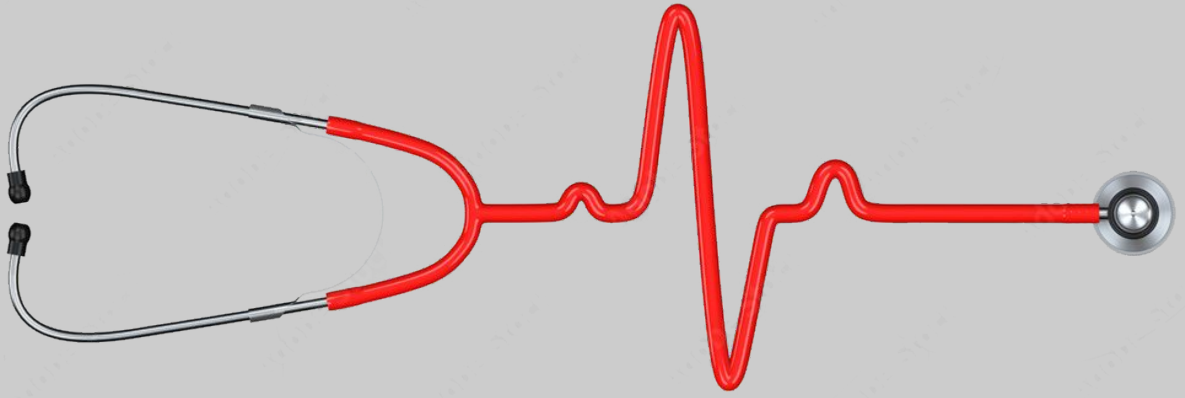


Stroke Predictor

University of Minnesota Data Visualization and Analytics Boot Camp

Team 2 — Janice Courtois, Alex Norgren, Tom Pankratz, Rachel Rautenberg

June 9, 2022

Team 2

Team 2 members all work at Mayo Clinic.

Janice Courtois

- Works in Healthcare Technology Management
- Lives on horse ranch
- Travels often to visit kids & grandson

Alex Norgren

Tom Pankratz

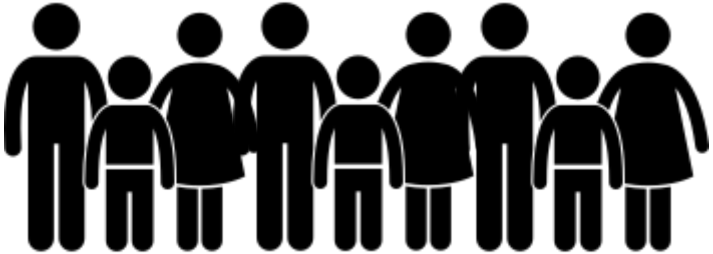
- 19 years at Mayo Clinic
- Manages a digital experimentation team
- Dad of 4

Rachel Rautenberg

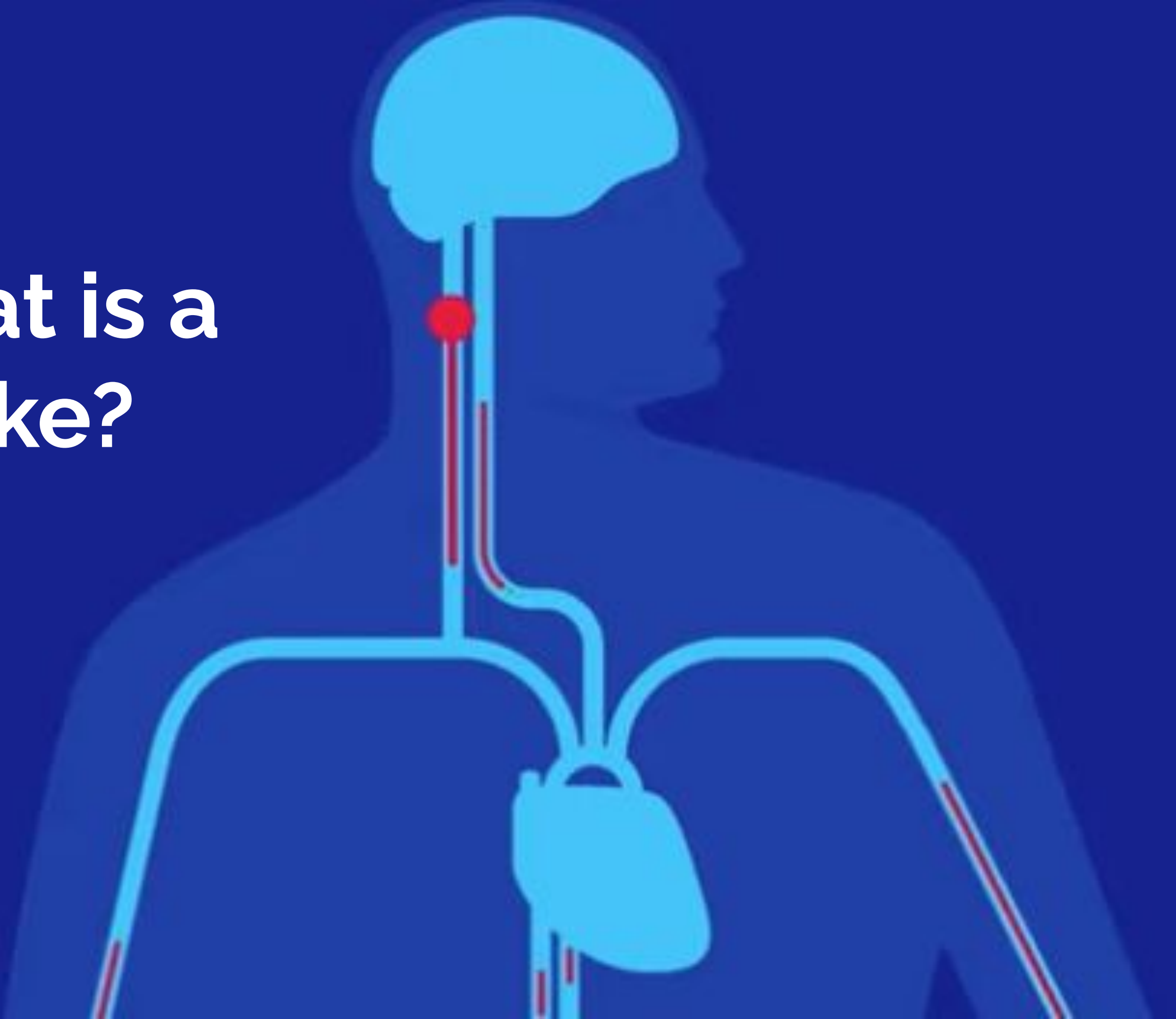
- Holds MHA
- 14 years at Mayo
- Mom of 4
- Enjoys the chaos

