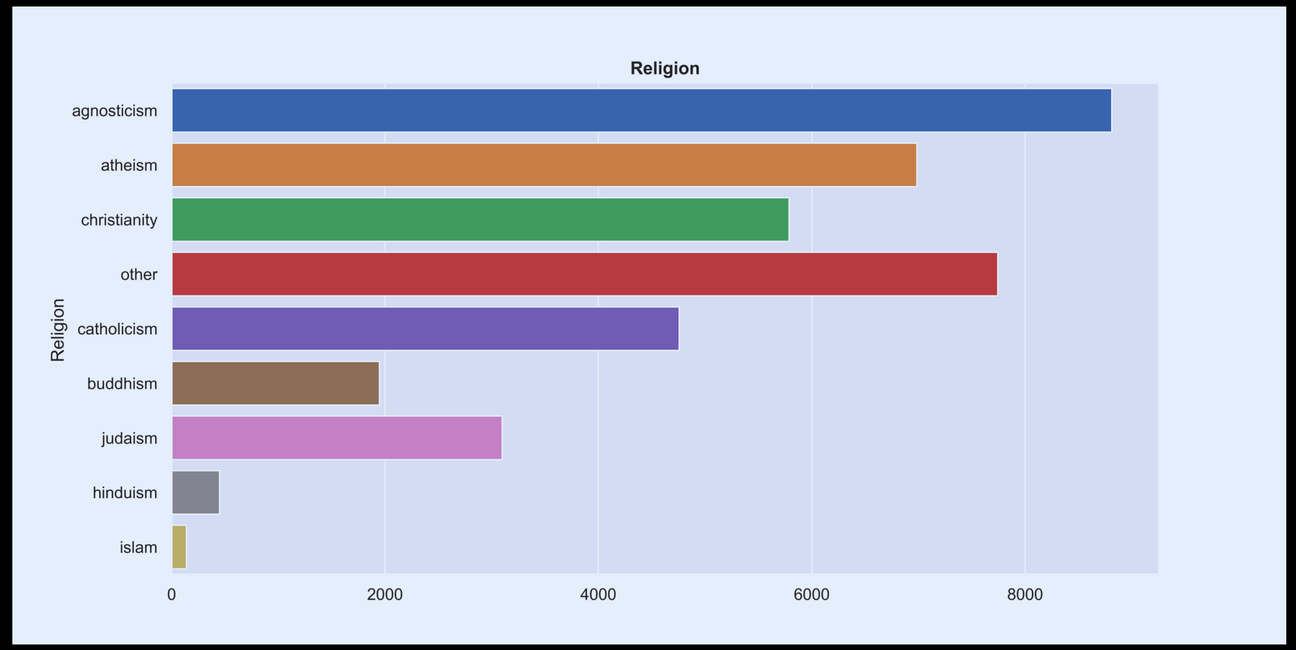
The project will use one dataset provided by Codecademy called profiles.csv. The data comprises user profiles, with each row representing an OkCupid user and the columns consisting of their responses to multi-choice and short answer questions.

Exploration data

In this project, I performed extensive data exploration to gain insights into the dataset. I used various visualization techniques to understand the distribution of features, their relationship with each other, and their correlation with the target variable.







Tests

Age

Was tested in this project is whether there is a statistically significant difference in the age between male and female users on a social network. However, I failed to reject the null hypothesis and concluded that there is no significant difference in the mean age between men and women on the dating app.

