# Analytics Startup Plan

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

|  |  |
| --- | --- |
| **Project** | *Key Indicators to prevent Heart Disease* |
| **Requestor** | *Centennial College* |
| **Date of Request** | *July 11, 2022* |
| **Target Quarter for Delivery** | *August 18, 2022* |
| **Epic Link(s)** | *Not Applicable* |
| **Business Impact** | *Improving overall healthcare and decreasing the number of deaths caused by heart disease* |

## 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. |

One of the top causes of death in the US for persons of most races is heart disease.

The analysis will be carried out to lower the number of heart-related fatalities in America.

In the field of medicine, it is crucial to identify and combat the causes that have the biggest influence on heart disease. In turn, advances in computation enable the use of machine learning techniques to identify "patterns" in data that can foretell a patient's state.

**The specific ask:**

*Looking for the key factors causing Heart Disease: risk factors, race etc Clearly articulate the specific task you will be conducting to help achieve the opportunity*

To analyze the likelihood that someone might experience specific symptoms of heart disease.

## 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? |

The Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to collect information on Americans' health state, is the primary source of the dataset, which comes from the CDC. As stated by the CDC: "BRFSS, which started out with 15 states in 1984, currently gathers data from all 50 states, the District of Columbia, and three U.S. territories. The BRFSS is the largest continually running health survey system in the world, conducting over 400,000 adult interviews each year ".

Heart disease is one of the major causes of death in the US for people of most races, according to the CDC. High blood pressure, high cholesterol, and smoking are the three main risk factors for heart disease that at least half of all Americans (47%) have. The presence of diabetes, obesity (high BMI), insufficient physical exercise, and excessive alcohol consumption are additional important markers. In the field of medicine, it is crucial to identify and combat the causes that have the biggest influence on heart disease. In turn, advances in computation enable the use of machine learning techniques to identify "patterns" in data that can foretell a patient's state.

## 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.* |

According to estimates, 17.9 million people die from heart disease each year, accounting for 32% of all fatalities worldwide. With the help of data, we can investigate and identify the factors that most influence a person's probability of developing heart disease.

## *Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, 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. |

## Which are the top 3 factors causing heart disease?

1. Race considering a factor risk for heart disease.
2. Age is likely a factor to have a heart disease.
3. Considering women or men are more likely to have a heart disease and justify.
4. Percentage of drinkers and smokers having heart disease.

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

Most of the sleep times are likely data errors (misreported or mis transcribed most likely). For example, sleep time within range of 18-25 hours.

## 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.* |
|  | 1. Hassle free checkup of person’s health and heart disease. 2. Convenient way to get suggestions on improving overall health to avoid heart problems. 3. People realize the importance of health checkups and improve accordingly. |
|  |  |

## 2.3 Methodology and Approach

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

**Type of Analysis: EDA, Visualization, Preprocessing, Modelling - Logistic regression, Decision tree, normalization, ML Models and comparison, correlation**

**I will begin with data cleaning and then creating basic visualizations to understand the data pattern. Then I will go for correlation to find out the top columns that are mostly leading to heart disease. After data visualization, I will begin with building models and will use other techniques to verify the results.**

**Methodology: I will begin with performing EDA to remove clean the data and remove outliers so that the data is ready to use for my actual analysis. The target variable is heart disease. I will start with data visualization and build the models and doing comparison among them. I will be showing the visualization of columns separately with the target variable and how likely it will be impacted with the variable.**

**Output: The output of this analysis will be predicting the likelihood of having a heart disease and recommendation on how to improve the current health condition to avoid heart disease.**

## 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: US residents**

**Observation window: 2020**

**Inclusions: 18 variables**

**Exclusions: NA, already original dataset was reduced from 300 variables.**

**Data Sources:** <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

**Audience Level:** NA

**Variable Selection:**

**Derived Variables:**

**Assumptions and data limitations:**

## 4.0 Dependencies and Risks

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
|  |  |  |  |

## 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. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | EDA – Data cleaning, exploration, and Analysis | *Data cleaning and analyzing* |  | *2* | *21-07-2022* |
| 2. | *Basic Visualization* | *Graphs and plots* |  | 2 | *23-07-2022* |
| 3. | *Modelling – Machine learning* | Building models – logistic regression, decision tree etc. |  | 4 | *30-07-2022* |
| 4. | Model comparison | Comparison and analysis |  | 1 | *31-07-2022* |
| 5. |  |  |  |  |  |
| 6. |  |  |  |  |  |
| 7. |  |  |  |  |  |
| 8. |  |  |  |  |  |
| 9. |  |  |  |  |  |
| 10. |  |  |  |  |  |
| 11. |  |  |  |  |  |