Topic: Stroke mortality

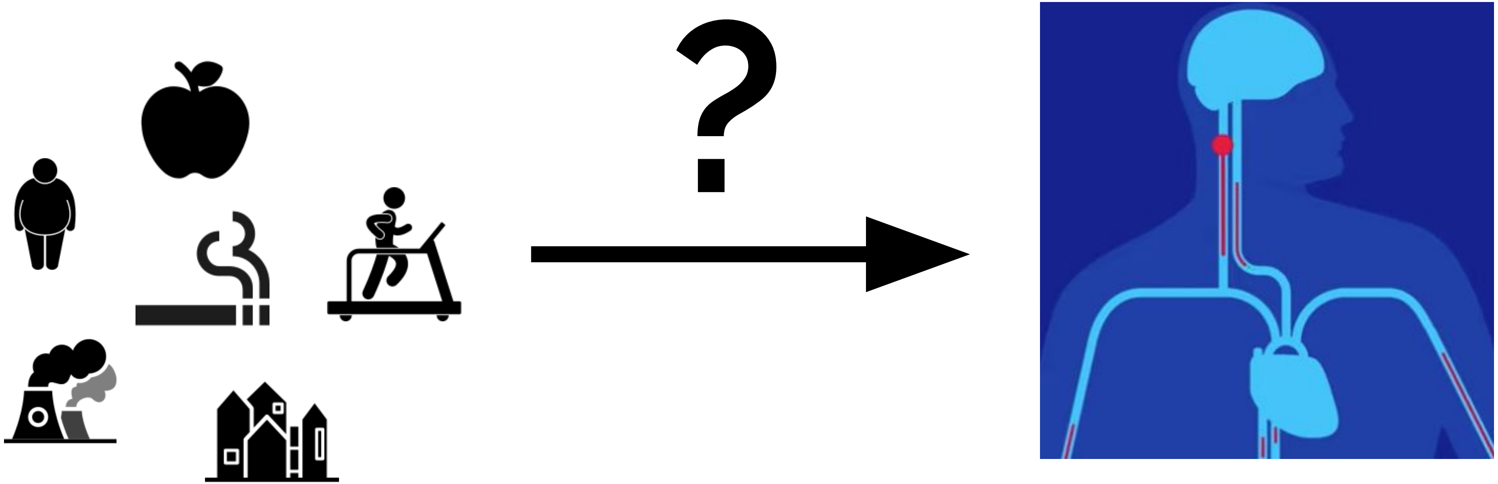
MAYO
CLINIC



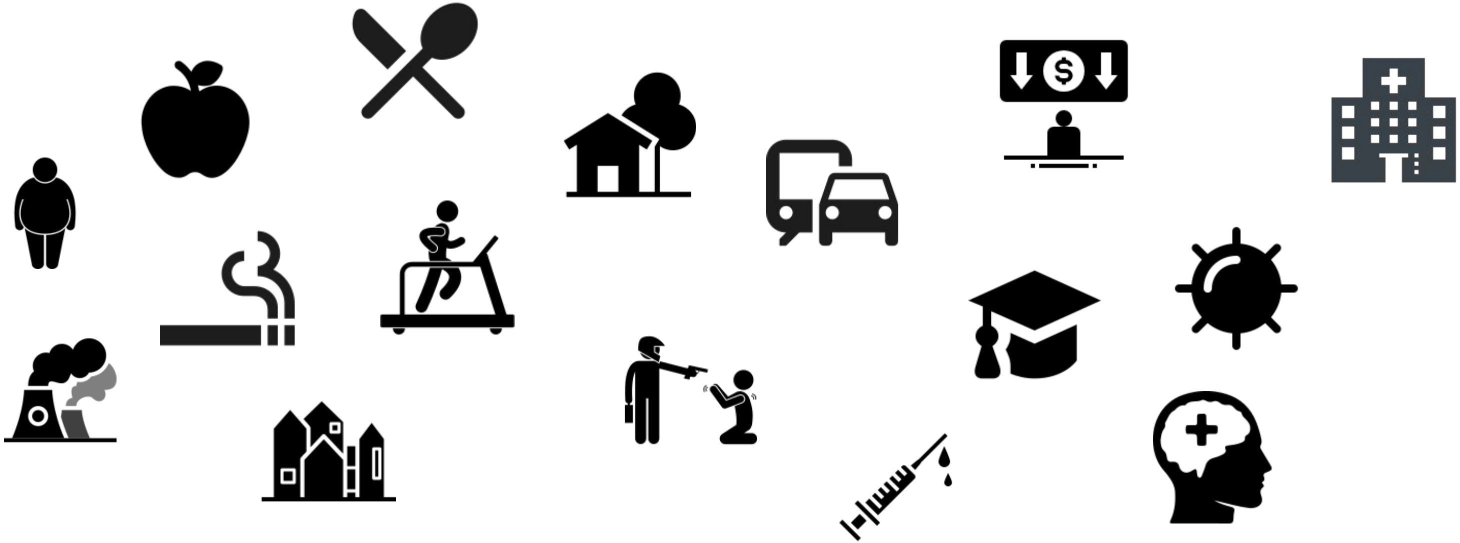
What is a stroke?



Goal of project & questions?



Brainstorming possible factors



Factors we landed on

Health-related:

- Smoking
- Obesity
- Access to healthy foods
- Access to exercise opportunities
- Primary care availability
- Availability of mental health providers

