Data Analytics Capstone Topic Approval Form

Student Name: Ashley Munguia

Student ID: 011276535

Capstone Project Name: Regression Analysis of Heart Attack Risk

Project Topic: Exploring features relevant to heart health and lifestyle choices to determine if any of the features, including indicators such as cholesterol and diabetes, have a higher impact on the presence or absence of a heart attack risk than the other features through logistic regression analysis.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: To what extent does cholesterol and diabetes affect heart attack risk?

Hypothesis:

- **Null Hypothesis-** Cholesterol and diabetes do not statistically significantly affect the risk of a heart attack.
- **Alternate Hypothesis-** Cholesterol and diabetes statistically significantly affect the risk of a heart attack.

Context: Data analysis through logistic regression will analyze the provided dataset of patient records to determine if there is a correlation between risk of heart attack and the features shown below. Many factors increase the risk of a heart attack, including lifestyle choices, genetic factors, and age. Perhaps the knowledge of what specific features to be cautious of after the analysis is conducted will be beneficial for the patient to know to lower the risk of a heart attack.

Data: The provided dataset has 8763 global patient records consisting of features related to heart health, lifestyle choices, lifestyle factors, and medical aspects to conduct a predictive analysis of the binary categorical feature, if the patient is at risk of a heart attack or not.

I will be using the "Heart Attack Risk Prediction" csv file from Kaggle. The link to the Kaggle dataset being used for analysis: https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset/data

This dataset is a made-up creation generated using ChatGPT that mimics a realistic experience. It helps beginners and data enthusiasts learn from a dataset that resembles real-world situations. The goal is to support learning and experimentation in data analysis.

Data Gathering: The data will be collected by downloading the "Heart Attack Risk Prediction" csv file from Kaggle.

Data Analytics Tools and Techniques: I will be using a logistic regression model to gain more insight to determine which features from the dataset correlate to a risk of a heart risk. I will be working in a Jupyter Notebook and using Python to perform this data analysis. I will be using the following libraries and packages for my analysis:

- pandas- to load datasets
- NumPy- to work with arrays
- Sci-kit Learn- for machine learning and to transform our data
- SciPy- for mathematical problems like checking for multicollinearity
- Matplotlib- for basic plotting generally consisting of bars, lines, pies, scatter plots, and graphs
- Seaborn- for a variety of visualization patterns

Justification of Tools/Techniques: Logistic regression is the appropriate technique to use to analyze the research question because the target variable, heart attack risk, is a binary categorical dependent variable. The multiple explanatory variables can be continuous and categorical. Performing logistic regression will determine if the explanatory variables have a positive or negative impact on the chosen target variable. This predictive

model will give an indication of what independent variables directly affect heart attack risk as we add or remove them.

Project Outcomes: The outcome of this data analysis is to perform a predictive analysis and develop a logistic regression model to gain insight on which features from the dataset have the most impact on the risk of having a heart attack. The deliverables will include a heatmap to check for multicollinearity, variance inflation factors of all the independent variables, the p-values to perform backward stepwise elimination, and the regression equation for the reduced model. The initial logistic model and the reduced model will be compared based on the following model evaluation metrics: log-likelihood, number of independent variables, and LLR p-value. The accuracy of the reduced model will be used to evaluate the model's performance.

Projected Project End Date: On or before 1/31/2025

Sources:

Banerjee, S. (2024, May 11). Heart attack risk prediction dataset. Kaggle. https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset/data

Course Instructor Signature/Date:

☒ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: 1/21/2025

Reviewed by:

Comments: Click here to enter text.