Capstone Project
Background and Objectives

Health Care Analytics

DATA SCIENCE
INSTITUTE

Background

1. **Background**

In recent years, the field of healthcare has seen a significant transformation with the advent of data science. Data science in healthcare involves the application of statistical methods, machine learning techniques, and computational algorithms to analyze and interpret complex healthcare data.

The dataset originally comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to collect data on the health status of U.S. residents. As described by the CDC: "Established in 1984 with 15 states, BRFSS now collects data in all 50 states, the District of Columbia, and three U.S. territories.

Background

2. Primary Objectives

1. **To s**tudy association between heart disease and various factors such as demographics, medical history and behavior
2. **To develop predictive model for estimating probability of heart disease**
3. Develop a dashboard to accept values of input variables and predict the probability of heart disease

Background

3. Data

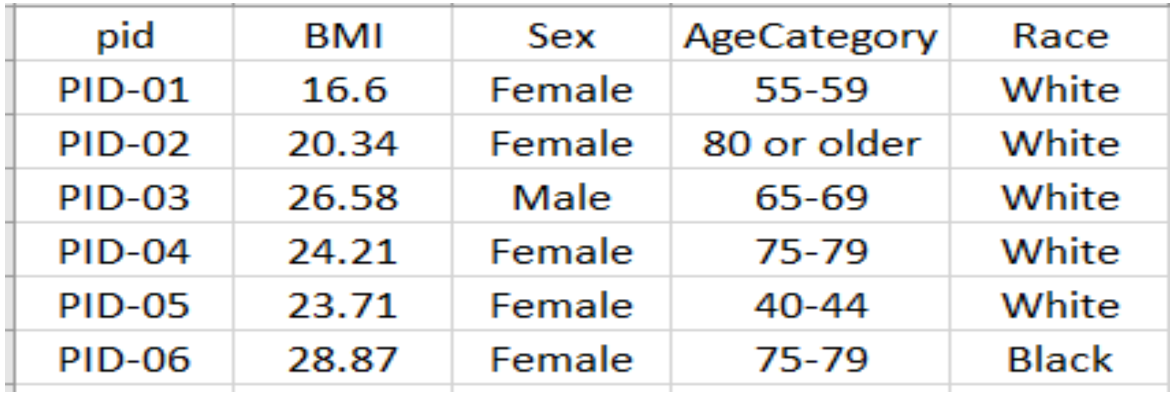
The following datasets are available:

1. Demographics
2. Behavior
3. **Medical History**
4. Heart Disease

Data : Demographics

Content

This dataset contains patients demographics



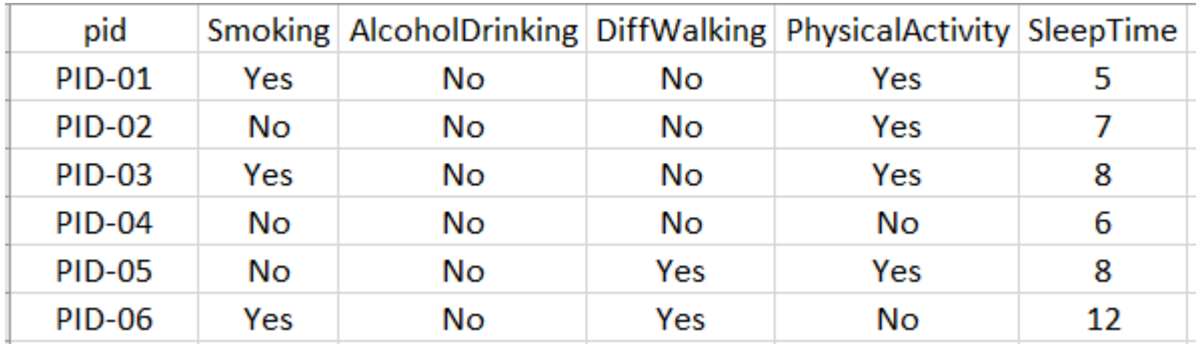
pid	BMI	Sex	AgeCategory	Race
PID-01	16.6	Female	55-59	White
PID-02	20.34	Female	80 or older	White
PID-03	26.58	Male	65-69	White
PID-04	24.21	Female	75-79	White
PID-05	23.71	Female	40-44	White
PID-06	28.87	Female	75-79	Black

Columns	Description	Type	Possible values
Pid	Patient ID	Alpha numeric	
BMI	Body Mass Index	Numeric	
Sex	Gender	Factor	Male or Female
AgeCategory	Age	Category	18-24,25-29,30-34 and so on
Race	Race	Factor	Black or White

Data : Behaviour

Content

This dataset contains patient behavior like smoking, alcohol drinking, physical activity and so on