Social-related:

- College education
- Unemployment
- Income
- Violent crime rate
- Air pollution
- Length and type of commute to work
- Urban vs. rural

Source data

CDC

Stroke Mortality Data Among US Adults (35+) by State/Territory and County (2018)

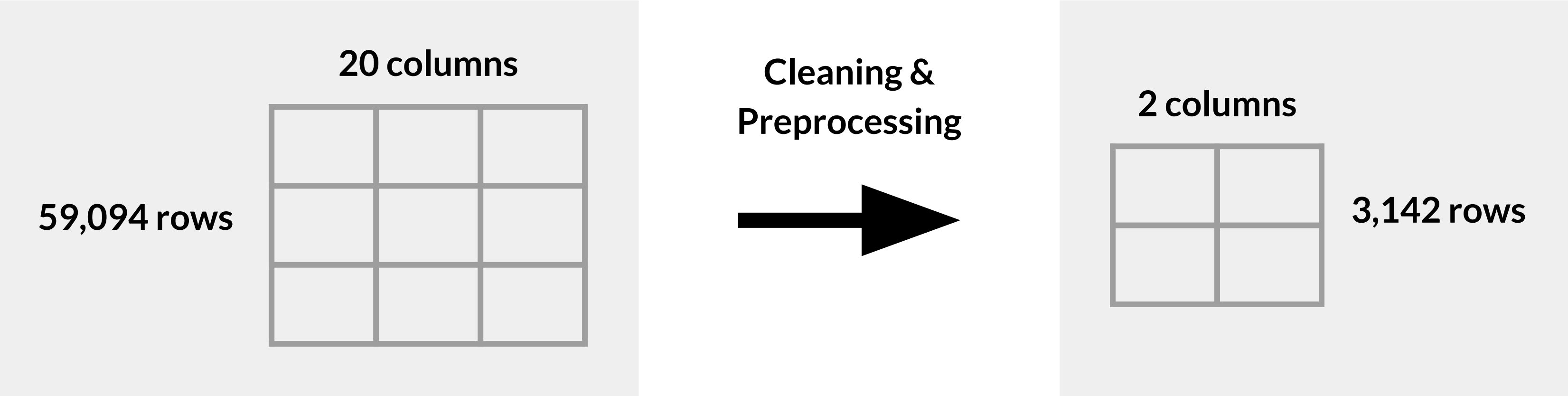
University of Wisconsin
Population Health Institute
SCHOOL OF MEDICINE AND PUBLIC HEALTH

County Health Rankings (2018)

