DATA MINING INDICATORS OF HEART DISEASE

CSPB 4502 PROJECT
GROUP 10

CYRO ESTEVAO, ERIC HERNANDEZ, LINDSAY KRIZ

GROUP 10 GITHUB

PROJECT DESCRIPTION

- OUR GOAL OF THIS PROJECT IS TO PRACTICE DATA MINING TECHNIQUES TO ANSWER INTERESTING QUESTIONS FROM A HEART DISEASE INDICATOR DATASET. WE WOULD LIKE TO ANSWER THE QUESTIONS:
  - WHICH STATES HAVE THE HIGHEST RATE OF HEART DISEASE?
  - DOES PHYSICAL ACTIVITY AFFECT HEART DISEASE?
  - WHICH FEATURES CORRELATE WITH HEART DISEASE?
  - IF SOMEONE HAS HEART DISEASE, WHAT IS THE LARGEST & MOST COMMON SET OF SYMPTOMS/FEATURES?
  - WHICH DEMOGRAPHICS HAVE HEART DISEASE?

PRIOR WORK

- THE NOT-A-NULL/BLANKS HAVE BEEN REMOVED
- DATA HAS BEEN BROKEN DOWN INTO BINARY ATTRIBUTES
- REVIEWED THE DATASET & DISCUSSED THE FEATURES
- PLANNING EDA & DATA PRE-PROCESSING

DATASET

- INDICATORS OF HEART DISEASE DATASET (KAGGLE)
- 400K+ RECORDS (SOME INCLUDE NAN VALUES)
- HEALTH INTERVIEW ANSWERS FROM 400K+ INDIVIDUALS
- "ESTABLISHED IN 1984 WITH 15 STATES, BRFSS NOW COLLECTS DATA IN ALL 50 STATES, THE DISTRICT OF COLUMBIA, AND THREE U.S. TERRITORIES. BRFSS COMPLETES MORE THAN 400,000 ADULT INTERVIEWS EACH YEAR, MAKING IT THE LARGEST CONTINUOUSLY CONDUCTED HEALTH SURVEY SYSTEM IN THE WORLD." –CDC
- "A HEART ATTACK IS OFTEN AN INDICATOR OF HEART DISEASE." - GOOGLE
- DATASET WILL BE DOWNLOADED ON ALL MEMBERS' MACHINES AND CONNECTED VIA GITHUB

PROPOSED WORK

- **DATA CLEANING:**
  - REVIEW & ANALYZE DATASET
  - ADDRESS MISSING VALUES (IF WE USE THE NAN-INCLUDED DATASET)
  - UPDATE BINARY ANSWERS
  - REVIEW & UPDATE DATA TYPES IF NEEDED
  - REMOVE DUPLICATES & OUTLIERS
- **DATA PREPROCESSING:**
  - ASSIGN UNIQUE IDENTIFIER
  - ASSIGN ‘HEART ATTACK: Y/N’ AS ‘HEART DISEASE: Y/N’
  - INITIAL VISUALIZATIONS
- **DATA INTEGRATION:**
  - IF ENOUGH TIME PERMITS:
    - INTEGRATE HEALTH INSURANCE RATES PER STATE TO FIND CORRELATIONS WITH HEART DISEASE
    - COMPARE 2022 & 2020 ANALYSIS

TOOLS

- KAGGLE
- GITHUB
- JUPYTERLAB
- EXCEL
- PYTHON
  - PANDAS
  - SKLEARN
  - MATPLOTLIB
  - SNS
- DISCORD

EVALUATION

- ANSWER THE INTERESTING QUESTIONS OUTLINED IN THE PROJECT DESCRIPTION
- COMPLETE VISUALIZATION FOR EACH QUESTION ANSWERED
- APPLY CROSS VALIDATION TECHNIQUES