101B Final Project Datasets

1 Diabetes Health Indicators Dataset

The diabetes health indicator dataset contains 25,000 responses from the CDC's survey. The target of the dataset is to predict whether the person has the diabetes or not. Thus, the response variable is ***Diabetes_binary***, which has 2 classes. 0 stands for no diabetes, and 1 stands for prediabetes or diabetes.
There are also 16 predictors. Three of them have numeric values. ***BMI*** is the body mass index for each person. ***MentHlth*** indicates the days of poor mental health in the past 30 days. ***PhysHlth*** indicates the the days of physical illness and injury in the past 30 days. The remaining 13 predictors are categorical variables. ***HighBP*** is binary with 0 representing no high blood pressure and 1 representing high blood pressure. ***HighChol*** is binary with 0 representing no high cholesterol and 1 representing high cholesterol. ***CholCheck*** is binary indicating whether the person has cholesterol checked in 5 years. ***Smoker*** is binary indicating whether the person smoked at least 100 cigarettes in the entire life or not. ***Stroke*** is binary indicating whether the person had a stroke. ***PhysActivity*** is also binary meaning whether the person had physical activity in the past 30 days other than job. ***Fruits*** is the binary response showing whether the person consumed fruit 1 or more times per day. ***Veggie*** also shows whether they consume vegetables 1 or more times per day. ***HvyAlcoholConsump*** shows whether the person has heavy alcohol consumption habit or not. ***GenHlth*** is a self evaluation score for respondents' general health with 5 levels, from 1 with the excellent condition to 5 with poor general health. ***DiffWalk*** shows whether the person have serious difficulty waking or climbing stairs. ***Sex*** is also binary showing the sex of the person. ***Age*** has 13 levels indicating different age stages.
All of the response variable and the predictors are summarized in the following table.

Diabetes_binary	response	categorical	0 = no diabetes; 1 = prediabetes or diabetes
BMI	predictor	numeric	Body Mass Index
MentHlth	predictor	numeric	days of poor mental health scale in the past 30 days
PhysHlth	predictor	numeric	physical illness or injury days in past 30 days
HighBP	predictor	categorical	0 = no high BP; 1 = high BP
HighChol	predictor	categorical	0 = no high cholesterol; 1 = high cholesterol
CholCheck	predictor	categorical	0 = no cholesterol check in 5 years; 1 = yes cholesterol check in 5 years
Smoker	predictor	categorical	Have you smoked at least 100 cigarettes in your entire life? (0 = no; 1 = yes)
Stroke	predictor	categorical	Have you had a stroke? (0 = no; 1 = yes)
PhysActivity	predictor	categorical	physical activity in past 30 days - not including job (0 = no; 1 = yes)
Fruits	predictor	categorical	Consume Fruit 1 or more times per day (0 = no; 1 = yes)
Veggies	predictor	categorical	Consume Vegetables 1 or more times per day (0 = no; 1 = yes)
HvyAlcoholConsump	predictor	categorical	(adult men $\geq$14 drinks per week and adult women$\geq$7 drinks per week) (0 = no; 1 = yes)
GenHlth	predictor	categorical	Would you say that in general your health is: scale 1-5 (1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor)
DiffWalk	predictor	categorical	Do you have serious difficulty walking or climbing stairs? (0 = no 1 = yes)
Sex	predictor	categorical	(0 = female 1 = male)
Age	predictor	categorical	13-level age categorical

2 Heart Disease Indicators Dataset

The heart disease health indicator dataset contains 25,000 responses from the CDC's survey. The target of the dataset is to predict whether the person has the heart disease or not. Thus, the response variable is ***HeartDisease***, which has 2 classes. 0 stands for no heart disease, and 1 stands for heart disease.
There are also 16 predictors. Four of them have numeric values. ***BMI*** is the body mass index for each person. ***MentalHealth*** indicates the days of poor mental health in the past 30 days. ***PhysicalHeallth*** indicates the the days of physical illness and injury in the past 30 days. ***Sleeptime*** indicates the average hours of sleep in a 24-hour period. The remaining 12 predictors are categorical variables. ***Asthma*** is binary with 0 representing no asthma and 1 representing asthma. ***KindeyDisease*** is