Data exploration and integration

- Cleaning
- Preprocessing
- Merging in PostgreSQL

Target: Stroke mortality dataset



20 columns

59,094 rows

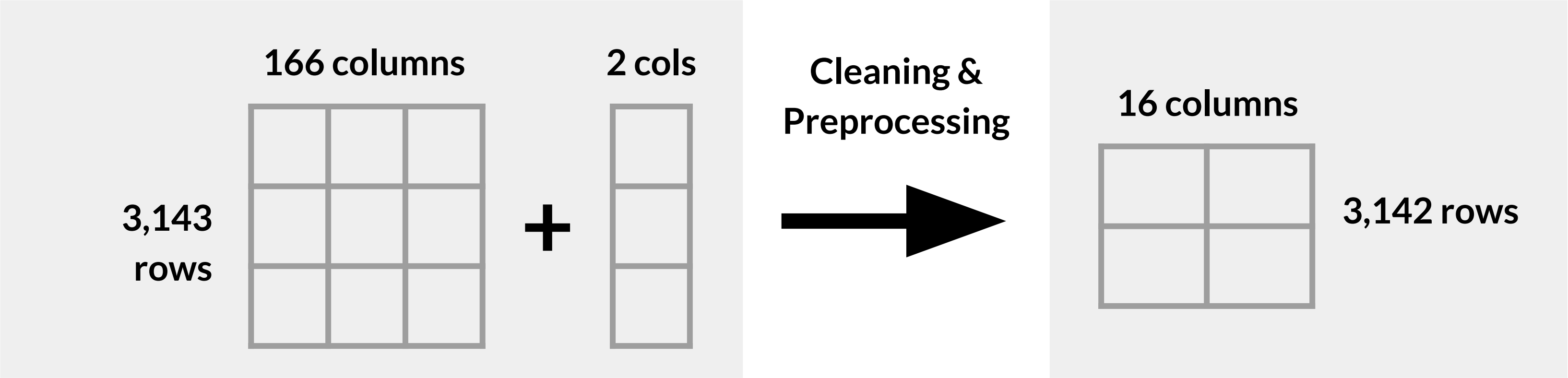
Cleaning & Preprocessing

2 columns

3,142 rows

[Python notebook file]

