Predicting County Level Disability from CDC PLACES Data – Dylan, Julian, Grace

A Comparative Study of Mobility and Cognitive Disability

CDC PLACES 2023 | 2,293 U.S. Counties | 1,100+ Models Trained

The Question

Can we predict which U.S. counties have high disability rates using only their chronic disease and behavioral health data?

And do different types of disability have different drivers?

Mobility Disability
Serious difficulty walking or climbing stairs

Cognitive Disability
Serious difficulty concentrating, remembering, or deciding

The Data

Binary target: Top 25% of counties = "High Disability" (class 1)

2,293 U.S. counties
from the 2023 CDC PLACES release

24 predictor features across 4 categories:

Chronic Disease (11) -	Stroke, Diabetes, Obesity, COPD, High BP, Cholesterol, etc.
Behavioral (4) -	Smoking, Physical Inactivity, Binge Drinking, Poor Health
Mental/Physical Distress (2) -	Frequent Mental Distress, Frequent Physical Distress
Social Needs + Access (7) -	Food Insecurity, Loneliness, Health Insurance, etc.

Feature Section

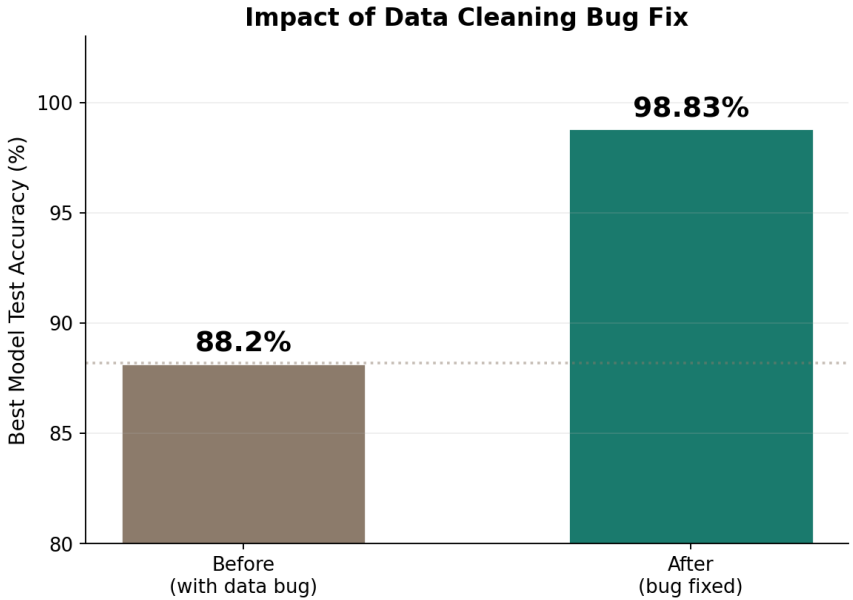
We chose these features due to them all having correlation to mobile and cognitive disability

S (Small, 8 features): Stroke, Diabetes, Obesity, COPD, Coronary Heart Disease, Arthritis, Smoking, and Physical Inactivity.

M (Medium, 15 features): Everything in S, plus High Blood Pressure, High Cholesterol, Depression, Current Asthma, Cancer, Mental Distress, and Physical Distress.

L (Large, 24 features): Everything in M, plus Binge Drinking, Poor General Health, No Health Insurance, Annual Checkup, Cholesterol Screening, Food Insecurity, Housing Insecurity, Loneliness, and No Social Support.

Biggest Challenge - Data Cleaning



Found and fixed a data cleaning bug where duplicate CDC records gave the same county two different class labels. About 20% of the training data had identical inputs pointing to opposite outcomes, which confused every model we trained. After fixing this, our best model accuracy jumped from 88% to 99%.

Instead of keeping both duplicate values as separate rows, we averaged them into one value per county. That way each county appears exactly once in the dataset with one clear label.

Methodologies for ranking

- We first researched and found the most important markers, along with their respective weights and why:
  - **15% - Accuracy:** looking for general correctness.
  - **35% - Macro F1**: Precision with recall.
  - **45% - Macro Recall:** Our primary driver. Our goal is to minimize "false negatives" (not missing any cases).
  - **5% - Training Time:** Medical data, so we would prefer a slow, accurate model, over a fast and inaccurate one, but still need to be considered.
- We then collected the z-score on our data, which was collected via a models %, subtracted by the median, and then divided by the 3rd quartile.
- This then allows us to apply our weights, to collect a "all around" score, which we then ranked and summed up, to provide us with our model rankings.

