COMPS224B: Quantitative Political Risk
Problem Set 1
Due: April 23, 2025 11:59pm ET

1 Forecasting the U.S. 2024 Election

: Using the 2024 county-level presidential results ("pres_2024_results.csv"), fit two classification models to predict whether a given county voted for Trump. You should use `sklearn` where possible to set-up your ML pipeline.

1. Create a binary variable indicating whether Trump won a given country (trumpwon) that is 1 if gop_vote_share > 0.5 and zero otherwise. Use train_test_split 0.85/0.15 and seed 2025.
2. Using `LogisticRegression()` in sklearn, estimate a logit model using the explanatory variables from the US census inputs listed on the next page. Print the accuracy, sensitivity, specificity, and F1 scores for train and test samples.
3. Using `RandomForestRegressor()` in sklearn, re-run the model using a random forest. Print the accuracy, sensitivity, specificity, and F1 scores for train and test samples.
4. Using `permutation_importance` in sklearn, what is the most important predictor in the logit model? What is the most important predictor in the RF?
5. Provide a table comparing the overall accuracy of these two models. Which of these methods is 'best' and why? Justify the metric you choose.
6. What changes to the model would you make to predict the 2026 midterms?

2 Create a 'Fragile' State Index

Principal Component Analysis: What makes a "weak" or "fragile" state? We will use PCA and K-Means clustering to identify what factors plausibly explain variation in state capacity using cross-sectional data ("statecap.csv) on different socio-economic and military indicators from the World Bank.

1. Handle any missing data. Justify your approach.
2. Make a correlation plot of the main attributes. What features have the strongest correlations? How are they related?
3. Perform PCA.
   - Plot the biplot. Label each observation with country name.
   - What loading vectors are close together? What attributes would you say best describe a 'weak' state and why?
   - What does each of the first two principal components seem to explain?
4. Create a 'fragile state' index from 0-100 based on the PCA results where higher values correspond to weaker states. How do the results compare to the Fragile State Index `https://fragilestatesindex.org/global-data/`?

K-Means Clustering

1. Use K-means clustering and $k = 5$ to create a series of country clusters. Label these clusters based on what you think is their defining characteristic. Are any 'weak' states?
2. Randomly sample a couple states from cluster 1 and 4. Comment on how similar or dissimilar they are.
3. Create a choropleth map of the world where states are colored according to their classification. Add a legend labeling each cluster with your short descriptive from (1).
4. If a company asked you to identify the most fragile states from the map – and where they should avoid investing operations – what cluster would you point to and why? How do the results compare to the Fragile State index?

Appendix

Question 1
Dataframe: pres_2024_results.csv
Relevant datapoints:

- medianincome: logged median income
- hsgradpct: percent of high school grads in the county,
- collegegradpct: percent of college grads in the county,
- internetaccesspct: percent of households with internet access, and
- whitepopulationpct: percent of non-white county population
- dem_vote_share_2020: Vote share that went to Biden in 2020 election

Question 2
Dataframe: statecap.csv
Relevant datapoints:

- gdp_per_capita: GDP per capita (constant 2015 USD) (Source: World Bank var: NY.GDP.PCAP.KD)
- infant_mortality: Infant mortality rate per 1000 live births (Source: World Bank var: SP.DYN.IMRT.IN)
- life_expectancy: life expectancy at birth (years) (Source: World Bank var: SP.DYN.LE00.IN)
- militaryexpenditures_pctgdp: military expenditures as percent of GDP (Source: World Bank var: MS.MIL.XPND.GD.ZS)
- land: land area in square km (Source: World Bank var: AG.LND.TOTL.K2)
- population: total population (Source: World Bank var: SP.POP.TOTL)
- military_personnel: total armed forces (active duty only) (Source: World Bank var: MS.MIL.TOTL.P1)
- badneighborhood: binary variable that is 1 if there is a neighboring civil war and zero otherwise (Source: Custom)
- rivalries: count variable of the number of geopolitical rivals the country has (Source: Dreyer and Thompson)
- cinc: composite measure of military strength based on population, pig steel production, military expenditures, (Source: Correlates of War)
- IGO_count: count of the number of intergovernmental organizations the state is a member of (Source: Correlates of War)