Features: Health rankings datasets



166 columns

2 cols

3,143 rows

+

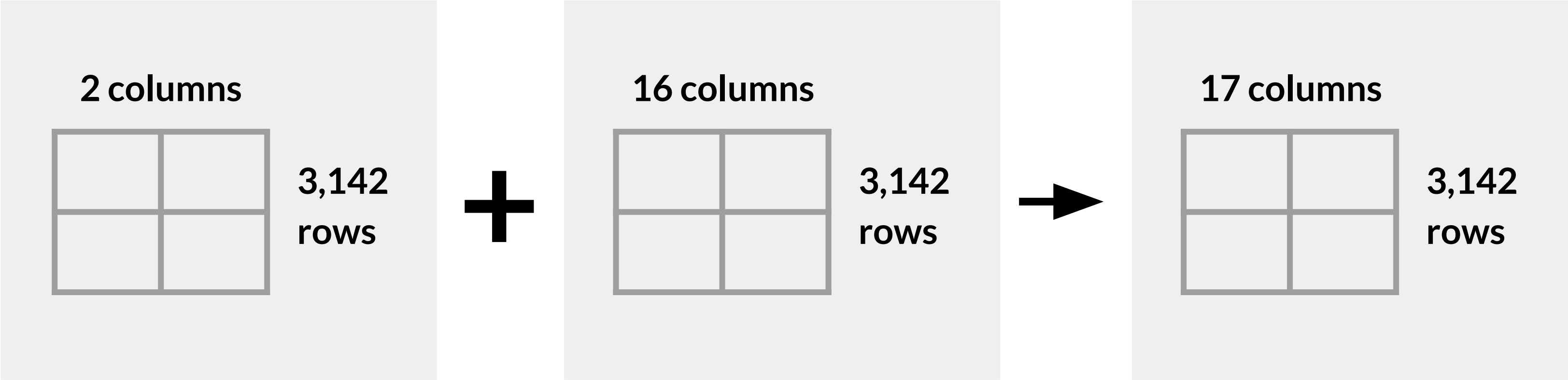
Cleaning & Preprocessing

16 columns

3,142 rows

Python notebook file

Datasets merge via PosgreSQL



2 columns

3,142 rows

+

16 columns

3,142 rows

17 columns

3,142 rows

Python notebook file

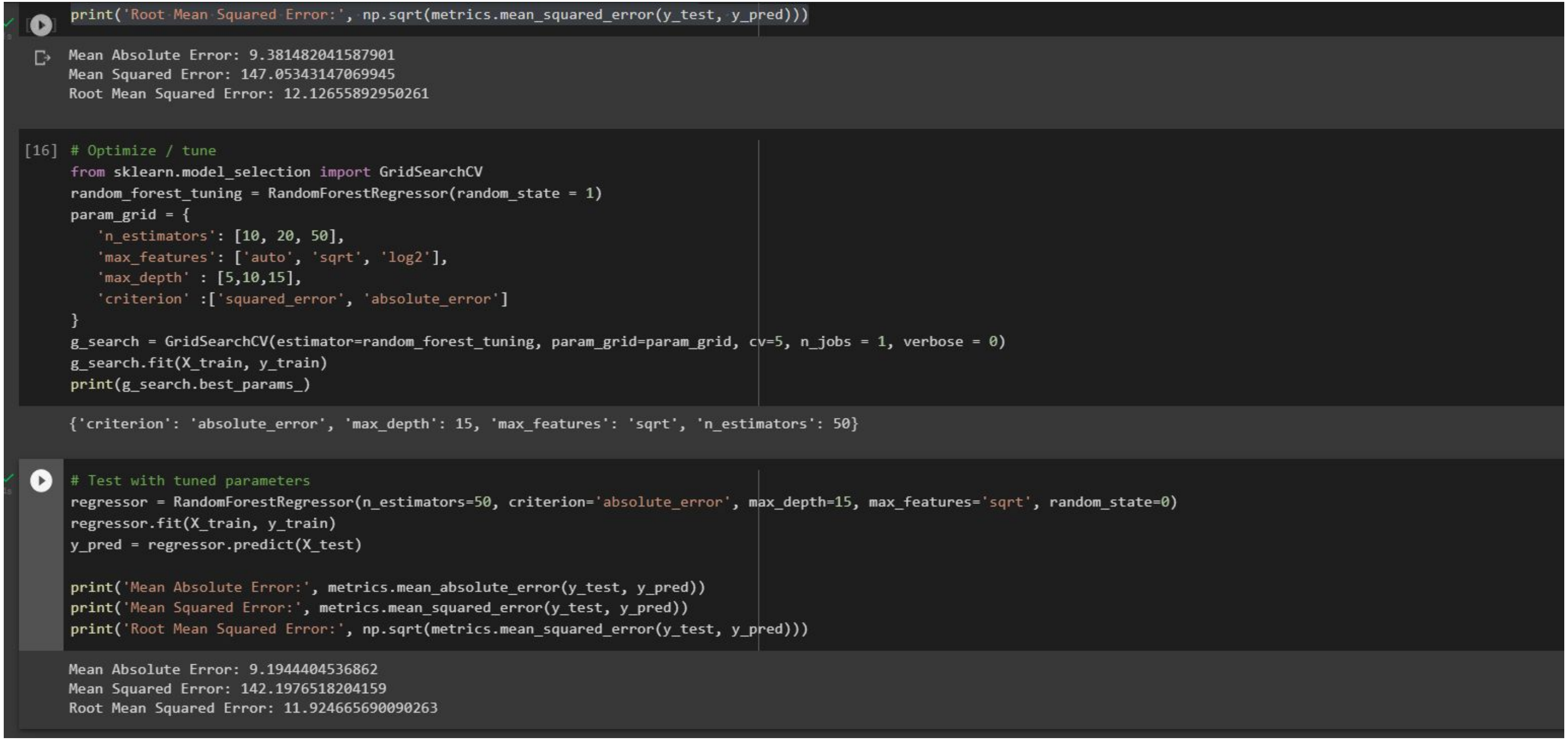
PostgreSQL post-join view

Analysis

- Machine learning model exploration
- Training and testing
- Model choice
- Model importances
- Model output and usage

Machine learning model exploration

Machine learning model exploration



```
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error: 9.381482041587901
Mean Squared Error: 147.05343147069945
Root Mean Squared Error: 12.12655892950261
```

```
# Optimize / tune
from sklearn.model_selection import GridSearchCV
random_forest_tuning = RandomForestRegressor(random_state = 1)
param_grid = {
   'n_estimators': [10, 20, 50],
   'max_features': ['auto', 'sqrt', 'log2'],
   'max_depth' : [5,10,15],
   'criterion' :['squared_error', 'absolute_error']
}
g_search = GridSearchCV(estimator=random_forest_tuning, param_grid=param_grid, cv=5, n_jobs = 1, verbose = 0)
g_search.fit(X_train, y_train)
print(g_search.best_params_)
```

```
{'criterion': 'absolute_error', 'max_depth': 15, 'max_features': 'sqrt', 'n_estimators': 50}
```

