

Project Title: Machine Learning Analysis of Global Billionaire Wealth, Demographics, and Industry Patterns

Group Members: JP Daliere, Will Hoover, Stellan Kilpatrick, Jordyn Paszamant, Genevieve Weisensel, AC Zylka

Dataset Description:

We will use the [*Global Billionaires 2026: Wealth & Demographics*](#) dataset from Kaggle, which includes information on billionaire net worth, age, gender, country, and industry. The dataset provides a rich opportunity to apply supervised and unsupervised learning methods to understand global wealth patterns and demographic trends.

Research Questions: 1. Which demographic and industry variables help predict billionaire net worth? 2. How do simple supervised learning models compare in predictive accuracy? 3. Does feature engineering improve model performance? 4. Is age a meaningful predictor of net worth?

Variables of Interest: NetWorth (numeric), Age (numeric), Gender (categorical), Country (categorical), Industry (categorical), Rank (numeric)

Methods & Analytical Plan:

1. Data Preparation: Use pandas to load and clean the dataset, handle missing values, convert categorical variables to numeric (one-hot encoding), use NumPy for array operations, and use matplotlib for visualizations

2. Supervised Learning: We will model NetWorth using: Linear Regression (Baseline model for numeric prediction), k-NN Regression (Nonparametric comparison to linear regression and Tune k using validation error), Decision Tree Regression (Useful for identifying important predictors), Kernel Regression (Allows nonlinear smoothing of NetWorth vs. Age or other numeric predictors)

3. Classification: If we create a binary target such as Top 50 vs. Not Top 50, we can apply: Logistic Regression, k-NN Classification, Decision Tree Classification, and One-Class Classification to identify unusually wealthy individuals

4. Model Assessment & Tuning: Using the “Basic Practice” units: Train/validation/test split, Cross-validation for hyperparameter tuning, Compare models using RMSE (regression) or accuracy (classification), Discuss underfitting, overfitting, and regularization

5. Feature Engineering: Create region categories from the country, simplify industry categories, standardize numeric variables, and evaluate whether engineered features improve model performance

Most Important Line of Code (Data Import): `df = pd.read_csv("Top_100_Richest_Billionaire(2026).csv", index_col = 0)`

Expected Outcomes: We expect to identify clear geographic and industry-based patterns in billionaire wealth, as well as demographic trends that may reflect broader global inequalities. Our analysis will contribute to understanding how wealth is concentrated and what factors correlate with extreme financial success.