Overall Model Ranking

Best Model Performers

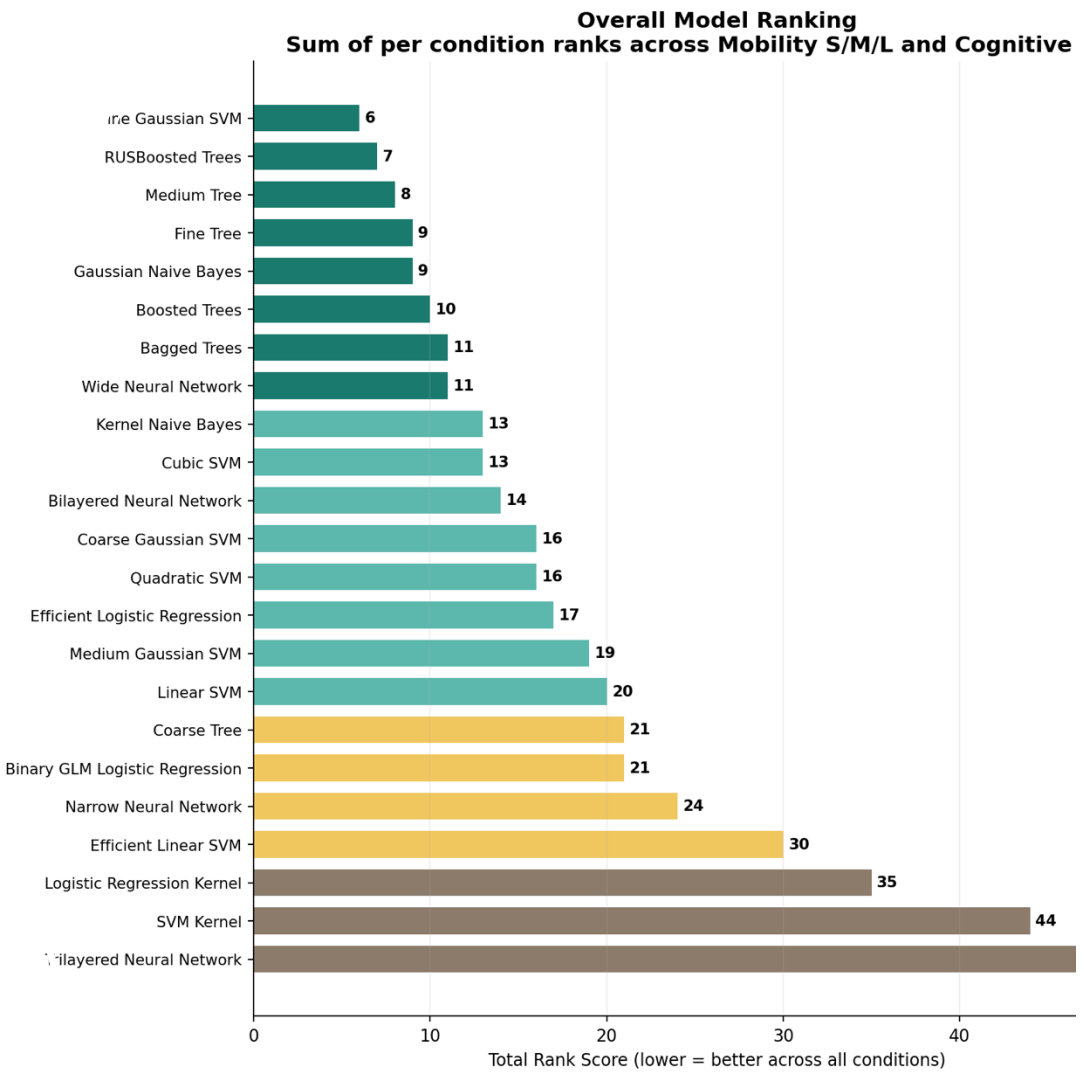
Fine Gaussian SVM

RUSBoosted Trees

Worst Model Performers

SVM Kernel

Trilayererd Nueral Network



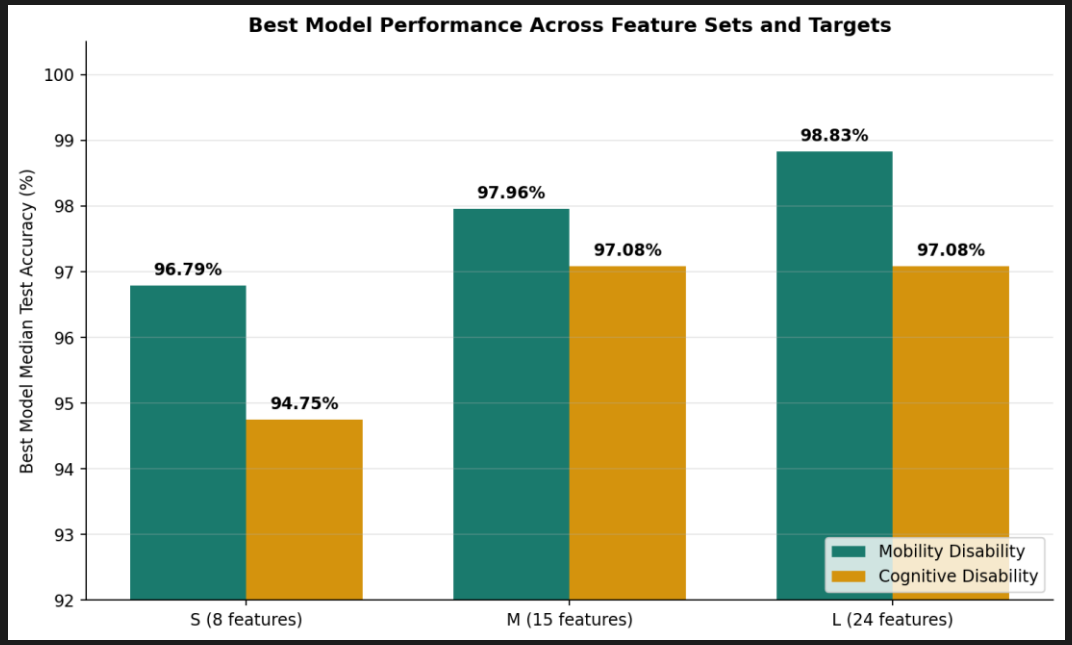
Best Model Performers

- **Fine Gaussian SVM** – Fast prediction speed, medium memory usage, hard interpretability, medium model flexibility