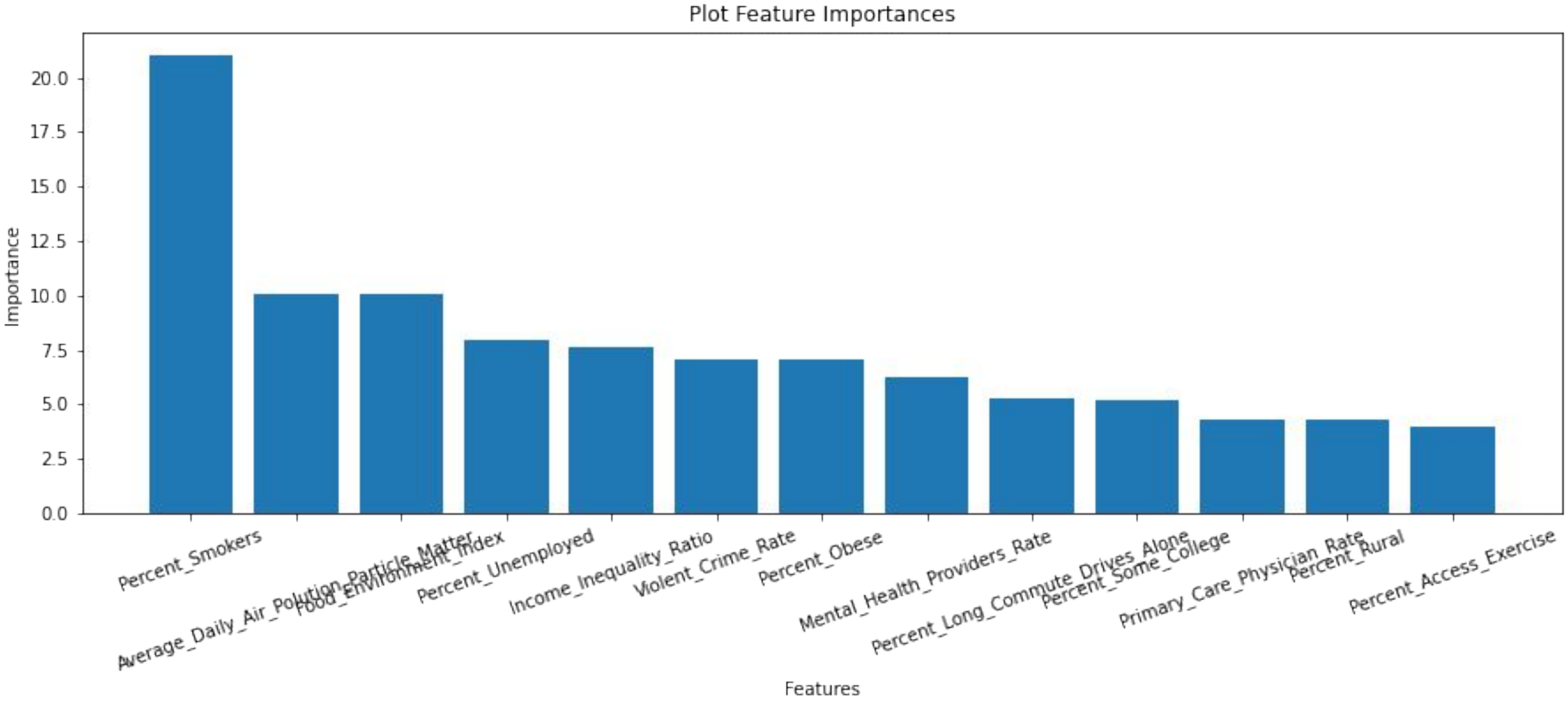
```
# Test with tuned parameters
regressor = RandomForestRegressor(n_estimators=50, criterion='absolute_error', max_depth=15, max_features='sqrt', random_state=0)
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)

print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error: 9.1944404536862
Mean Squared Error: 142.1976518204159
Root Mean Squared Error: 11.924665690090263
```

