Capstone Project
Heart Disease Prediction

Yumemi Kinsella

Agenda

01. Overview of the subject area and the problem statement

02. Overview of my proposed vision

03. Estimate of the potential impact

04. Introduction to the dataset

05. Next steps in terms of data processing

01. Overview of the subject area and the problem statement

- Heart Disease and Medical Bills in the US

Firstly, according to the Centers for Disease Control and Prevention (CDC), about 695,000 people die of heart disease in the United States in 2021 – that's 1 in every 5 deaths.

Secondly, about 22% of Americans have avoided some sort of medical care — including doctor visits, medications, vaccinations, annual exams, screenings, vision checks and routine blood work — because of the expense

02. Overview of My Proposed Vision

1.Identifying the problem and defining the purpose of this project

Researching the current situation about heart disease and medical expense, such as "how many people in the US die from heart disease", "what percentage of American population are low-income household" and so forth. Then, deciding specific target of this project.

2. Identifying relevant data sources and collecting data

Based on the target which I defined earlier, identifying relevant data. Then, collecting best dataset for this project.

3. Understand the data and data preparation

For understanding the data, doing EDA. In EDA, I generate visualizations and graphs to find relationships between the data that is collected. After understanding the data, the next step is preparing the data. Checking missing data and columns, substituting the data, and removing the data that is not required.

4. Developing predictive models and evaluate them

The target variable is that if you have heart disease, and it shows "Yes / No" values. Therefore, machine learning approach is categorical prediction. Developing predictive models with right method and predicting the probability of having a heart disease.

5. Creating an application

Based on the predictive model, creating an application which people can know the possibility of having a heart disease. Giving some advice based on their information which they enter into the application and the results.

03. Estimate of the potential impact

The target user : Ordinally people who cannot afford medical care due to their low-income.

Impact : People don’t have to spend money for medical bills if you don’t suspect that you have heart disease. On the other hand, it can lead people who suspect that you have heart disease to go see a doctor and reduce the number of deaths from heart disease. People also can change their lifestyle or habit to make the chances of getting heart disease lower if they need.

04. Introduction to the dataset

- **The dataset**: CDC
- **The respondents**: The residents of the United States

- **Numerical variables**: BMI, PhysicalHealth, MentalHealth, SleepTime
- **Categorical variables**: Smoking, AlcoholDrinking, DiffWalking, Sex, AgeCategory, Race, PhysicalActivity, GenHealth, Asthma, HeartDisease

- **The Target variable**: HeartDisease
  - **Cleaned dataset:** "Yes" : 8.6%(27,269rows) / "No" : 91.6%(291,804 rows)
  - **Sampled dataset:** "Yes" : 50%(27,269 rows) / "No" : 50%(27,269 rows)

04. Introduction to the dataset

The Cleaned Dataset: 319073 Respondents' data

Numerical variables

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	319073.000000	319073.000000	319073.000000	319073.000000
mean	28.264732	3.355618	3.887211	7.088594
std	6.167446	7.929974	7.940627	1.394768
min	12.020000	0.000000	0.000000	1.000000
25%	24.030000	0.000000	0.000000	6.000000
50%	27.320000	0.000000	0.000000	7.000000
75%	31.380000	2.000000	3.000000	8.000000
max	59.970000	30.000000	30.000000	16.000000

04. Introduction to the dataset

The Sampled Dataset: 54538 Respondents' data

Numerical variables

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	54538.000000	54538.000000	54538.000000	54538.000000
mean	28.758882	5.351681	4.218453	7.101452
std	6.291098	9.939388	8.515041	1.541498
min	12.210000	0.000000	0.000000	1.000000
25%	24.390000	0.000000	0.000000	6.000000
50%	27.790000	0.000000	0.000000	7.000000
75%	32.077500	5.000000	3.000000	8.000000
max	59.970000	30.000000	30.000000	16.000000

04. Introduction to the dataset

The Cleaned Dataset: 31907 Respondents' data

Categorical variables

	HeartDisease	Smoking	AlcoholDrinking	DiffWalking	Sex	AgeCategory	Race	PhysicalActivity	GenHealth	Asthma
count	319073	319073	319073	319073	319073	319073	319073	319073	319073	319073
unique	2	2	2	2	2	13	6	2	5	2
top	No	No	No	No	Female	65-69	White	Yes	Very good	No
freq	291804	187479	297340	275038	167368	34099	244825	247606	113770	276395

