



# Predicting Zodiac Signs from OKCupid Profiles



A Machine Learning Exploration



# Project Goals and Scope

---

- Goal: Predict a user's zodiac sign using profile data
- Motivation: Zodiac signs are important to many users; predicting them could improve match recommendations
- Dataset: **profiles.csv** from Codecademy — 59,946 users, 31 columns

# What Are We Trying to Predict?

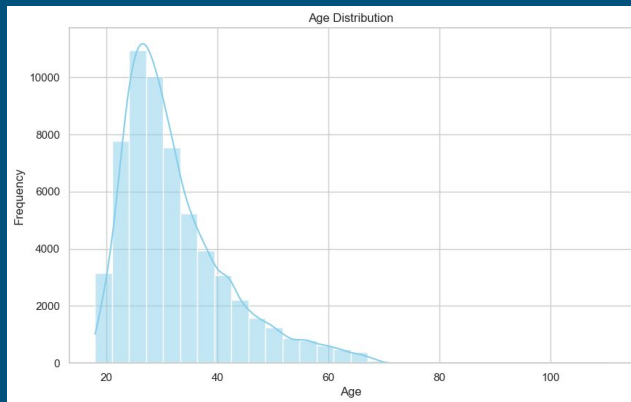
---

- Primary question: Can we predict a user's zodiac sign based on their profile responses?
- Approach: Classification models on structured + text features
- Challenge: Zodiac signs may not correlate strongly with lifestyle data

# Age Distribution by Gender

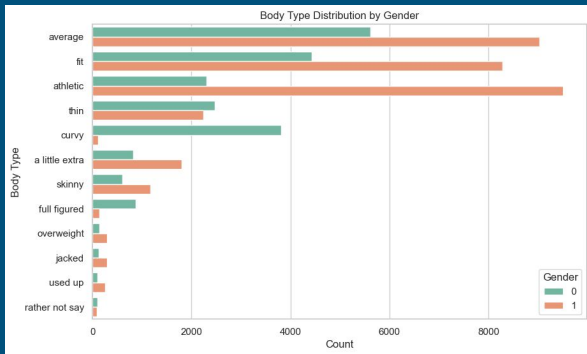
---

- Histogram showing age distribution
- Insight: Most users are in their late 20s to early 30s
- Gender breakdown shows slightly fewer females



# Body Type Preferences by Gender

- Bar chart of **body\_type** vs. **sex**
- Insight: Certain body type labels are gendered (e.g., “curvy” and “full figured” for women, “overweight” for men)
- Demonstrates categorical feature differences across genders



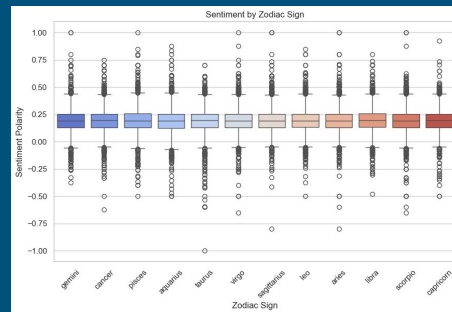
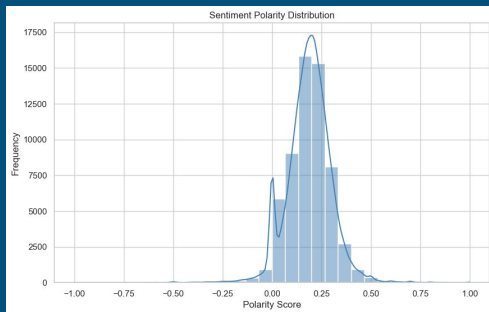
# Preprocess Data

---

- Handle missing values (drop/impute, treat lifestyle attributes as “unknown”)
- Encode categorical variables (Label/One-Hot Encoding)
- Standardize numeric features (**age, height, income**)
- Feature engineering:
  - Target = **zodiac\_clean**
  - Essays combined into **essays\_combined** for NLP vectorization

# Sentiment Features from Essays

- New column: sentiment (TextBlob polarity score)
- Graphs:
  - Sentiment distribution histogram
  - Boxplot of sentiment by zodiac sign
- Insight: Sentiment varies across users, but no strong predictive signal by zodiac



# Classification Model Comparison

---

- Models: Logistic Regression, KNN, Linear SVC, Naive Bayes, Random Forest (optimized)
- Runtime: Naive Bayes fastest ( $\sim 0.2s$ ), Logistic Regression slowest ( $\sim 7.5s$ )
- Accuracy: All models  $\sim 8-9\%$  (near random baseline)
- Show bar chart of precision, recall, F1-score
- Insight: No model significantly outperforms chance

# Final Insights and Takeaways

---

- Zodiac prediction is not feasible with structured + sentiment features
- Best model (Random Forest) achieved only 8.7% accuracy
- Importance of problem framing: not all questions are learnable
- Show confusion matrix for Random Forest

# Future Directions

---

- Add essay text features using TF-IDF or embeddings
- Reframe task: predict lifestyle traits instead of zodiac
- Explore unsupervised clustering to find dating archetypes
- Additional data needed: user preferences, match outcomes, personality scores