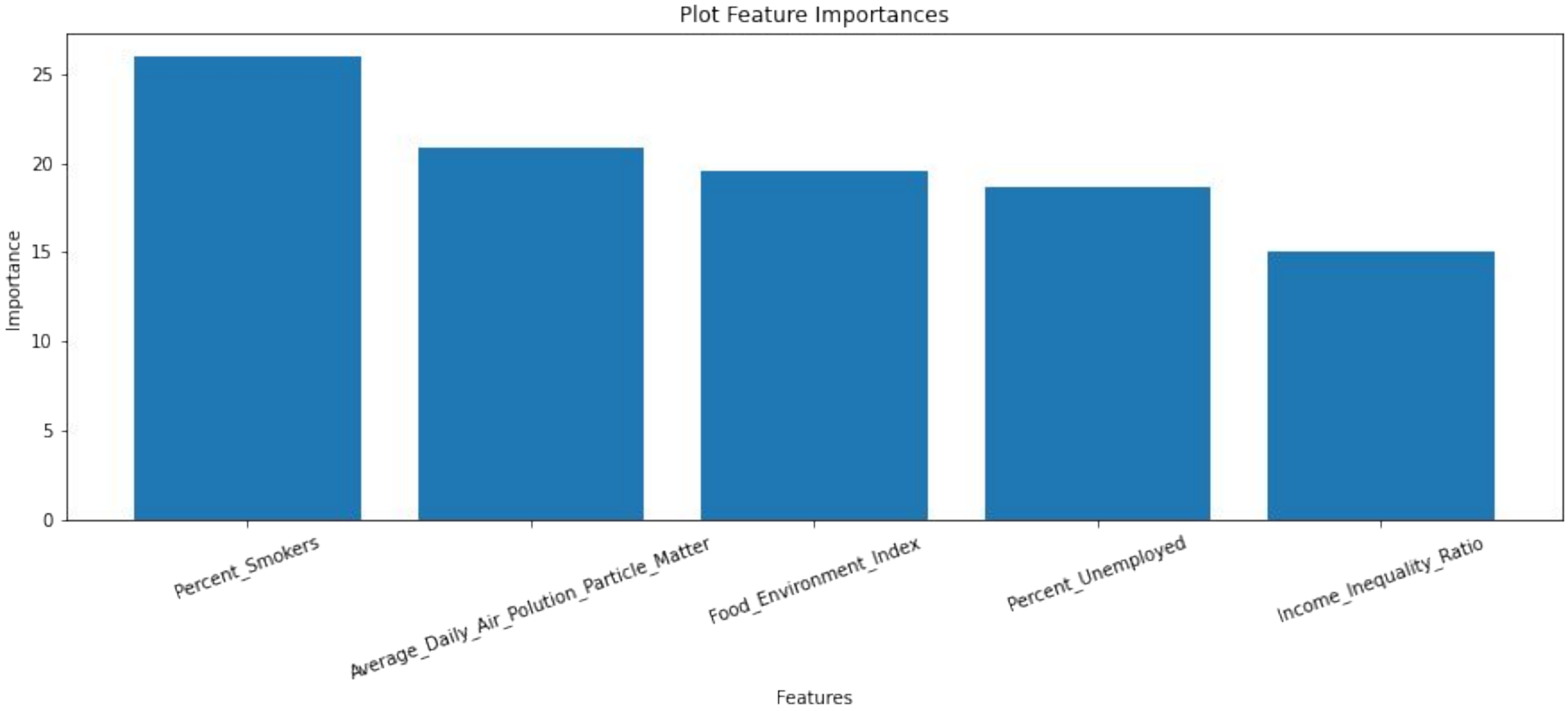
Training and testing

Model choice

Model importances



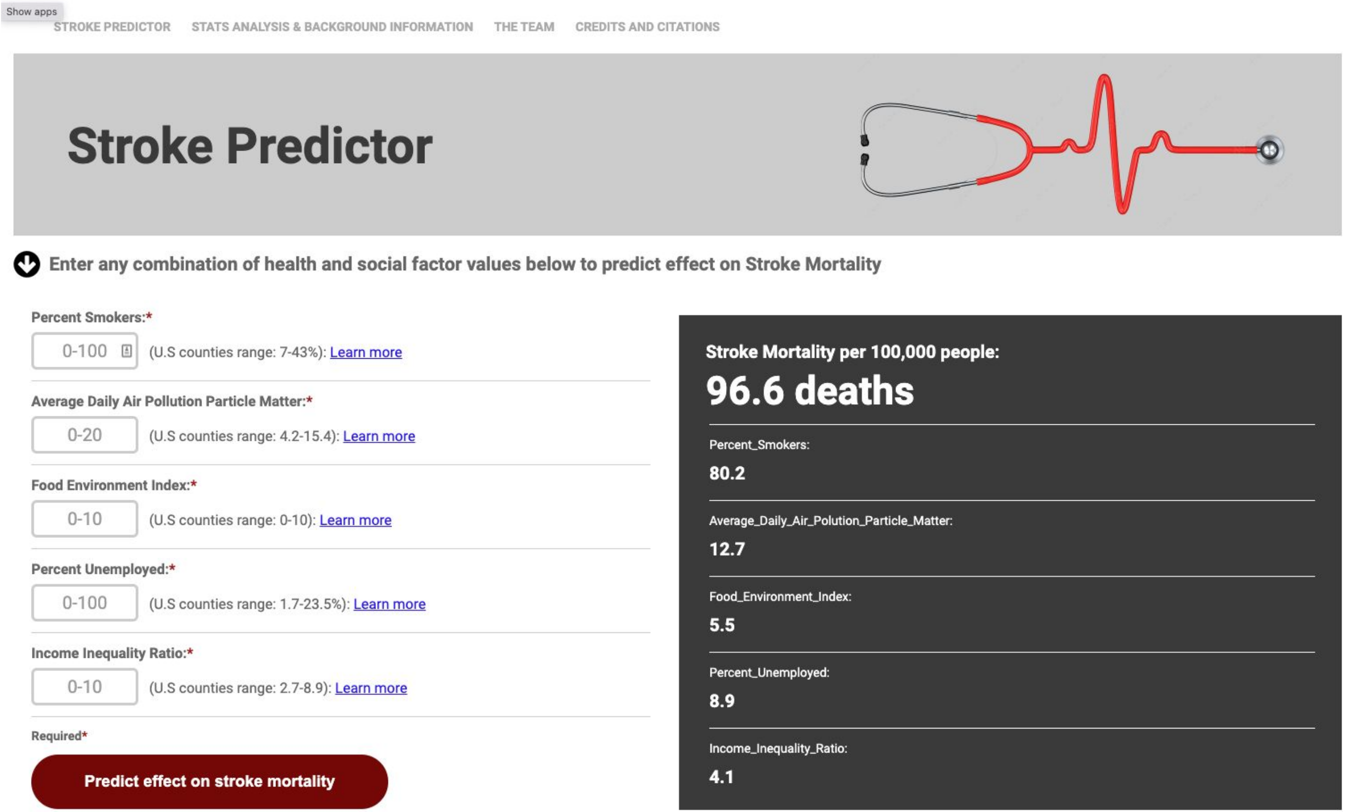
Model importances: Top 5



Model output and usage

Question: Could we input variations of the feature data to determine what sort of effect it would have on stroke mortality?

