DATA MINING INDICATORS OF HEART DISEASE

CSPB 4502 PROJECT
GROUP 10

CYRO ESTEVAO, ERIC HERNANDEZ, LINDSAY KRIZ

GROUP 10 GITHUB

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

QUESTIONS SOUGHT TO ANSWER

- OUR GOAL FOR THIS PROJECT WAS TO PRACTICE DATA MINING TECHNIQUES TO ANSWER INTERESTING QUESTIONS FROM A HEART DISEASE INDICATOR DATASET. WE SET OUT TO ANSWER THE QUESTIONS:
  - WHICH STATES HAVE THE HIGHEST RATE OF HEART DISEASE?
    - Arkansas, Florida, Maine, and West Virginia exhibit the highest percentage of their populations affected by Heart Disease, comprising over 7% of each state's population.
  - DOES PHYSICAL ACTIVITY AFFECT HEART DISEASE?
    - Indirectly, yes. In our Bayesian Networks Analysis, we've discerned that while there isn't a direct link between Physical Activity and Heart Disease, there's a significant correlation between Difficulty Walking - attribute associated with Physical Activity - and the presence of Angina, which is strongly correlated with Heart Diseases.
  - WHICH FEATURES CORRELATE WITH HEART DISEASE?
    - Only HasAngina is directly correlated. Hence, HasAngina is correlated with Age Category and Difficulty Walking.
  - IF SOMEONE HAS HEART DISEASE, WHAT IS THE LARGEST & MOST COMMON SET OF SYMPTOMS/FEATURES?
    - The person with Angina (chest pain), difficulty walking and usually age 50 and above.
  - WHICH DEMOGRAPHICS HAVE HEART DISEASE?
    - Individuals typically affected by heart disease are those who experience angina and are aged 50 years or older, with the likelihood of occurrence increasing with age, difficulty walking and physical activities.

DATA PREPARATION WORK

- REVIEWED THE DATASET & DISCUSSED THE FEATURES
- RENAMED ‘HADHEARTATTACK’ TO ‘HASHEARTDISEASE’
- EDIT ENTRIES THAT WERE NOT ‘YES’ OR ‘NO’
- CONVERTED ‘YES’/’NO’ FEATURES TO BINARY
- CONVERTED ‘AGECATEGORY’ TO A SINGLE AGE VALUE
- ONE-HOT ENCODE ‘AGECATEGORY’ FOR MCA
- PLANNING EDA & DATA PRE-PROCESSING

TOOLS USED

- KAGGLE
- GITHUB
- JUPYTERLAB
- EXCEL
- DISCORD

- PYTHON
  - PANDAS
  - SKLEARN
  - MATPLOTLIB
  - SNS
  - KMEANS
  - MCA
    - PRINCE
  - CONFUSION MATRIX
  - PGMPY
    - BAYESIANNETWORK
    - MAXIMUMLIKELIHOODESTIMATOR

CLASSIFICATION/CLUSTERING/ETC.

- USED MATH TO CALCULATE WHICH STATES HAD THE HIGHEST PERCENTAGE OF HEART DISEASE
- USED CORRELATION MATRICES TO CHOOSE WHICH FEATURES TO FOCUS ON
- APPLIED K-MEANS CLUSTERING
  - LED TO LEARNING THAT K-MEANS ISN'T EFFECTIVE FOR BINARY DATA
  - MULTIPLE CORRESPONDENCE ANALYSIS (MCA), NOT PCA, MUST BE USED FOR DIMENSION REDUCTION FOR BINARY DATA
- APPLIED BAYESIAN NETWORK
  - Bayesian Networks Analysis helped identify key attributes directly and indirectly linked to heart disease, offering insights into the relationships and predictive modeling of attributes.

KNOWLEDGE GAINED

- A SIGNIFICANT AMOUNT OF TIME MUST BE DEDICATED TO EDA IN ORDER TO HAVE A SUCCESSFUL OUTCOME
  - SOME DATA EXPLORATION NEEDS TO BE PERFORMED BEFORE DECIDING WHICH QUESTIONS TO ANSWER AND WHICH MODELS TO USE
- WHICH MODELS/TOOLS WORK FOR WHICH TYPE OF DATASET
  - K-MEANS AND MCA AREN'T SUITABLE FOR BINARY/CATEGORICAL DATA
- RESULTS FROM ONE MODEL CAN LEAD TO THE NEXT APPROPRIATE STEP IN THE DATA MINING PROCESS

HOW THIS KNOWLEDGE CAN BE APPLIED

- This new knowledge can easily be applied in our daily jobs
  - Allocating more time/resources to EDA before choosing a project direction
  - Presenting EDA findings to leadership before moving forward
- This data mining project seeks to address critical questions regarding Heart Disease, given its profound impact on public health. By understanding the geographical distribution of Heart Disease prevalence, the influence of physical activity, correlated features, prevalent symptoms, and demographic predispositions, targeted interventions and preventive measures can be developed to mitigate its burden and enhance public health outcomes.