- **RUSBoosted Trees** – Medium prediction speed, medium to high memory usage, moderate interpretability, high model flexibility

Worst Model Performers

- **SVM Kernel** – Medium prediction speed, medium memory usage, easy interpretability, low to medium flexibility
- **Trilayered Neural Network** – Medium to slow prediction speed, high memory usage, hard interpretability, high flexibility

Results



98.83%

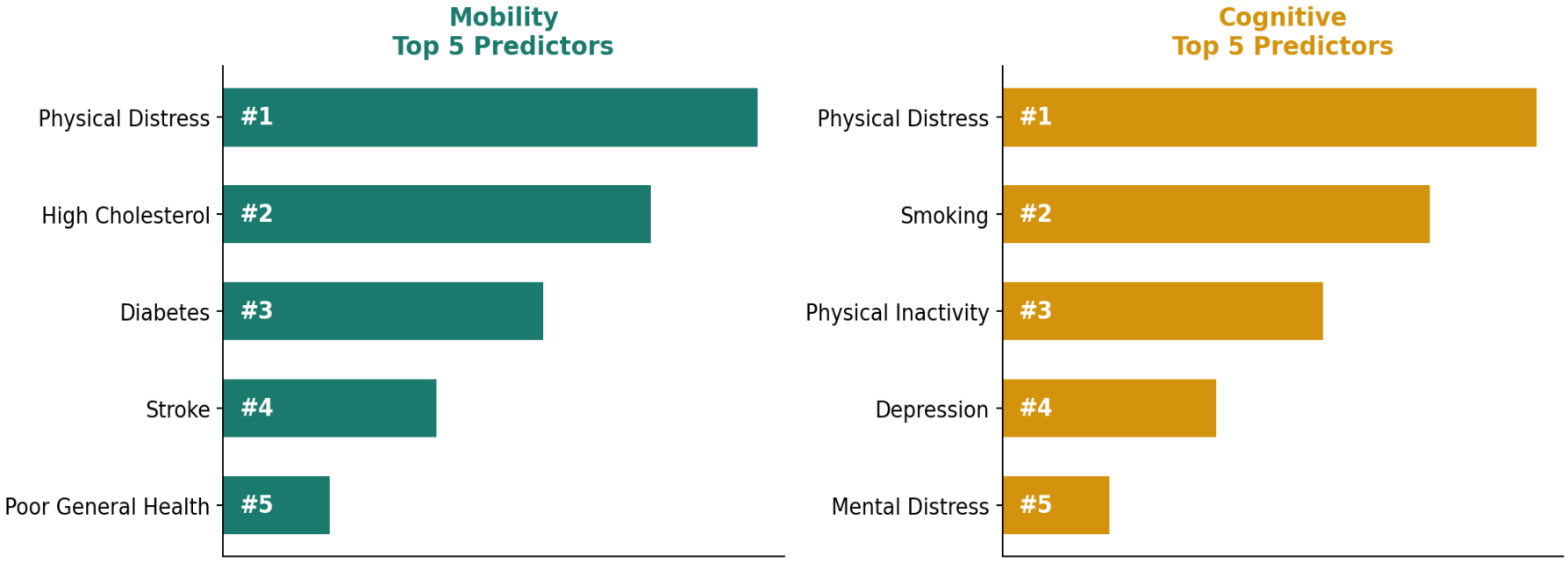
Mobility (Cubic SVM)

