DELIVERABLE 1 PROJECT PRESENTATION

By: Jason Leonard, Mamata Khadka, Susmitha Abbaiahvari
(Team A)

Presentation Overview

- **Considered Datasets**
  - Yearly Air Quality Index
  - In Hospital Mortality
  - Heart Failure (Cardiovascular Disease)
  - BRFSS Health Survey
  - COVID-19 Estimated ICU Beds
- **Selected Dataset**
  - Breast Cancer
  - Opportunities and Challenges

Considered Datasets

Yearly Air Quality Index

- Retrieved from Kaggle, data originally from the Environmental Protection Agency (EPA)
  - Reports the yearly air quality of a specific county
- ~34k rows/observations **(*n* ≈ 34,000)**
- 4 predictor variables **(*p* = 4)**
- Outcome variable - ***Number of Good Days***
  - type : numerical

Yearly Air Quality Index

Predictor Variable Table

Variable Name	Type
State	Categorical
County	Categorical
Year	Numerical
Days with AQI (Air Quality Index)	Numerical

Yearly Data Air Quality from the EPA

- Fairly applicable to Healthcare
  - Climate change has a well documented impact on healthcare
- Help healthcare experts better prepare for worsening conditions
  - Assist in proper allocation of resources
- **Not selected because the required model would prove too complicated**
  - Involves a time series dataset

In Hospital Mortality

- Retrieved from Kaggle, originally from MIMIC-III Database
- Consisted of 1,177 rows/observations **(*n* = 1,177)**
- 8 Predictor Variables **(*p* = 8)**
- Outcome Variable - ***Outcome***
  - Type : categorical
  - Binary classification of patient's current status
  - 0 for Alive
  - 1 for Dead

In Hospital Mortality

Predictor Variable Table

Variable Name	Type
Age	Numerical
Gender	Categorical
BMI	Numerical
Hypertensive	Categorical
Diabetes	Categorical
Respiratory Rate	Numercial
Ethnicity	Categorical
Mean Blood Pressure	Numercial

In Hospital Mortality

- Extremely applicable to Healthcare
    - Allow Healthcare professionals to prioritize patients at the most risk before their condition endangers their life
- **Not selected because of substantially missing data**
    - BMI variable is missing data for 28% of observations
    - Not suitable for case-wise deletion

Heart Failure (Cardiovascular Disease)

- Retrieved from Kaggle, originally from a study on Heart Failure
- Consists of 299 observations **(*n*=299)**
- Selected 8 predictor variable **(*p*=8)**
- Outcome variable - ***Death Event***
  - Type : categorical
  - 0 for No Heart Failure
  - 1 for Heart Failure

Heart Failure (Cardiovascular Disease)

Predictor Variable Table

Variable Name	Type
Age	Numerical
Sex	Categorical
Diabetes	Categorical
High Blood Pressure	Categorical
Smoking	Categorical
Anaemia	Categorical
Ejection Fraction	Numerical
Platelets	Numerical

Heart Failure (Cardiovascular Disease)

- Similar to the In Hospital Mortality Dataset
- Narrower variables more specific to Heart Disease
    - Still applicable in a hospital environment or to identify individuals before they succumb to heart failure
- **Not selected due to low sample size**
    - Additionally, multiple publicly available analysis already conducted

BRFSS Health Survey

- Retrieved from the CDC (Centers for Disease Control & Prevention) website
- Consists of 401,958 rows/observations **(*n* = 401,958)**
- 8 predictor variables **(*p* = 8)**
- Outcome Variable - ***BMI_5*** (or simply BMI)
    - Type : numerical

BRFSS Health Survey

Predictor Variable Table

Variable	Variable Description	Type
SEXVAR	Sex of respondent	Categorical
GENHLTH	General Health	Categorical
PHYS HLTH	Number of days physical health is not good in the past 30 days	Numerical
MENT HLTH	Number of days mental health is not good in the past 30 days	Numerical
EXERANY2	Exercise in the past 30 days	Numerical
SLEPTIM1	Average hours of sleep in a 24-hour period	Numerical
EDUCA	Education Level	Categorical
EMPLOY1	Employment Status	Categorical

BRFSS Health Survey

- Interesting application in Healthcare by estimating an individual's BMI based on socioeconomic status and biological factors
- A variety of variables and a good amount of observations
- **Not selected because of missing values (above 5%)**
    - Could not verify if missing data was systematically removed

COVID-19 Estimated ICU Beds Occupied

- Retrieved from HealthData.gov
- Consists of 1,613 rows/observations (***n* = 1,613**)
- 5 predictor variables (***p* = 5**)
- Outcome Variable - ***Number of Available Beds***
    - Calculated by *Subtracting Total Staffed Beds - Staffed adult ICU beds Occupied*
    - Type : numerical

COVID-19 Estimated ICU Beds Occupied

Predictor Variable Table

Variable	Type
State	Numerical
Collection Date	Numerical
Staffed Adult ICU beds Occupied Estimated	Numerical
Total Staffed Adult ICU Beds	Numerical
Percentage of staffed adult ICU beds occupied estimated	Numerical

COVID-19 Estimated ICU Beds Occupied

- A model created from this dataset could help predict Bed shortages before then happen given the current trend of availability
- Model created for the COVID-19 pandemic could be helpful in future surge of hospitalizations
- **Not selected because like the EPA Air Quality dataset, this dataset relies on a time series**

Selected Dataset

Breast Cancer

- Retrieved from Kaggle, originated from a study on Breast Cell Cytology
- Consists of 683 rows/observations (***n* = 683**)
- 9 predictor variables (***p* = 9**)
- Outcome Variable - ***Class***
  - 2 for Benign
  - 4 for Malignant
  - Type : categorical

Breast Cancer

Predictor Variable Table

Variable Name	Type
Clump_Thickness	Categorical
Uniformity of Cell Size	Categorical
Uniformity of Cell Shape	Categorical
Marginal Adhesion	Categorical
Single Epithelial Cell Size	Categorical
Bare Nuclei	Categorical
Bland Chromatin	Categorical
Normal Nucleoli	Categorical
Mitoses	Categorical

Breast Cancer

- Help Healthcare workers quickly identify patients who have a higher likelihood to have a malignant tumors
- **Opportunity**
  - Provide a good opportunity to learn more about the application of statistical methods under the lense of classification
- **Challenge**
  - Data has a relatively small sample size compared to other datasets considered for the project; Could provide the team with a better understanding of how to handle small data sets

Questions?