

Milestone 3 (Preliminary Analysis) - Week 6

Milestone 3 should include all the information from Milestone 2, updated if necessary.

Milestone #2

- **What types of model or models do you plan to use and why?**
 - A logistic model will be used on the dataset to determine which features are most related to or most predictive of heart disease, where the HeartDisease feature is the target. This will be a supervised learning model where “the supervisor is the target variable, a column in the data representing values to predict from other columns in the data. The target variable is chosen to represent the answer to a question the organization would like to answer or a value unknown at the time the model is used that would help in decisions. Sometimes supervised learning is also called predictive modeling. The primary predictive modeling algorithms are classification for categorical target variables or regression for continuous target variables.” (Abbott, 2014, p. 5).
 - In addition to the logistic model with 1 target and 17 features, a second logistic model will be used with the 5 best features only, where the 5 features are derived with highest chi-squared statistics.
- **How do you plan to evaluate your results?**
 - I plan to calculate the accuracy, precision, recall and F1 score of both the logistic regression model with 17 features and best 5 features, as well as visualize a confusion matrix and an ROC curve where "The area under the ROC curve (AUC) results were considered excellent for AUC values between 0.9-1, good for AUC values between 0.8-0.9, fair for AUC values between 0.7-0.8, poor for AUC values between 0.6-0.7 and failed for AUC values between 0.5-0.6." (El Khouli, 2009). All of these together will be used to evaluate the results of the logistic regression models.

- **What do you hope to learn? Why is this data useful to solve the problem?**
 - This data references individual lifestyle and other disease information, which is useful as they could have an impact on heart disease risk potential. I hope to learn what features in the dataset, if any, are predictive of heart disease.
- **Assess any risks with your proposal**
 - Logistic regression models are not perfect and have their disadvantages, “Logistic Regression is a statistical analysis model that attempts to predict precise probabilistic outcomes based on independent features. On high dimensional datasets, this may lead to the model being over-fit on the training set, which means overstating the accuracy of predictions on the training set and thus the model may not be able to predict accurate results on the test set. This usually happens in the case when the model is trained on little training data with lots of features. So on high dimensional datasets, Regularization techniques should be considered to avoid over-fitting (but this makes the model complex). Very high regularization factors may even lead to the model being under-fit on the training data.” (Grover, 2020). If overfitting is identified as an issue via cross-validation, two types of regularization techniques will be used: Lasso (L1 Regularization) or Ridge (L2 Regularization), where λ is called the regularization parameter and controls the trade-off between fitting the training data well and keeping the parameters small to avoid overfitting.
- **Identify a contingency plan if your original project plan does not work out**
 - If this plan does not work out, the contingency plan will be to use a similar medical dataset on HCC (Hepatocellular Carcinoma dataset) survival. This dataset has more features (50) than the selected Heart Disease dataset (17), with less samples (165). The target variable would be the “Class” feature (nominal - 1 if patient survives, 0 if patient

died), and a logistic regression model would be used to predict patient survival from HCC.

- **Include anything else you believe is important:**
- **Problem Statement**
 - Which dataset feature(s) are most related to or most predictive of heart disease (single characteristics and interactions)?
- **Explain why the problem is important/interesting**
 - This problem is important because many people die of heart disease every year “Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One person dies every 36 seconds in the United States from cardiovascular disease. About 659,000 people in the United States die from heart disease each year—that’s 1 in every 4 deaths. Heart disease costs the United States about \$363 billion each year from 2016 to 2017. This includes the cost of health care services, medicines, and lost productivity due to death.” (CDC, 2022).

Understanding what factors contribute to heart disease could save lives and healthcare costs.
- **Who would be interested in solving this problem, i.e., who would you be trying to sell this project to?**
 - Doctors, healthcare professionals, patients, and life & health insurance companies would all be interested in understanding what factors are most predictive of heart disease. Life & health insurance companies would be interested in solving this problem to better understand how risky a particular customer would be to insure. Doctors, healthcare professionals and patients would be interested to ensure to be able to advise

on heart disease risk factors and reduce heart disease events (coronary heart disease (CHD) or myocardial infarction (MI)).