97.08%

Cognitive (Wide Neural Net)

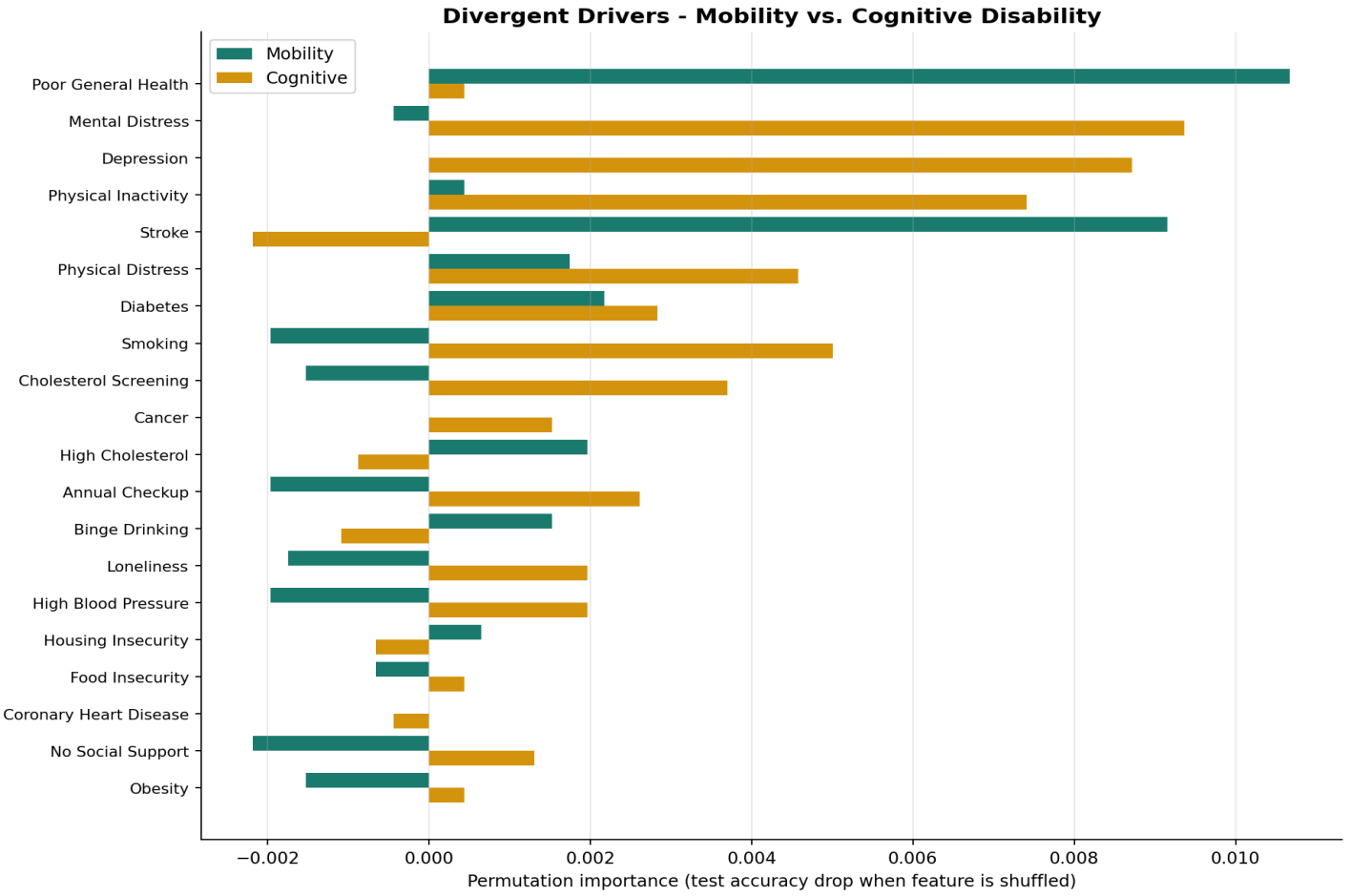
1,100+ models trained
across 35 MATLAB sessions

Different Disabilities, Different Drivers



Only 1 of 5 top predictors appears in both lists. The two disabilities have almost entirely different upstream causes!

