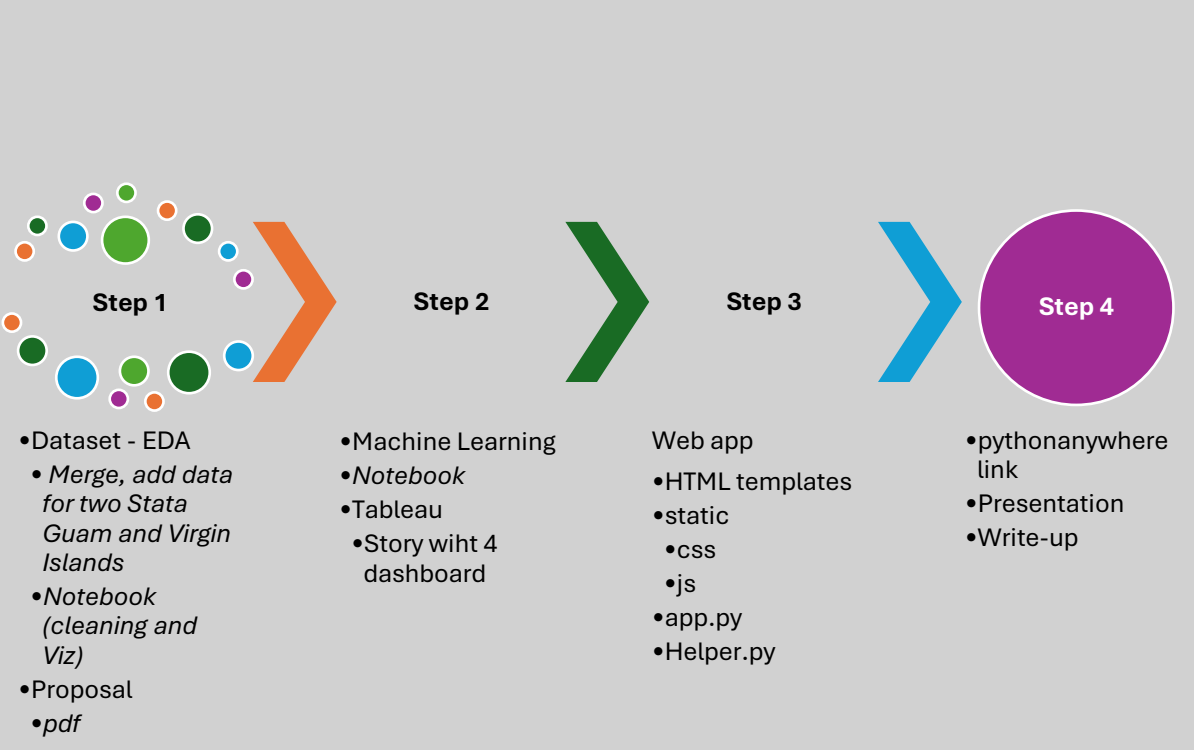
Write-Up
Project 4-Group 12

Gavin Plemon, Isbelis Castro, Sam Hoemann, Stephen Ferrier
October 3th, 2024

1. Introduction

According to C.D.C Heart disease is the leading cause of death for men, women, and most racial/ethnic groups in the United States, including African Americans, Hispanics, and whites. About 660,000 people in the U.S. die from heart disease every year—that's 1 in every 4 deaths and on average 1 person every 40 seconds. For this reason, in this project involved to work on heart attack prediction using machine learning and Tableau because of the growing need for accurate and data-driven healthcare solutions. Machine learning allows us to analyze complex clinical and demographic data, uncovering patterns that can improve early prediction and intervention for heart attack risks. By integrating this with Tableau, we can create intuitive and interactive visualizations that make the insights from machine learning models accessible to healthcare providers and decision-makers. This combination offers a powerful approach to enhancing heart disease prevention, ultimately improving patient outcomes.



2. Data Cleaning

Dataset:

1. heart_2022_no_nans: https://www.kaggle.com/code/alibinkashif/heart-disease-indicators-eda/input?select=2022
2. heart_2020_cleaned: https://www.kaggle.com/code/alibinkashif/heart-disease-indicators-eda/input?select=2020
3. US_GeoCode.csv: https://simplemaps.com/data/us-zips
4. world_country_and_usa_states_latitude_and_longitude_values.csv (add Guam and Virgin Islands): https://www.kaggle.com/datasets/paultimothymooney/latitude-and-longitude-for-every-country-and-state

Cleaning process:

During this project, we cleaned the dataset **`heart_2022_no_nans`,** which was updated by adding latitude and longitude columns. The process included merging it with the dataset **`US_GeoCode.csv`** and manually adding data for Guam and the Virgin Islands from **`world_country_and_usa_states_latitude_and_longitude_values.csv`**. This dataset was prepared to create visualizations and a story dashboard in Tableau.

