Methodology:

1. Social Security Number Data

Social Security Administration (SSA) [1] collects all names from social security card applications for births in the United States that occurred after 1879. It's worth noting that many people born before 1937 never applied for a Social Security card; thus, their names aren't in the database. We have collected these data from the SSA website to identify the gender demographics of the Twitter data by comparing the first name.

2. Census Data

The United States Census Bureau compiles annual estimates of race and Hispanic origin shares for each county in the country according to the respondent's surname. These estimates are based on the most recent decennial census as well as estimates of population changes (deaths, births, and migration) since that time. Respondents can choose from one of six racial groups on the census questionnaire: White, Black, or African American, American Indian, Alaska Native, Asian, Native Hawaiian, and Other Pacific Islander, or Other, which creates a wide range of variations. While race/ethnicity is a complicated topic, we simplify it for the sake of this study by focusing on only four categories: Asian, Black, Hispanic, and White. We use the 2010 estimates for this study [2].

3. Details Steps for Predicting Gender from First Names And Race From Last Names

Data Collection:

- The United States Census Bureau provides annual estimates of race and Hispanic origin shares for each county based on the respondent's surname.
- The data used for this study comes from a CSV file containing last names (for race prediction) and first names (for gender prediction) and their corresponding race/gender labels.

Data Preprocessing:

- The last names are read from the CSV file and loaded into a Data Frame.
- A function is defined to count the occurrences of each letter in a name.
- The letter counts are used to generate an alphabet matrix with additional race/gender labels for each name.

Feature Engineering:

- The alphabet matrix is constructed by converting letter counts into concatenated strings for each name.
- Each string is associated with a race label for four categories: Asian, Black, Hispanic, and White.
- Each string is associated with a gender label for two categories: Male and Female.

Model Training and Evaluation:

- The alphabet matrix is divided into feature matrix X (containing letter counts) and target vector y (containing race labels).

- The data is split into training and testing sets using a 70-30 train-test split. (Ultimately ten-fold cross validation has been performed)
- Different machine learning classifiers such as Random Forest, Decision Tree, K-nearest neighbor, Support Vector Machine, Naïve Bayes etc. were trained on the training data to predict race labels from letter counts.
- The model's accuracy is evaluated on the test set and the out-of-bag (OOB) accuracy is computed.

Results and Analysis:

- The accuracy score provides an indication of how well the model predicts race labels based on last names.
- A confusion matrix is generated to visualize the number of correct and incorrect predictions for each race category.
- The classification report presents precision, recall, F1-score, and support for each race category.

Note: The model's performance and predictions should be interpreted cautiously as race/gender prediction based on names alone may introduce biases and should be handled responsibly and ethically.

Reference:

1. Popular Baby Names https://www.ssa.gov/oact/babynames/limits.html (accessed Jul 30, 2021)
2. Decennial Census by Decades https://www.census.gov/programssurveys/decennialcensus/decade.2010.html (accessed Jul 30, 2021)