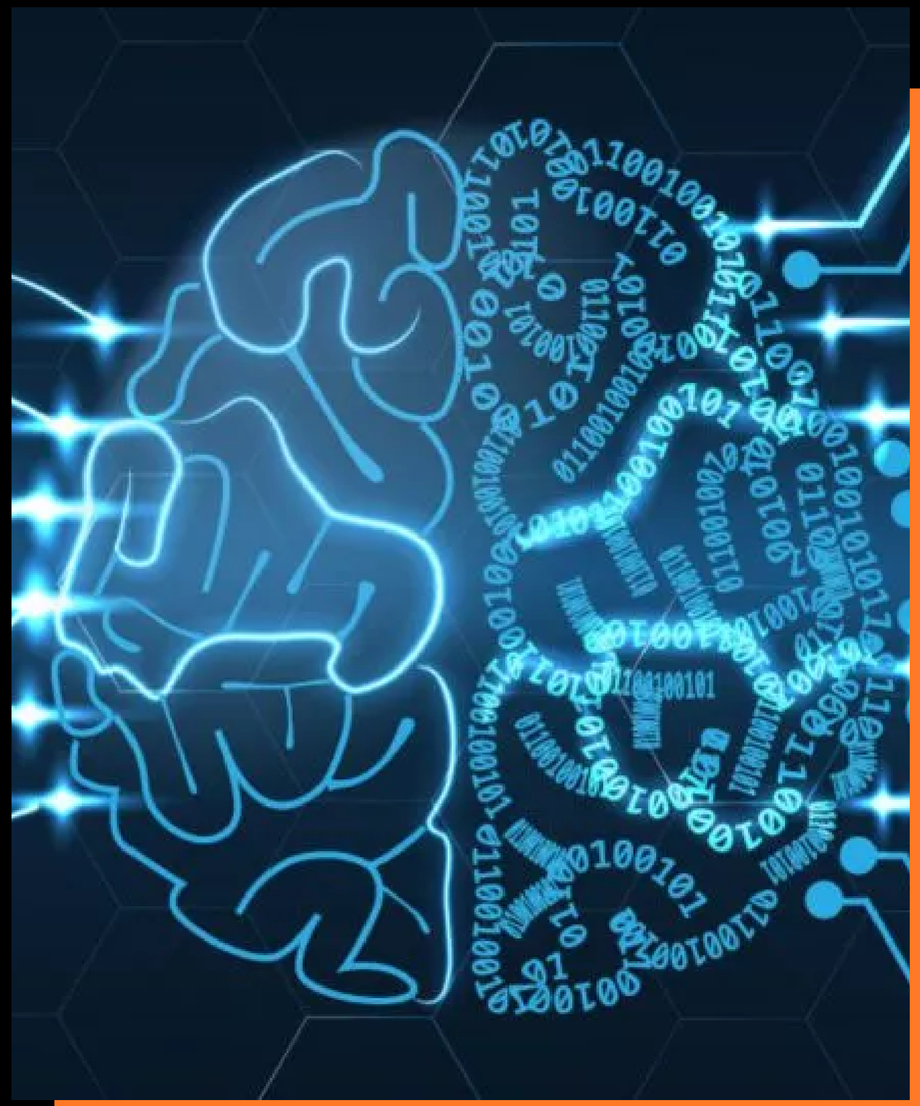
Sex and drugs

I found a significant association between gender and drug use using the chi-squared test, with men more likely to report drug use than women. These results have implications for public health and drug prevention efforts.

Idea of model

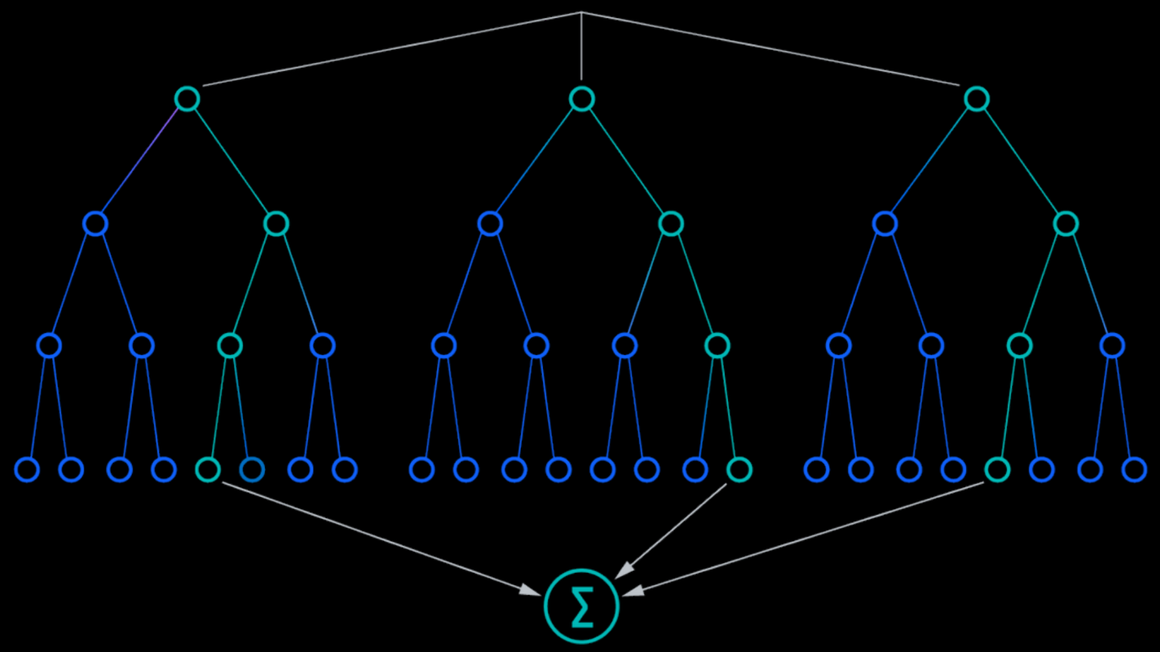
In addition to exploring patterns and relationships within the OKCupid dataset, we wanted to see if it was possible to use the data to predict a user's religion.

I selected a set of columns that I believed could be relevant in determining a user's religious affiliation, including body type, diet, sexual orientation, and job, among others.