- **Where did you get your data?**

- This dataset was obtained from an open source data website, Kaggle

(<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?resource=download>). The dataset contains 319,795 rows and 18 columns or features (9 booleans, 5 strings and 4 decimals), which are:

- HeartDisease: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) - either yes or no
- BMI: Body Mass Index (BMI) - float
- Smoking: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] - either yes or no
- AlcoholDrinking: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week - either yes or no
- Stroke: (Ever told) (you had) a stroke? - either yes or no
- PhysicalHealth: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days have you exercised - integer
- MentalHealth: Thinking about your mental health, for how many days during the past 30 days was your mental health not good? - integer
- DiffWalking: Do you have serious difficulty walking or climbing stairs? - either yes or no
- Sex: Are you male or female? - string
- AgeCategory: Fourteen-level age category - string

- Race: Imputed race/ethnicity value- string
- Diabetic: (Ever told) (you had) diabetes? - either yes, no, no-borderline diabetes, or yes pregnancy
- PhysicalActivity: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job - either yes or no
- GenHealth: Would you say that in general your health is... - Excellent, Very Good, Good, Fair, Poor
- SleepTime: On average, how many hours of sleep do you get in a 24-hour period? - integer
- Asthma: (Ever told) (you had) asthma? - either yes or no
- KidneyDisease: Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease? - either yes or no
- SkinCancer: (Ever told) (you had) skin cancer? - either yes or no

Milestone #3

- **Will I be able to answer the questions I want to answer with the data I have?**
 - **Problem Statement:** Which dataset feature(s) are most related to or most predictive of heart disease (single characteristics and interactions)?
 - Yes, I will be able to answer my question with the dataset that I have. The dataset consists of 17 features (both numerical and categorical) with “HeartDisease” being the target (Yes/No option transformed into 1/0, respectively). There are no missing data points, and the categorical features will have dummy variables for the modeling.

Observations

- The dataset has 319795 rows and 18 columns
- All of the features will be useful in this analysis. Any categorical features will have dummy variables assigned to them for modeling
- There is a mix of numerical and categorical data
- The target of the model will be 'HeartDisease': No means no heart disease reported, Yes means heart disease reported

○

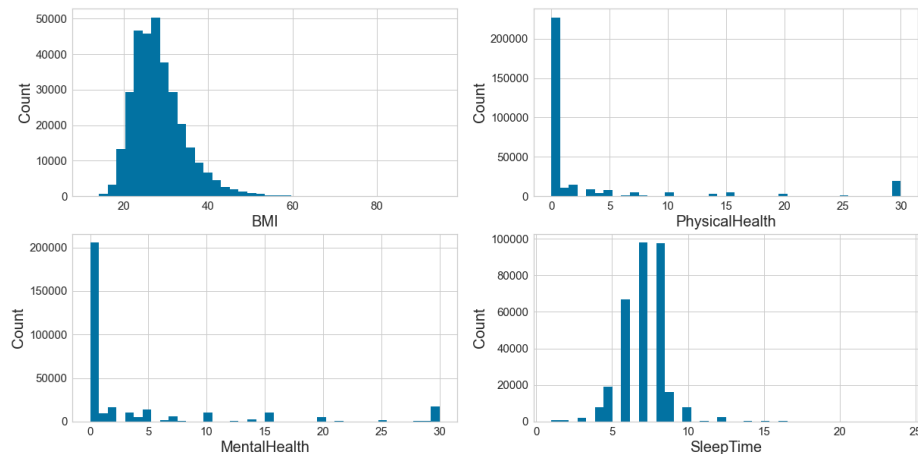
Observations

- There is no missing data to deal with (all features have 319795 observations which make them complete)

○

• What visualizations are especially useful for explaining my data?

- Histograms, bar charts, Pearson's correlation matrix (numerical data) and Spearman's correlation matrix (categorical & numerical data) will be used to explain my data.
- **Numerical Data:** To show the distribution of the numerical data (BMI, PhysicalHealth, MentalHealth, and SleepTime), histograms were used

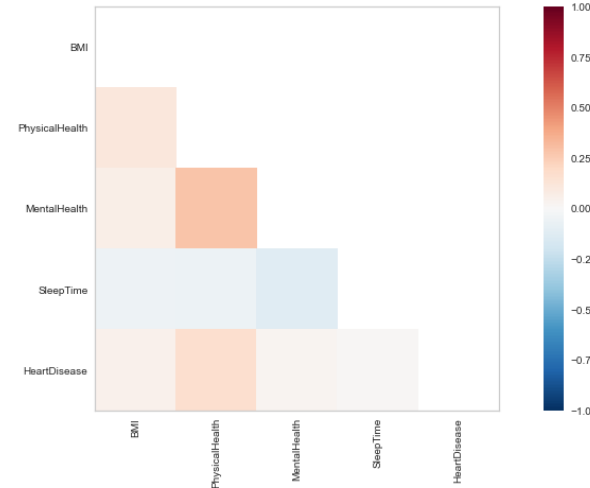


Observations

- Most people in the dataset have a BMI between 20 and 40
- Most people in the dataset exercise between 0-5 days in the past 30 days
- Most people in the dataset have 0-5 bad mental health days in the past 30 days
- Most people in the dataset get between 5-9 hours of sleep a night