binary with 0 representing no kidney disease and 1 representing kidney disease. ***SkinCancer*** is binary indicating whether the person has skin cancer or not. ***Stroke*** is binary indicating whether the person had a stroke. ***Smoking*** is binary indicating whether the person is a heavy smoker or not. ***AlcoholDrinking*** shows whether the person has heavy alcohol consumption habit or not. ***PhysicalActivity*** is also binary meaning whether the person had physical activity in the past 30 days other than job. ***GenHealth*** is a self evaluation score for respondents' general health with 5 levels, from 1 with the excellent condition to 5 with poor general health. ***DiffWalking*** shows whether the person have serious difficulty waking or climbing stairs. ***Sex*** is also binary showing the sex of the person. ***AgeCategory*** has 13 levels indicating different age stages. ***Race*** has 6 levels showing the ethnicity of the person.
All of the response variable and the predictors are summarized in the following table.

HeartDisease	response	category	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) (0= no; 1 = yes)
BMI	predictor	numeric	Body Mass Index (BMI)
PhysicalHealth	predictor	numeric	how many days during the past 30 days was your physical health not good? (0-30 days)
MentalHealth	predictor	numeric	how many days during the past 30 days was your mental health not good? (0-30 days)
SleepTime	predictor	numeric	On average, how many hours of sleep do you get in a 24-hour period?
Asthma	predictor	category	(Ever told) you had asthma? (2 levels)
KidneyDisease	predictor	category	were you ever told you had kidney disease? (2 levels)
SkinCancer	predictor	category	(Ever told) you had skin cancer? (2 levels)
Stroke	predictor	category	(Ever told) you had a stroke? (2 levels)
Smoking	predictor	category	have you smoked at least 100 cigarettes in your entire life? (2 levels)
AlcoholDrinking	predictor	category	Heavy drinkers or not (2 levels)
PhysicalActivity	predictor	category	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job (2 levels)
GenHealth	predictor	category	Would you say that in general your health is. (5 levels)
DiffWalking	predictor	category	Do you have serious difficulty walking or climbing stairs? (2 levels)
Sex	predictor	category	Are you male or female? (2 levels)
AgeCategory	predictor	category	Fourteen-level age category
Race	predictor	category	race/ethnicity (6 levels)

3 Cardiovascular Diseases Dataset

The cardiovascular diseases dataset consists of records of 25,000 patients and the target is to predict whether the patients have the cardiovascular diseases or not. The response variable is ***Cardio_Disease*** indicating the presence or absence of cardiovascular disease.
There are also 11 predictors. Five of them have numeric values. ***Age*** shows the age of the person which takes integer values. ***Height*** and ***Weight*** show the height and weight of the patients in centimeter and kilogram, respectively. ***AP_high*** and ***AP_low*** shows the systolic and diastolic blood pressure of the patient. The remaining six predictors are categorical. ***Gender*** is binary indicating the sex. ***Cholesterol*** has three levels with 1 meaning normal cholesterol level, 2 meaning above normal cholesterol level and 3 meaning well above normal cholesterol level. ***Glucose*** also has three levels indicating normal, above normal and well above normal glucose level, similar as the Cholesterol predictor. ***Alcohol*** is binary indicating whether the person consume alchohol or not. ***Pyhsical_Activity*** is also binary meaning whether the person did physical activity or not.
All of the response variable and the predictors are summarized in the following table.

Cardio_Disease	response	categorical	0: no, 1: yes
Age	predictor	numeric	years of age
Height	predictor	numeric	height in cm
Weight	predictor	numeric	weight in kg
AP_high	predictor	numeric	systolic blood pressure
AP_low	predictor	numeric	diastolic blood pressure
Gender	predictor	categorical	1: female, 2: male
Cholesterol	predictor	categorical	1: normal, 2: above normal, 3: well above normal
Glucose	predictor	categorical	1: normal, 2: above normal, 3: well above normal
Smoke	predictor	categorical	0: no, 1: yes)
Alcohol	predictor	categorical	0: no, 1: yes
Physical_Activity	predictor	categorical	0: no, 1: yes