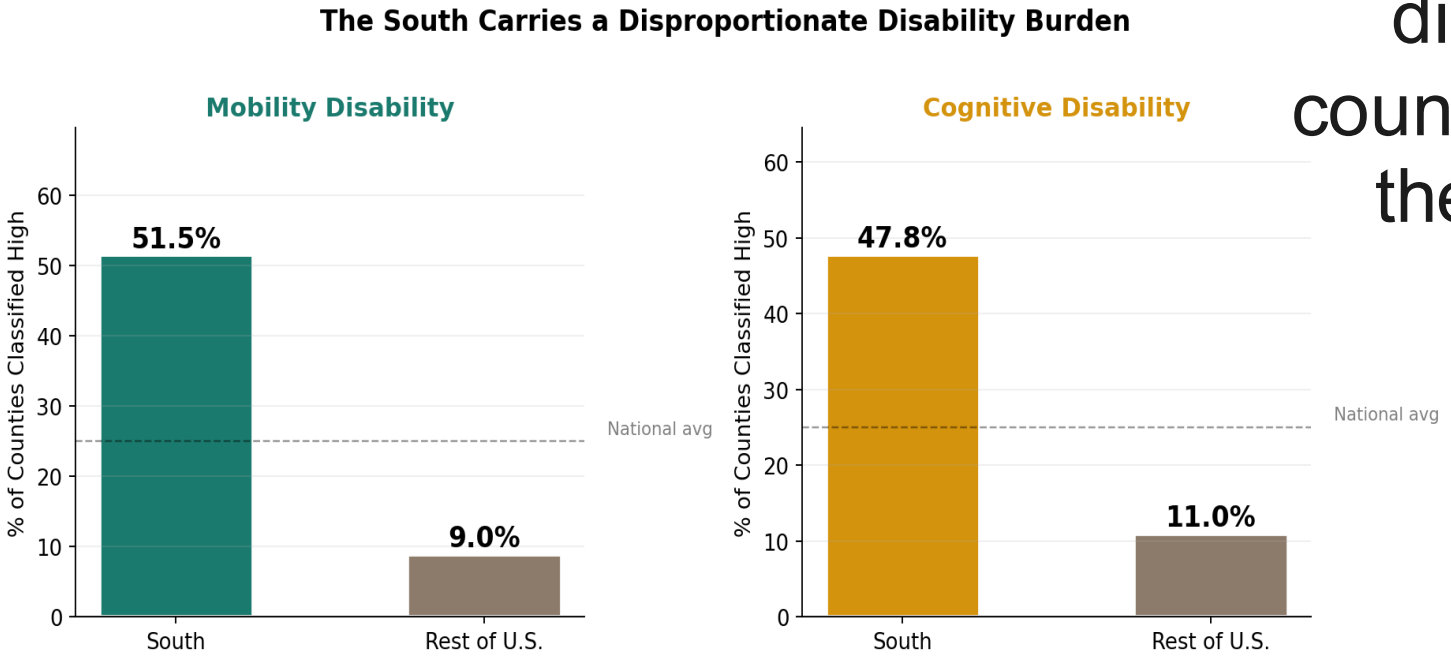
Full Feature Importance Comparison



Geographic Concentration

51.5%

of high mobility disability counties are in the South



Which part of project enhances your skills in analytics?

Data Cleaning: Identifying & Resolving Contradictions in the dataset that significantly improved model accuracy

Machine Learning : Learned importance of model tests and training in MATLAB

Predictive Analytics: Developed models to classify high-disability counties and evaluated performance using accuracy, recall, and F1-score

Statistical Analysis: Conducted comparative driver analysis to identify key differences between mobility and cognitive disability predictors

Key Takeaways

1. County disability is highly predictable from health data (98.83% accuracy)
2. Mobility and Cognitive disability have almost completely different predictors
3. ”Any disability" measures brings distinct problems to analysis
4. Regional differences are fully explained by measurable health factors
5. Data cleaning properly had a large accuracy impact (+10%)