○

- A Pearson's Correlation Matrix was then created to understand the correlation between 'HeartDisease' and the numerical features (BMI, PhysicalHealth, MentalHealth, SleepTime)



Observations

- The numerical feature most highly correlated with 'HeartDisease' is 'PhysicalHealth'
- There is not significant collinearity between the features (Rule of thumb: If the correlation > 0.8 then severe multicollinearity may be present)

- **Categorical Data:** In this section, we will start out first with the Spearman's correlation

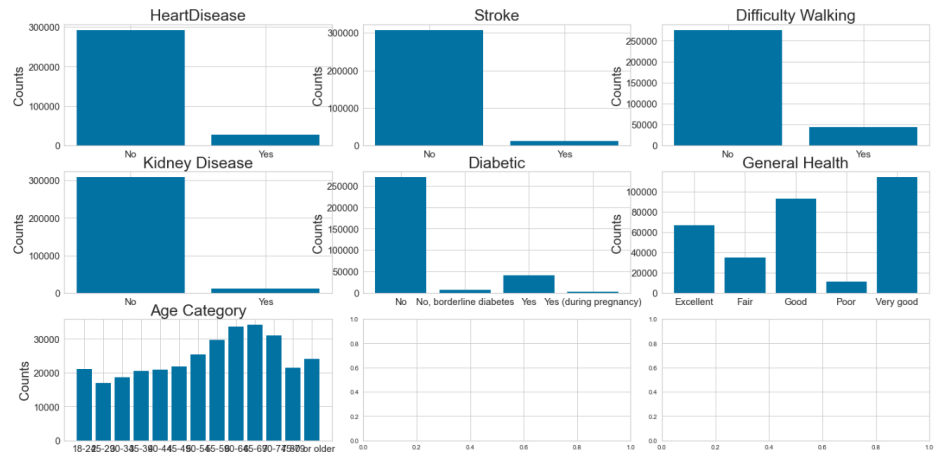
matrix to find the features most correlated to the target variable, 'HeartDisease'

