Analytics Finalized Plan

Synopsis: ***This document provides a high-level walkthrough of the activities required to guide the completion of the analysis.***

Project	*A predictive model for identifying individuals at high risk of heart disease.*
Requestor	*Centennial College*
Date of Request	*July 15, 2024*
Target Quarter for Delivery	*2nd Quarter of 2024 (August 2024)*
Epic Link(s)	*Kaggle. (2022). Heart Disease Health Indicators Dataset: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset* *American Heart Association. (2024). Understand Your Risks to Prevent a Heart Attack: https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack#:~:text=An%20inactive%20lifestyle%20is%20a,blood%20pressure%20in%20some%20people.* *National Library of Medicine. (Circa). Risk Factors for Coronary Artery Disease: Historical Perspectives:* *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5686931/* *World Health Organization. (2024). Cardiovascular diseases: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1*
Business Impact	***The project aims to create a predictive model for identifying individuals at high risk of heart disease.*** *By analyzing the key health indicators to build a model capable of predicting heart disease risk.*
Prepared by	***Jason S. Yap - 301293413***

1.0 Business Opportunity Brief

Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment.

Business Problem

Knowledge is power. By understanding the risks of heart disease, individuals can take proactive steps to improve their health. Analyzing how various indicators contribute to heart disease can help identify key factors that could reduce the risk. Heart disease remains a leading global cause of mortality, claiming an estimated over 17 million lives

annually (World Health Organization, 2024). Identifying at-risk individuals is crucial for prevention and early intervention.

Business Solution

Inspired by the Behavioral Risk Factor Surveillance System (BRFSS) in the United States, this project aims to create a similar but more localized and digitized survey system. By utilizing simple questionnaires to identify individuals at risk of heart disease, we can mimic the BRFSS model. Initially, the heart disease indicator dataset from BRFSS will serve as a foundation for predictive modeling. Ultimately, this system can be used by communities, non-profit organizations, and healthcare institutions.

Specific Tasks

1. Data Cleaning and Preparation

- Handle missing values through imputation as needed.
- Address inconsistencies or outliers using capping and flooring if necessary.
- Convert categorical variables into a numerical format using one-hot or label encoding.
- Normalize or standardize numerical features to ensure consistent scale.

2. Exploratory Data Analysis

- Calculate descriptive statistics (mean, median, standard deviation) for each lifestyle factor.
- Visualize feature distributions and relationships with heart disease risk using skewness, histograms, and box plots.

3. Correlation and Feature Selection

- Conduct correlation analysis to identify potential relationships between features.
- Employ a Random Forest classifier to identify significant features for modeling.

4. Predictive Modeling

- Build multiple models (Decision Trees, Logistic Regression, Neural Networks) using selected features.
- Evaluate model performance using ROC AUC, accuracy, recall, precision, and F1 score.
- Apply cross-validation to assess model robustness and generalizability.

5. Model Selection and Interpretation

- Select the best-performing model based on evaluation metrics.
- Interpret model results to identify key indicators of heart disease risk.

6. Recommendations and Insights

- Utilize insights from the best indicators to inform heart disease risk identification and prevention.
- Create visualizations to effectively communicate findings.
- Prepare a comprehensive report summarizing the analysis, results, and recommendations.

1.1 Supporting Insights

Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally?

According to the American Heart Association
The common traditional risk factors for heart attack are Smoking, High blood pressure, High cholesterol, Diabetes, and being Overweight or obese **which are then complemented by** a Family history of early atherosclerotic cardiovascular disease (less than 55 years old for men while less than 65 years old for women), High cholesterol (ranging from 160-189 mg of LDL-C/ 190-219 mg/dl of non-HDL-C), metabolic syndrome, chronic kidney disease, Chronic inflammatory conditions, History of preeclampsia or early menopause, High-risk ethnicity (esp. from South Asian ancestry), and Higher than normal triglycerides (over 175 mg/dl), ankle-brachial index (ABI) and other lab tests.

National Library of Medicine: According to a **Framingham study**

Cholesterol Level: In 1953, an association between cholesterol levels and CHD mortality was reported in various populations. Animal and clinical observation have suggested such a relationship. This association was confirmed by epidemiological studies showing a strong relation between serum total cholesterol and cardiovascular risk.