04. Introduction to the dataset

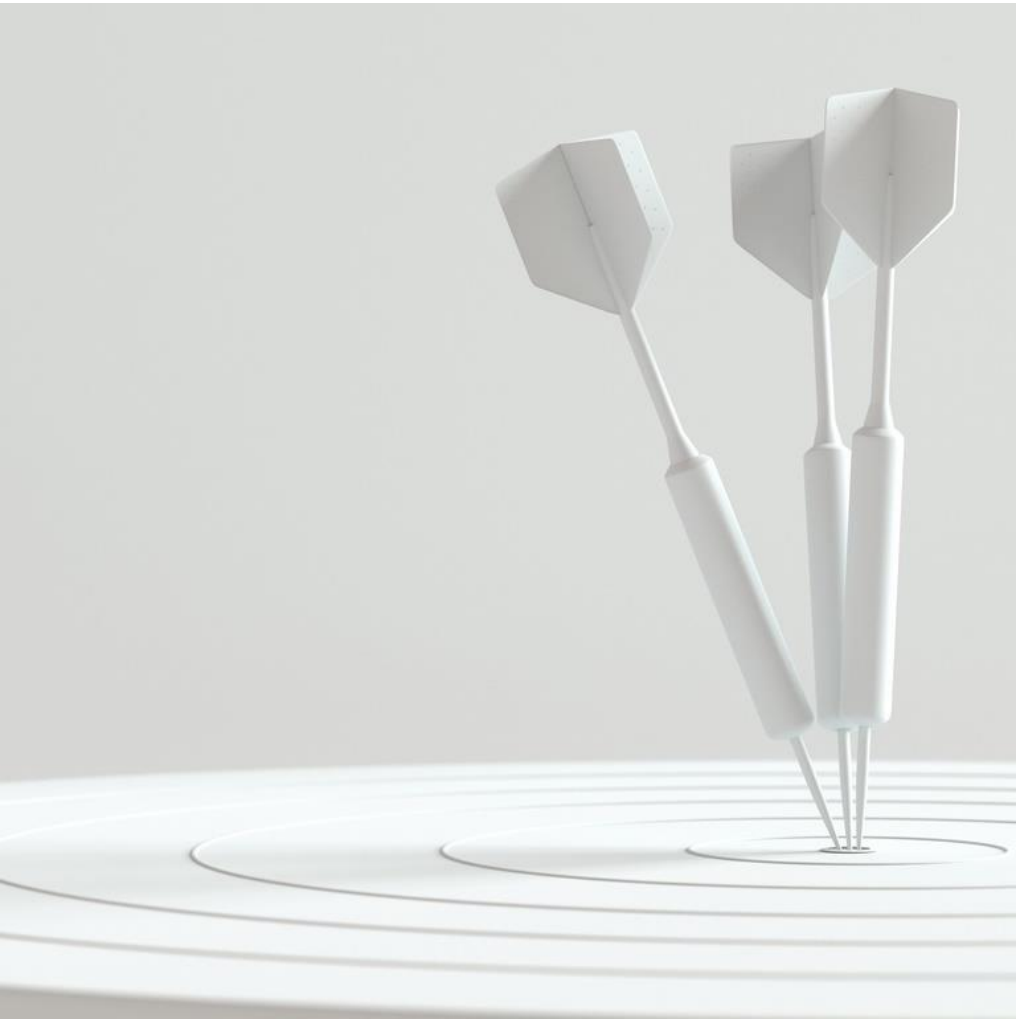
The Sampled Dataset: 54538 Respondents' data

Categorical variables

	Smoking	AlcoholDrinking	DiffWalking	Sex	AgeCategory	Race	PhysicalActivity	GenHealth	Asthma	HeartDisease
count	54538	54538	54538	54538	54538	54538	54538	54538	54538	54538
unique	2	2	2	2	13	6	2	5	2	2
top	No	No	No	Male	70-74	White	Yes	Good	No	No
freq	27829	51529	41442	28705	7351	43296	38874	17317	46108	27269

05. Next steps in terms of data processing

- **Deciding Machine Learning Method**
- **Data transformation**: The dataset still has many categorical variables such as "Yes/No" or "Male/Female", so converting those values into 0/1 or getting dummies if I need it.
- **Developing predictive models and evaluate them**



Thank You
For Your Attention