The stroke predictor web input form


Show apps
STROKE PREDICTOR
STATS ANALYSIS & BACKGROUND INFORMATION
THE TEAM
CREDITS AND CITATIONS
Predict effect on stroke mortality

Stroke Predictor

Enter any combination of health and social factor values below to predict effect on Stroke Mortality

Percent Smokers:*
0-100 (U.S counties range: 7-43%): Learn more

Average Daily Air Pollution Particle Matter:*
0-20 (U.S counties range: 4.2-15.4): Learn more

Food Environment Index:*
0-10 (U.S counties range: 0-10): Learn more

Percent Unemployed:*
0-100 (U.S counties range: 1.7-23.5%): Learn more

Income Inequality Ratio:*
0-10 (U.S counties range: 2.7-8.9): Learn more

Required*

Stroke Mortality per 100,000 people:

96.6 deaths

Percent_Smokers:
80.2

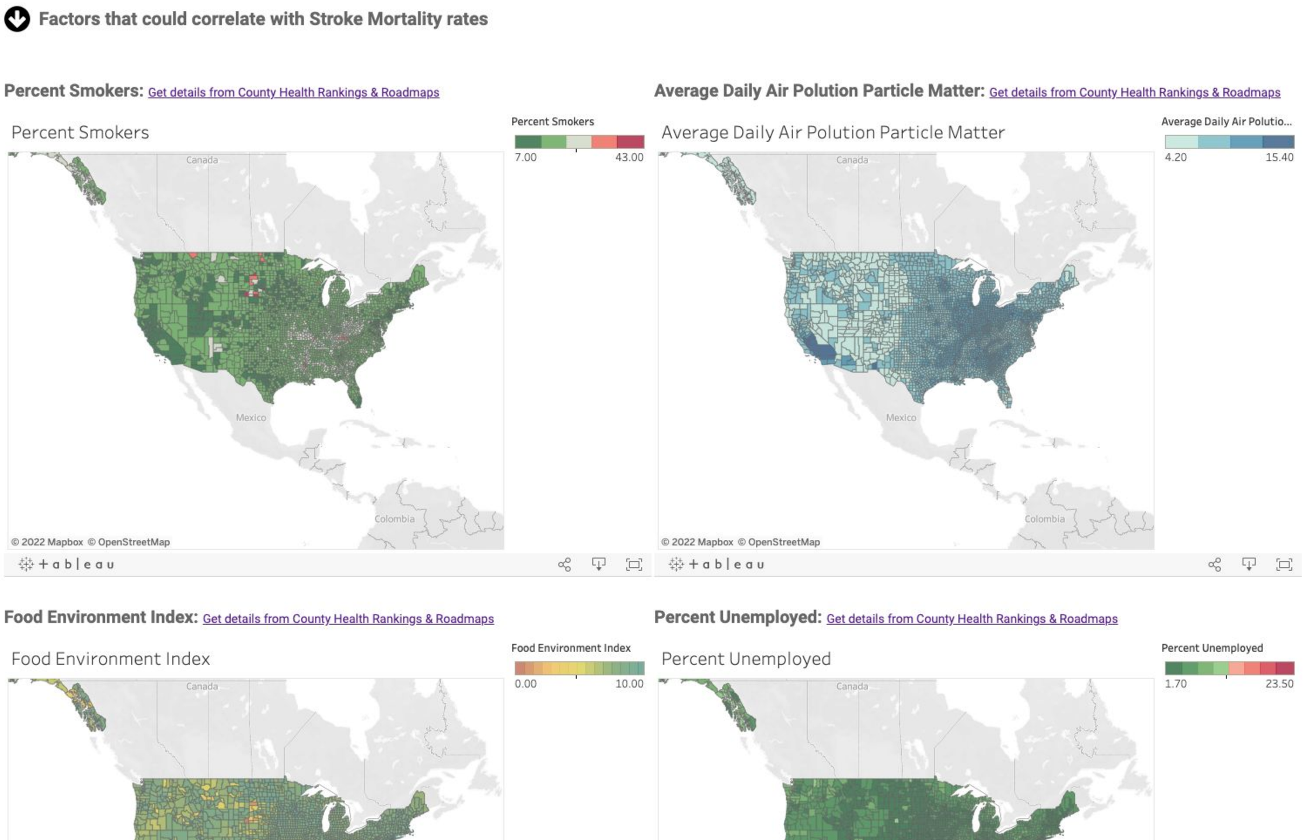
Average_Daily_Air_Polution_Particle_Matter:
12.7

Food_Environment_Index:
5.5

Percent_Unemployed:
8.9

Income_Inequality_Ratio:
4.1

Features data maps dashboard



Factors that could correlate with Stroke Mortality rates

Percent Smokers: Get details from County Health Rankings & Roadmaps

Percent Smokers

Percent Smokers
7.00 – 43.00

Average Daily Air Polution Particle Matter: Get details from County Health Rankings & Roadmaps

Average Daily Air Polution Particle Matter

Average Daily Air Polutio...
4.20 – 15.40

Food Environment Index: Get details from County Health Rankings & Roadmaps

Food Environment Index

Food Environment Index
0.00 – 10.00

Percent Unemployed: Get details from County Health Rankings & Roadmaps

Percent Unemployed

Percent Unemployed
1.70 – 23.50

Result of analysis

- Results following model inputs
- Recommendation for future analysis
- What could we have done differently?

Results

Percent Smokers appeared to have the largest impact on stroke mortality, but beyond that, it was difficult to determine impacts from the other features. It was less a matter of machine learning model choice, and more a matter of the choice of features/factors.

Recommendations

Choose features/factors that have already been determined by the health care community to have a larger impact on predicting stroke mortality, as a starting point.