On the other hand, the dataset **`heart_2020_cleaned`** did not need any cleaning, as it was used directly for the machine learning experiment.

3. Color design considerations

In this project for overall outputs of visualizations as EDA, Dashboards (Tableaus) and Webpage. They were used the following colors and design:

- Colors – Autumn
- Orange
- Red
- Brown
- Bootswatch 4.5.2 -united

4. ML Experiment

Creating the Model

Given the data we used was already cleaned the only major step to getting the data model ready was to encode the various categorical data and then to scale the few numerical data columns.

Data Example

HeartDise	BMI	Smoking	AlcoholDr	Stroke	PhysicalH	MentalHe	DiffWalkir	Sex	AgeCateg	Race	Diabetic	PhysicalA	GenHealth	SleepTime	Asthma	KidneyDis	SkinCance
No	16.6	Yes	No	No	3	30	No	Female	55-59	White	Yes	Yes	Very good	5	Yes	No	Yes
No	20.34	No	No	Yes	0	0	No	Female	80 or olde	White	No	Yes	Very good	7	No	No	No
No	26.58	Yes	No	No	20	30	No	Male	65-69	White	Yes	Yes	Fair	8	Yes	No	No
No	24.21	No	No	No	0	0	No	Female	75-79	White	No	No	Good	6	No	No	Yes
No	23.71	No	No	No	28	0	Yes	Female	40-44	White	No	Yes	Very good	8	No	No	No
Yes	28.87	Yes	No	No	6	0	Yes	Female	75-79	Black	No	No	Fair	12	No	No	No
No	21.63	No	No	No	15	0	No	Female	70-74	White	No	Yes	Fair	4	Yes	No	Yes
No	31.64	Yes	No	No	5	0	Yes	Female	80 or olde	White	Yes	No	Good	9	Yes	No	No
No	26.45	No	No	No	0	0	No	Female	80 or olde	White	No, borde	No	Fair	5	No	Yes	No
No	40.69	No	No	No	0	0	Yes	Male	65-69	White	No	Yes	Good	10	No	No	No
Yes	34.3	Yes	No	No	30	0	Yes	Male	60-64	White	Yes	No	Poor	15	Yes	No	No

Info Pre-Encoding

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 319795 entries, 0 to 319794
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   HeartDisease      319795 non-null  int64
 1   BMI               319795 non-null  float64
 2   PhysicalHealth    319795 non-null  float64
 3   MentalHealth      319795 non-null  float64
 4   SleepTime         319795 non-null  float64
 5   AgeCategory       319795 non-null  float64
 6   Smoking           319795 non-null  int64
 7   AlcoholDrinking   319795 non-null  int64
 8   Stroke            319795 non-null  int64
 9   DiffWalking       319795 non-null  int64
 10  Sex               319795 non-null  int64
 11  Race              319795 non-null  int64
 12  Diabetic          319795 non-null  int64
 13  PhysicalActivity  319795 non-null  int64
 14  GenHealth         319795 non-null  int64
 15  Asthma            319795 non-null  int64
 16  KidneyDisease     319795 non-null  int64
 17  SkinCancer        319795 non-null  int64
dtypes: float64(5), int64(13)
memory usage: 43.9 MB
```

Info Post-Encoding

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 319795 entries, 0 to 319794
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   HeartDisease      319795 non-null  object
 1   BMI               319795 non-null  float64
 2   Smoking           319795 non-null  object
 3   AlcoholDrinking   319795 non-null  object
 4   Stroke            319795 non-null  object
 5   PhysicalHealth    319795 non-null  float64
 6   MentalHealth      319795 non-null  float64
 7   DiffWalking       319795 non-null  object
 8   Sex               319795 non-null  object
 9   AgeCategory       319795 non-null  object
 10  Race              319795 non-null  object
 11  Diabetic          319795 non-null  object
 12  PhysicalActivity  319795 non-null  object
 13  GenHealth         319795 non-null  object
 14  SleepTime         319795 non-null  float64
 15  Asthma            319795 non-null  object
 16  KidneyDisease     319795 non-null  object
 17  SkinCancer        319795 non-null  object
dtypes: float64(4), object(14)
memory usage: 43.9+ MB
```

All the objects you see above in the info pre-encoding were encoded 1 or 0 as they are yes or no answers, with the exceptions of:

- Age category was averaged between the range categories given then added to the numerical features to be scaled
- GenHealth was on a 5-category scale from poor to very good and was assigned 5 different numbers
- Race was changed to 6 different numbers corresponding to the 6 race options given in the data