Cigarette smoking doubles the risk of morbidity and mortality from ischemic heart disease compared with a lifetime non-smoker, and the risk is related to the duration and amount of smoking. Additionally, Smokers below 50 years old is 10 times more likely to develop CHD than nonsmokers of the same age. Passive smoking also increases the risk of CHD.

Approximately 68% of people over 65 years old with diabetes die due to heart-related disease. Adults with diabetes are two to four times more likely to die from heart disease than adults without diabetes. Thus, The American Heart Association considers diabetes to be one of the seven major controllable risk factors for CVD.

The association between **obesity** and CHD was first noticed over 50 years ago. Which is an independent risk factor for all-cause mortality. It is a metabolic disorder associated with comorbidities such as CHD, type 2 diabetes, hypertension, and sleep apnea. A recent study reported that higher body mass index (BMI) during childhood is associated with an increased risk of CHD in adulthood.

1.2 Project Gains

Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI's.

• **Early Intervention and Prevention:** Early detection of heart disease risk indicators can significantly reduce the likelihood of devastating consequences such as heart attacks and strokes.

• **Personalized Healthcare:** By effectively managing and mitigating risk factors, individuals can significantly enhance their quality of life by preventing or delaying the onset of debilitating advanced heart disease.

• **Economic Benefits:** A healthy population is a productive population. By reducing the prevalence of heart disease, we can minimize productivity losses due to illness and disability.

• **Empowerment and Awareness:** Empowering individuals with knowledge about heart disease risks is essential for promoting proactive health management. Increased health literacy encourages individuals to take control of their well-being.

Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resources, time, priority and data availability.

2.0 Analytics Objective

List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.

Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed.

Key Questions

- Which health indicators are the most significant predictors of heart disease?
- How do lifestyle choices like physical activity, diet, and smoking impact the risk of heart disease?
- Can demographic factors such as age, sex, education, and income level be used to accurately predict heart disease risk?
- What are the implications of these findings for public health policies and individual behavior changes?

Assumptions

- The data is accurate, assuming that respondents accurately report their health behaviors, conditions, and demographic details. Misreporting or recall bias could affect the validity of the data.
- The sample is represented & consistent with the population: The sample is representative and consistent with the general population.
- Independence of Observations: One's data does not influence another's within the dataset.
- Balance of the Target Variable: Using down sampling techniques.
- Association vs. Causation: correlation-related (e.g., if ice cream sales and drowning incident increases, it means both tend to increase).

2.1 Other related questions and Assumptions:

List any assumptions that may affect the analysis

- Confounding Factor: A variable that influences both the dependent variable and independent variable, leading to a spurious association. In this context, confounding factors are those not captured in the dataset due to the time constraints of telephone surveys.

2.2 Success measures/metrics

What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.

Key Performance Indicators (KPIs):
The model with the highest percentage or score across all specified metrics below will be considered the best model.

1. ROC-AUC: Best ability to distinguish between classes
2. Accuracy: Measures the correct proportion of predictions among all instances
3. Recall: Very good at identifying true positive cases.
4. Precision: Measuring the proportion of correct predictions among all instances
5. F1 Score: Balances precision and recall, providing a comprehensive evaluation metric

2.3 Methodology and Approach

Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.

Please refer to the specific tasks under the Business Opportunity Brief Section

3.0 Population, Variable Selection, considerations

Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today.