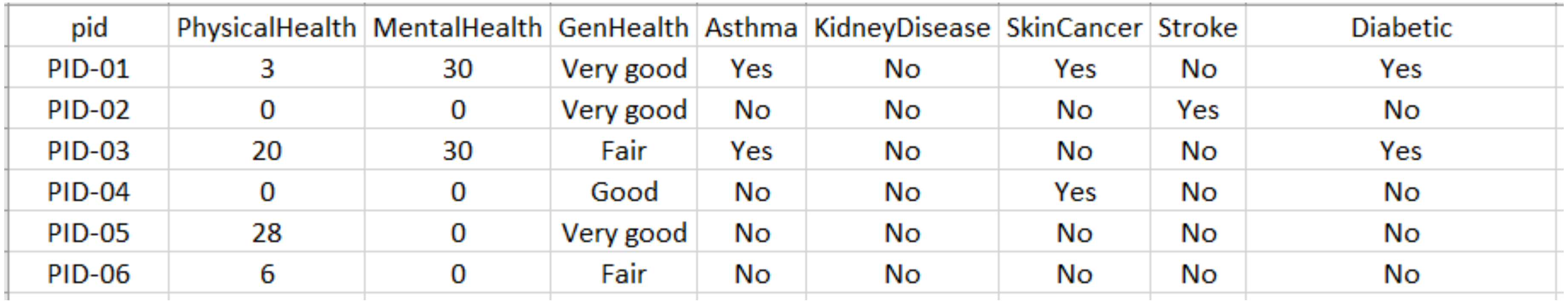
pid	Smoking	AlcoholDrinking	DiffWalking	PhysicalActivity	SleepTime
PID-01	Yes	No	No	Yes	5
PID-02	No	No	No	Yes	7
PID-03	Yes	No	No	Yes	8
PID-04	No	No	No	No	6
PID-05	No	No	Yes	Yes	8
PID-06	Yes	No	Yes	No	12

Columns	Description	Type	Possible values
Pid	Patient ID	Alpha numeric	
Smoking	Smoker	Factor	Yes or No
AlcoholDrinking	Does a patient consume alcohol	Factor	Yes or No
DiffWalking	Any difficulty in walking ?	Factor	Yes or No
PhysicalActivity	Physical activity such as running, walking, skipping, etc	Factor	Yes or No
SleepTime	Average sleep time in hours	numeric	

Data : Medical History

Content

This dataset contains medical history of patients



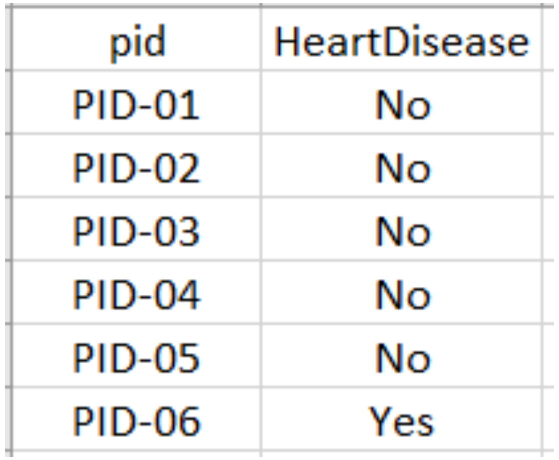
pid	PhysicalHealth	MentalHealth	GenHealth	Asthma	KidneyDisease	SkinCancer	Stroke	Diabetic
PID-01	3	30	Very good	Yes	No	Yes	No	Yes
PID-02	0	0	Very good	No	No	No	Yes	No
PID-03	20	30	Fair	Yes	No	No	No	Yes
PID-04	0	0	Good	No	No	Yes	No	No
PID-05	28	0	Very good	No	No	No	No	No
PID-06	6	0	Fair	No	No	No	No	No

Columns	Description	Type	Possible values
Pid	Patient ID	Alpha numeric	
PhysicalHealth	For how many days during the past 30 days was your physical health not good?	Numeric	0-30
MentalHealth	For how many days during the past 30 days was your mental health not good?	Numeric	0-30
GenHealth	General health of a patient	Factor	Poor, fair, good, very good, excellent
Asthma	Whether a patient is suffering from Asthma	Factor	Yes or No
KidneyDisease	Whether a patient has kidney disease	Factor	Yes or No
SkinCancer	Whether a patient has skin cancer	Factor	Yes or No
Stoke	Whether a patient has any stroke	Factor	Yes or No
Diabetic	Whether a patient is suffering from diabetes	Factor	Yes, No, Yes(during pregnancy), No, borderline

Data : Heart Disease

Content

This dataset set contains information of patients who is suffering from heart disease

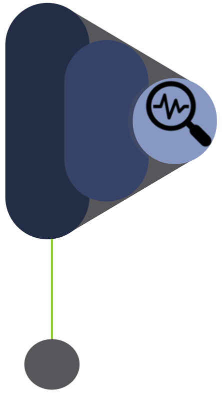


pid	HeartDisease
PID-01	No
PID-02	No
PID-03	No
PID-04	No
PID-05	No
PID-06	Yes

Columns	Description	Type	Possible values
Pid	Patient ID	Alpha numeric	
HeartDisease	Whether patient is suffering from a heart disease	Factor	Yes or No

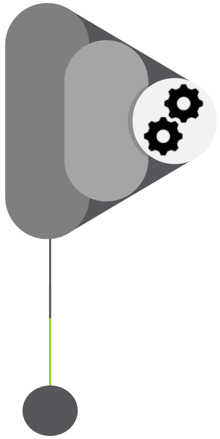
Next steps

Data management



- Compile 4 datasets using Patient ID
- Data cleaning , Handling missing values and completing Basic Data checks
- Check if any variables needed to be feature coded i.e made into groups or want to be left as continuous variables

Descriptive Statistics & Data visualization



- Summarize heart disease rate for various subgroups in the data such as gender, age group, health, etc
- Explore data for heart disease rate, which are the key indicators of heart disease
- How can this data be presented better visually ?
- Once again post Data visualization check if any variable needs to be feature coded

Predictive modelling

- Develop a model to predict the probability of heart disease
- Using different Predictive model techniques to find Significant variables
- Ensure you follow all steps like Train and test data , checking for Multicollinearity
- Check if any other ML technique fits better

DATA SCIENCE INSTITUTE