Then we ran through all the available models to find the best model to predict whether someone had heart disease to find whether someone was at a high risk. What we were looking for specifically in the model were the ones that had the highest predicted positive correct result as we felt that predicted positive and false were not as important as a predicted negative and false. Now we still wanted the accuracy in those categories to be within reason, but we wanted to make sure as many people as we could who are high risk got a positive test result. Given this we chose the decision tree classifier model as it gave the highest tested score on accuracy for predicted positive results, at 25%. Now this is very low but going in we knew that this was going to be low. Predicting diseases based on general health and lifestyle questions is inherently inaccurate. High accuracies can really only be obtained by extensive medical testing with detailed medical histories. This data is better for finding correlating data than it is for predicting data.

Putting Model Into Website and Model Interaction With User

The website is designed to take in a set of inputs that correspond to the categories that the model trained on, through a series of dropdown menus and numerical input slots. Once the submit button is hit the logic.js file assigns all the variables from the form, which then makes a request to the app.py with ajax. The app.py then reassigns the variables with correct data types so that the variables can be correctly called by the modelHelper function. Which then in turn loads the pickled model and scaler and gives an output string to be displayed on the website, giving the predicted result to whoever submitted the form.

5. Dashboard design concepts

Dashboard Structure

The structure of the dashboard is a Story dashboard, which contains four sections titled as follows:

- General Information
- Geographical
- Individual Factor
- Healthcare and Epidemic Factor

Before creating the dashboard, the dataset used was `**heart_2022_no_nans**`, which was merged
with `**US_GeoCode.csv**` and `**world_country_and_usa_states_latitude_and_longitude_values.csv**`. The latter was included to add data for Guam and the Virgin Islands, but it was not necessary to perform a merge for this purpose.

Organization of columns:

a. Group calculation Diseases type:

WHO Group	Subclassification	Variables
Noncommunicable Diseases (NCDs)	**Cardiovascular Diseases**	HadHeartAttack, HadAngina, HadStroke
	Respiratory Diseases	HadAsthma, HadCOPD
	Cancers	HadSkinCancer
	Metabolic Disorders	HadDiabetes
	Musculoskeletal Diseases	HadArthritis
	Kidney Diseases	HadKidneyDisease
Mental Health Disorders	**Mental Health Disorders**	HadDepressiveDisorder
Disability and Functional Difficulties	**Sensory Impairments**	DeafOrHardOfHearing, BlindOrVisionDifficulty
	Cognitive and Physical Disabilities	DifficultyConcentrating, DifficultyWalking, DifficultyDressingBathing, DifficultyErrands

To create groups in Tableau, subclassification was used for diseases because it is more understandable. In contrast, classification was applied to the columns for disabilities and sensory impairments. Finally, the visualization should provide more information about the relationship between heart attacks and multiple factors. The classification was based on the World Health Organization (WHO).

b. Group calculation preventive health:

This group includes all columns about vaccination, chest scan, and HIV testing.

Filters:

The filters used in the story dashboard were "Had Heart Attack" as the target variable, along with "Sex," "Race," "Ethnicity Category," and "State." These filters allowed for the examination of the associations between different factors and the occurrence of heart attacks.

Finally, Enjoy story dashboard the following link:

https://public.tableau.com/app/profile/isbelis.castro/viz/Project_4_Group_12/HeartDiseases

6. How does your dashboard and ML answer any research questions?

Machine Learning:

1. What factors indicate the highest likelihood for heart disease
2. Can we use a machine learning model to accurately predict based on health factors whether someone has or is likely to get heart disease

Dashboard:

1. What are the key demographic factors (age, gender, ethnicity) that significantly influence heart attack risk?

   This dashboard provides information about various demographic factors derived from the 2022 dataset of survey respondents from the Behavioral Risk Factor Surveillance System (BRFSS). It shows that over 60% of respondents who had a heart attack were aged 60 to over 80. Additionally, the story dashboard reveals that the highest percentage of white respondents who had a heart attack is in the age category of over 80, at 20.36% (n=2,153), while the highest percentages for other races, including Multiracial, Black, and Hispanic respondents, occur in the 65 to 69 Age category. However, gender does not show a statistically significant relationship with heart attacks in this analysis.

   The geographical factors highlighted in this dashboard emphasize the principal states. It provides a visual analysis by state, displaying various maps that illustrate heart attacks, total diseases, preventive actions, and health behaviors. Notably, the top five states with the highest percentage of heart attacks are Washington at 5.22%, Ohio at 4.38%, Florida at 4.15%, Maryland at 3.69%, and Texas at 3.17%. The concentration of total diseases features the same top states, except for Minnesota, which replaces Texas in the last position. Thus, this geographical dashboard offers valuable insights based on areas of interest and targets.