Audience/population selection: General Community

Category	Data Variables	Type	Context
Output	HeartDiseaseorAttack	Binary	Experienced HD?
Health Details	HighBP	Binary	HBP confirmed by medical professional?
Health Details	HighChol	Binary	HChol confirmed by medical professional?
Health Details	CholCheck	Binary	Chol check-up w/in 5yrs
Health Details	BMI	Category/ Int (18.5, 24.9, 29.9, 34.9, 40)	Underweight to Obese 3
Health Details	Smoker	Binary	Smoked over 100 cigars?
Health Details	Stroke	Binary	Experienced Stroke?
Health Details	Diabetes	Binary	Has diabetes?
Lifestyle	PhysActivity	Binary	Active w/in 30 days? Except work
Lifestyle	Fruits	Binary	Eat 1 or more/ day
Lifestyle	Veggies	Binary	Eat 1 or more/ day
Lifestyle	HvyAlcoholConsump	Binary	Over 14 drink/ week? (7 women)
Access to Healthcare	AnyHealthcare	Binary	Healthcare?
Access to Healthcare	NoDocbcCost	Binary	Avoid medical attention due to cost?
Health Status	GenHlth	Category/ Int (1-5)	1: Excellent, 5: Poor
Health Status	MentHlth	Binary	Days of MentHlth/ Month
Health Status	PhysHlth	Binary	Days of PhysHlth/ Month
Health Status	DiffWalk	Binary	Difficulty walking
Personal Details	Sex	Binary	Male/ Female
Personal Details	Age	Category/ Int (1-5)	1. 18-24, 2. 25-29…13. over 80
Socioeconomic	Education	1: Only kindergarten, 2-4: Grade - Highschool, 5 : Diploma/ Vocational, 6: Bachelor's Degree	Education levels are varied, with a concentration in higher education
Socioeconomic	Income	1: Less than $10,0000, 2: $10,0000- $15,00008… 8 : over $75,000, With $5,000 gap.	Income levels are spread across several categories, with distinct peaks

Data Sources: Kaggle (*https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset*).

4.0 Dependencies and Risks

Identification of key factors that may influence the outcome of the project and likelihood of it happening:

Dependency	Details	Influence	Likelihood
Data Quality and Availability	Access to a comprehensive and high-quality dataset that accurately represents the population of interest.	Poor quality or incomplete data can lead to inaccurate results and unreliable conclusions.	Moderate, since data has no missing values
Technical Infrastructure:	Adequate computational resources and technical infrastructure to perform data processing, analysis, and modeling.	Insufficient technical resources can slow down the analysis process and affect the study's timeline.	Low to moderate, since we are only using approximately eight thousand records.
Collaboration with Advisors	Collaboration with data advisor to validate findings and ensure the relevance of recommendations.	Lack of expert input can reduce the credibility and applicability of the study's results.	Low to moderate, depending on the availability and willingness of experts to collaborate.
Risk	**Details**	**Impact**	**Mitigation**
Data Bias and Limitations	May have inherent biases or limitations that affect the generalizability of the findings.	High, as biased data can lead to incorrect conclusions and ineffective recommendations.	Perform data exploration and validation and use techniques to identify and address biases.
Model Overfitting or Underfitting:	Predictive models may overfit to the training data or underfit, failing to capture important patterns.	High, as poor model performance can result in inaccurate risk predictions and recommendations.	Use cross-validation, regularization techniques, and proper model selection to ensure robust performance.

5.0 Deliverable Timelines

List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process.

Items	Week	Action Items	Due Date	Status
1	**2**	**Analysis Plan & Data Finalization**	**15-Jul**	**Completed**
2	3	Analysis Plan & Data Assessment/ Revision	18-Jul	
3	3	Data Exploration Review	19-Jul	
4	3-4	Data Exploration Revision	21-Jul	
5	**3-4**	**Data Exploration Final**	**22-Jul**	**Completed**
6	4	Analysis and Interpretation Review	23-Jul	
7	2-4	Correlation and Regression Analysis Final	24-Jul	
8	2-4	Interpretation/ Recommendation/ Insights Final	24-Jul	
9	**5**	**Peer Review Week 4**	**29-Jul**	**Completed**
10	**6**	**Modeling**	**5-Aug**	**Completed**
11	**7**	**Governance**	**15-Aug**	**Completed**
12	**7**	**Documentation**	**15-Aug**	**Completed**
13	**7**	**Peer Review Week 6**	**15-Aug**	**Completed**
14	**7**	**Presentation**	**15-Aug**	**Completed**
15	**7**	**Portfolio**	**19-Aug**	**Completed**
16	**8**	**Finalized all documents**	**19-Aug**	**Completed**

Note: Highlighted rows are class due dates. While unhighlighted are internal timelines, which are subject to change.