Random Forest Tree

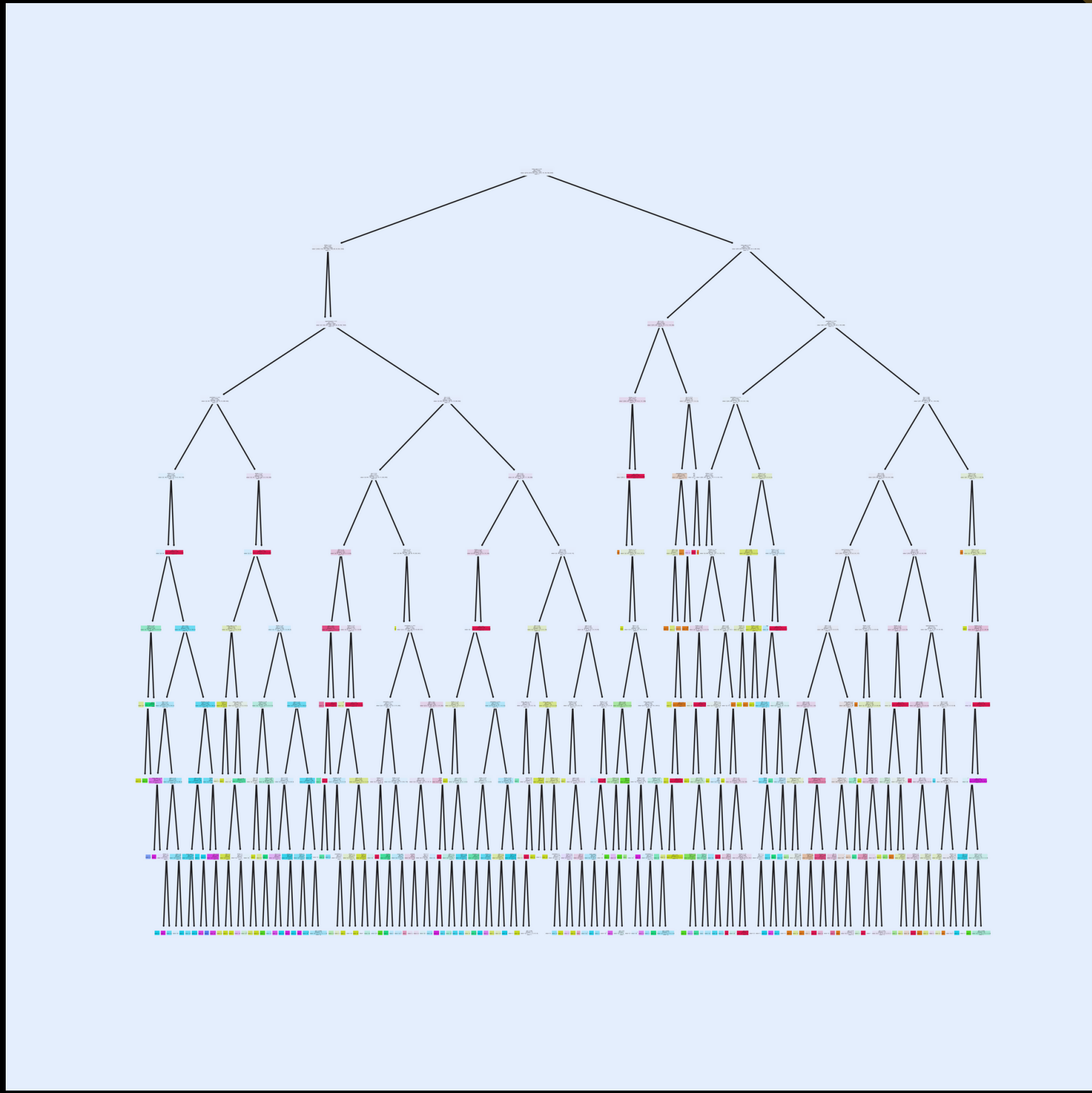
The Random Forest Tree model is a powerful machine learning algorithm that we used to predict religion based on various user attributes. It works by creating multiple decision trees and combining their results to make predictions.



My results

My Random Forest Tree model achieved an accuracy of 0.98, with high precision and recall for most classes. However, there were some issues with recall and precision for a few classes.

If you would like to learn more about our results and other models we tried, please refer to my Jupyter Notebook.



Conslusion

In conclusion, I have successfully developed a model to predict an individual's religion using their personal characteristics. My analysis showed that the Random Forest Tree model achieved an accuracy of 0.98, with strong performance across most classes. However, there were some limitations in predicting certain classes, indicating the need for further exploration and refinement of the model. Overall, this project highlights the potential for machine learning techniques to be used in understanding and predicting complex societal factors, such as religion, and provides a foundation for further research in this area.

Thank You