	HeartDisease	BMI	PhysicalHealth	MentalHealth	SleepTime	Smoking_No	Smoking_Yes	AlcoholDrinking_No	AlcoholDrinking_Yes	Stroke_No	Stroke_Yes	DiffWalking_No	DiffWalking_Yes	Sex_Female	Sex_Male	AgeCategory_18-24	AgeCategory_25-29	AgeCategory_30-34
HeartDisease	1.00000	0.027355	0.143084	-0.034444	0.007521	-0.107764	0.107764	0.032080	-0.032080	-0.198893	0.198893	-0.201238	0.201238	-0.070440	0.070440	-0.073385	-0.088739	-0.088811
BMI	0.027355	1.00000	0.089124	0.034011	-0.064533	-0.028788	0.028788	0.039295	-0.039295	-0.021597	0.021597	-0.154506	0.154506	-0.044332	0.044332	-0.126873	-0.034811	-0.034537
PhysicalHealth	0.143084	0.089124	1.00000	0.276993	-0.079914	-0.002819	0.002819	0.012833	-0.012833	-0.116252	0.116252	-0.355052	0.355052	0.062744	-0.062744	-0.024341	-0.027438	-0.023344
MentalHealth	-0.034444	0.034011	0.276993	1.00000	-0.132038	-0.000838	0.000838	-0.056525	0.056525	-0.026438	0.026438	-0.107651	0.107651	0.137540	-0.137540	0.019557	0.084244	0.088833
SleepTime	0.007521	-0.064533	-0.079914	-0.132038	1.00000	0.033726	-0.033726	0.003927	-0.003927	-0.006237	0.006237	0.032884	-0.032884	0.017853	-0.017853	0.018542	-0.019822	-0.041772
Smoking_No	-0.107764	-0.028788	-0.002819	-0.000838	0.033726	1.00000	-1.00000	0.111768	-0.111768	0.081226	-0.081226	-0.120074	0.120074	0.083052	-0.083052	-0.138397	0.052149	0.015226
Smoking_Yes	0.107764	0.028788	0.002819	0.000838	-0.033726	-1.00000	1.00000	-0.111768	0.111768	-0.081226	0.081226	-0.120074	-0.120074	-0.083052	0.083052	-0.138397	-0.052149	-0.015226
AlcoholDrinking_No	0.032080	0.039295	0.012833	-0.056525	0.003927	0.111768	-0.111768	1.00000	-1.00000	0.019838	-0.019838	-0.033228	0.033228	0.054200	-0.054200	-0.004334	-0.023089	-0.019302
AlcoholDrinking_Yes	-0.032080	-0.039295	-0.012833	0.056525	-0.003927	-0.111768	0.111768	-1.00000	1.00000	0.019838	-0.019838	-0.033228	-0.033228	-0.054200	0.054200	0.004334	0.023089	0.019302
Stroke_No	-0.198893	-0.021597	-0.116252	-0.026438	-0.002817	0.081226	-0.081226	-0.019838	0.019838	1.00000	-1.00000	0.174143	-0.174143	-0.003091	0.003091	0.048352	0.048208	0.046279
Stroke_Yes	0.198893	0.021597	0.116252	0.026438	0.002817	-0.081226	0.081226	0.019838	-0.019838	-1.00000	1.00000	-0.174143	0.174143	0.003091	-0.003091	-0.048352	-0.048208	-0.046279
DiffWalking_No	-0.201238	-0.154506	-0.355052	-0.107651	0.032884	0.032884	-0.120074	-0.120074	0.033228	0.033228	0.174143	-0.174143	1.00000	-1.00000	-0.000000	-0.000000	-0.000000	-0.000000
DiffWalking_Yes	0.201238	0.154506	0.355052	0.107651	-0.032884	-0.032884	0.120074	0.120074	-0.033228	-0.033228	-0.174143	0.174143	-1.00000	1.00000	0.000000	-0.000000	-0.000000	-0.000000
Sex_Female	-0.070440	-0.044332	0.062744	0.137540	0.017853	0.083052	-0.083052	0.004200	-0.004200	-0.003091	0.003091	-0.000000	0.000000	1.00000	-1.00000	0.040865	-0.026288	-0.018028
Sex_Male	0.070440	0.044332	-0.062744	-0.137540	-0.017853	-0.083052	0.083052	-0.004200	0.004200	0.003091	-0.003091	0.000000	-0.000000	-1.00000	1.00000	-0.040865	0.026288	0.018028
AgeCategory_18-24	-0.073385	-0.126873	-0.024341	0.119557	0.018542	0.084244	-0.119557	-0.119557	-0.004334	0.004334	-0.048352	0.048352	-0.047173	-0.047173	1.00000	-0.062331	-0.062331	-0.062331
AgeCategory_25-29	-0.088739	-0.034811	-0.027438	0.084244	-0.019822	0.052149	-0.052149	-0.032089	0.032089	0.048352	-0.048352	0.003030	-0.003030	-0.032639	0.032639	-0.062331	1.00000	-0.059056
AgeCategory_30-34	-0.088811	-0.034537	-0.023344	0.088833	-0.041772	0.015226	-0.015226	-0.015902	0.015902	0.048352	-0.048352	0.003030	-0.003030	-0.032639	0.032639	-0.062331	-0.059056	1.00000
AgeCategory_35-39	-0.086651	0.048033	-0.023381	0.056060	-0.048279	-0.004200	0.004200	-0.021545	0.021545	0.008863	-0.008863	0.046903	-0.046903	-0.008302	0.008302	-0.069586	-0.062006	-0.063498
AgeCategory_40-44	-0.091188	0.032081	-0.018086	0.036191	-0.044768	-0.010880	0.010880	-0.018816	0.018816	0.001818	-0.001818	0.056260	-0.056260	-0.001790	0.001790	-0.070408	-0.062738	-0.068178
AgeCategory_45-49	-0.049733	0.047434	-0.008614	0.026357	-0.040508	0.006637	-0.006637	-0.009937	0.009937	0.025682	-0.025682	0.046195	-0.046195	0.000091	-0.000091	-0.071806	-0.063864	-0.067482
AgeCategory_50-54	-0.032648	0.032373	0.008397	0.013537	-0.039921	0.011867	-0.011867	-0.010388	0.010388	0.011968	-0.011967	0.018367	-0.018367	0.011943	-0.011943	-0.003080	-0.068473	-0.073284
AgeCategory_55-59	-0.013276	0.043527	0.019442	0.000321	-0.033470	-0.008701	0.008701	-0.008189	0.008189	0.001344	-0.001344	-0.016113	0.016113	-0.022744	0.022744	-0.083555	-0.073790	-0.079943
AgeCategory_60-64	0.016152	0.033760	0.027004	-0.028692	-0.010893	-0.011862	0.011862	-0.001621	0.001621	-0.011316	0.011316	-0.038116	0.038116	-0.021220	0.021220	-0.091115	-0.081190	-0.085641
AgeCategory_65-69	0.042628	0.027622	0.011022	-0.063254	0.011382	0.033387	-0.033387	0.008026	-0.008026	-0.022110	0.022110	-0.074311	0.074311	0.002296	-0.002296	-0.091816	-0.081815	-0.088300
AgeCategory_70-74	0.062578	0.004235	0.011988	-0.076074	0.054875	-0.045200	0.045200	0.021730	-0.021730	-0.039817	0.039817	-0.057190	0.057190	0.016668	-0.016668	-0.087100	-0.077613	-0.081688
AgeCategory_75-79	0.088890	-0.021008	0.013039	-0.078903	0.083770	-0.048240	0.048240	0.027891	-0.027891	-0.039840	0.039840	-0.074948	0.074948	0.022299	-0.022299	-0.071258	-0.065486	-0.068977
AgeCategory_80 or older	0.143084	-0.091123	0.018317	-0.126253	0.084476	-0.011969	0.011969	0.043228	-0.043228	-0.008039	0.008039	-0.156933	0.156933	0.046396	-0.046396	-0.073898	-0.067631	-0.073139
Race_American Indian/Alaskan Native	0.028547	0.028768	0.019108	0.141166	-0.088032	0.023967	-0.023967	0.004240	-0.004240	-0.014037	0.014037	-0.021203	-0.021203	0.003404	-0.003404	0.003728	0.003697	0.003838