   According to the visualizations, the story dashboard indicates significant disparities in heart attack occurrences based on demographic and geographical factors in the U.S.

2. How do lifestyle factors (smoking, physical activity, diet) and medical history (Musculoskeletal, Infectious, Respiratory, Mental Health, Cardiovascular, Metabolic disorders, Cancers, Kidney diseases) correlate with heart attack occurrences in different demographic groups?

The Individual Factor and Healthcare and Epidemic Factor indicate that people who had a heart attack perceive their general health status differently, with 42.44% rating it as good and 34.18% as fair, while the percentage for excellent is the lowest. In contrast, among those who did not have a heart attack, 49.32% rated their health as good and 27.77% as excellent, with poor being the lowest percentage. There was a significant change in the percentage of individuals rating their health as excellent, dropping from 26.24% to 5.43%, while the fair category remained at 34.18%.

Additionally, the dashboards provide information about the presence of other diseases in individuals who had a heart attack. However, no disease exceeded 20% prevalence, except for cardiovascular diseases, which were reported by 59% of respondents. In other words, 41% of the respondents had other types of diseases, such as stroke.

Furthermore, around 50% of individuals who had a heart attack reported that their mental health days were not good, which is less than those who did not have a heart attack. The highest percentage of poor mental health days—50%—was reported for 15 days, while 28 physical health days were also reported as not good.

However, individual behaviors show a predominance of alcohol consumption, but this visualization does not indicate any relationship with having a heart attack. Notably, 57.80% of respondents who had a heart attack also had musculoskeletal diseases (Arthritis), and 40.88% had respiratory diseases (Asthma). These findings are significant because each group contains only one disease type, meaning there are no duplicates. Overall, health outcomes indicate a possible correlation with heart attacks.

7. Bias/Limitations

There are two primary limitations to this project. The data only covers patients in the United States and there was limited time to conduct tests. First is that the data itself is limited.
While the United States has a large population, it is not large enough to get an idea of how the data would be affected around the world. The factors that contribute to heart problems in the United States may not exist in the same manner as other countries. Along with this, there was limited time to gather data, run tests, and generate a well performing product. To counter this, we chose quick and simple solutions that were manageable to implement within the given time frame. While this method did allow us to produce an informative product, it also limited the functionality.

The primary bias to this project ties into the limited time. Anchoring bias occurs when the first or most available piece of data is used without considering other sources. Due to the time constraints, we did not have long to decide on what kind of project we were going to complete. This type of bias narrows how much the data can inform users.

8. Conclusions/Reflection

- ✓ The dashboard highlights important demographic and geographical factors related to heart attack occurrences based on the 2022 BRFSS dataset. It shows that a significant number of individuals who had heart attacks are aged 60 and older, with variations in prevalence among different racial groups. The geographical analysis identifies states with the highest heart attack rates, indicating where public health efforts should focus. Overall, the dashboard emphasizes the need to understand the complexities of heart health disparities in the U.S. to guide targeted interventions and resource allocation.

- ✓ the analysis indicates that heart attack survivors generally perceive their health status more negatively than those who have not experienced a heart attack, with fewer rating their health as excellent. There is a notable prevalence of cardiovascular diseases among these individuals, along with other health issues such as musculoskeletal and respiratory diseases. Mental health is also impacted, with many survivors reporting poor mental health days. Although alcohol consumption is common among respondents, its connection to heart attacks is not clearly established. Overall, the findings highlight the complex relationship between heart attacks and various health outcomes, pointing to the necessity for comprehensive strategies to address both physical and mental health in affected individuals.

9. References

- ➢ https://public.tableau.com/app/search/vizzes/heart%20attack
- ➢ https://public.tableau.com/app/profile/adrian.tan2691/viz/HeartAttackDistributions/Demographics
- ➢ https://www.quantizeanalytics.co.uk/tableau-healthcare-dashboard-examples/
- ➢ https://www.analyticsvidhya.com/blog/2022/06/machine-learning-for-heart-disease-prediction/
- ➢ https://www.analyticsvidhya.com/blog/2022/02/heart-disease-prediction-using-machine-learning-2/
- ➢ https://www.analyticsvidhya.com/blog/2022/02/heart-disease-prediction-using-machine-learning/
- ➢ https://github.com/g-shreekant/Heart-Disease-Prediction-using-Machine-Learning/blob/master/Heart_disease_prediction.ipynb
- ➢ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10378171/