Observations

- The categorical features most highly correlated with 'HeartDisease' are 'Stroke_Yes', 'DiffWalking_Yes', 'AgeCategory_80 or older', 'Diabetic', 'GenHealth_Fair', 'GenHealth_Poor', 'KidneyDisease_Yes'
- There is not significant collinearity between the features (Rule of thumb: If the correlation > 0.8 then severe multicollinearity may be present)

- From here, histograms were created for only the most correlated features to 'HeartDisease' ('Stroke', 'DiffWalking', 'AgeCategory', 'Diabetic', 'GenHealth', 'KidneyDisease') to show their distributions



Observations

- Most people in the dataset do not develop heart disease
- Most people in the dataset do not develop stroke
- Most people in the dataset do not have difficulty walking
- The dataset age categories span from the lowest at 16955 (ages 25-29) and highest at 34151 (ages 65-59)
- Most people in the dataset do not have diabetes, pre-diabetes or gestational diabetes
- Most people in the dataset have Excellent, Good or Fair general health
- Most people in the dataset do not have diabetes

- Additionally, stacked bar charts were created to compare heart disease and no heart disease by the most correlated features found in the Spearman's Correlation Matrix



Observations

- Most people in the dataset do not have heart disease, stroke, difficulty walking, kidney disease, or diabetes
- Heart disease rates are relatively the same in both men and women (however women have slightly less occurrences of heart disease)
- General Health and Age appear to be a significant factor in predicting heart disease

- **Do I need to adjust the data and/or driving questions?**
 - Although the dataset is fairly simple and not overly complex with no missing data, the data does not need to be replaced or the driving questions changed.
 - The only adjustments to the data were the target variable ('HeartDisease') was converted into a binary option (No is 0 and Yes is 1), and the categorical features were

transformed into dummy variables. This was all done to support the Correlation Matrices and the Logistic Regression Modeling.

- **Do I need to adjust my model/evaluation choices?**
 - As the problem statement looks to understand what features are most predictive of heart disease (a binary field), a classification algorithm would need to be used as it predicts a class/category (heart disease detection). Therefore, a logistic regression model will be used on the dataset to determine which features are most related to or most predictive of heart disease, where the HeartDisease feature is the target.
 - The model choice does not need to be adjusted as this problem statement lends itself to a categorical decision which is supported by a logistic regression model.
- **Are my original expectations still reasonable?**
 - Yes, after analyzing and visualizing the data, my original expectations that the Heart Disease dataset will be able to help us understand which features are most predictive of heart disease is still valid.
 - There is no missing data, all categorical data has been transformed into dummy variables, and the data is ready to be prepared for the logistic regression model.

References

1. CDC. (2022, February 7). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved March 21, 2022, from <https://www.cdc.gov/heartdisease/facts.htm>
2. Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. Wiley.

3. El Khouli, R. H., Macura, K. J., Barker, P. B., Habba, M. R., Jacobs, M. A., & Bluemke, D. A. (2009, November). Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. *Journal of magnetic resonance imaging : JMRI*. Retrieved February 20, 2022, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2935260/>
4. Grover, K. (2020, June 23). *Advantages and disadvantages of logistic regression*. OpenGenus IQ: Computing Expertise & Legacy. Retrieved March 21, 2022, from <https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/>