Group Three Pep Two PRD

Scrum Master: Finn Mikkola, Max Ross, Jason Fearnall

Introduction

We will be examining Cardiovascular disease data provided by Kaggle.com and setting up a data pipeline using AWS services to examine the data, store the data, and provide an analytical report. We will be loading the data into an S3 storage unit, creating an RDS SQL database, cleaning the data, and providing insights on the data.

Features

* S3 Bucket Data Storage
* RDS Instance with Public VPN
* Lambda Loading and Data Cleaning Function
* Statistical Analysis
* Webpage using an AWS EC2
* Machine Learning Classification Model

Deliverables

* Statistical analysis to show insights on data
* Webpage to present findings
* Machine Learning Classification Model

Technical Requirements

* AWS Services
* Python, SQL, HTML, scikit-learn
* User Interface (UI)

For our Extra task, if there is time, we will be creating a website hosted on an AWS EC2 displaying our analytical findings.

Data Management

* S3, SQL database in RDS.

Personas

* For individuals seeking information about health conditions that may indicate cardiovascular disease.

Success Metrics

* Webpage is live displaying the statistical analysis
* Machine learning model has a high score for our chosen evaluation metric
* All Team members can access S3 bucket storage
* All Team members can access RDS instance
* Lambda function successfully places cleaned data into RDS
* Data analysis has 5 meaningful insights

Timeline & Release Planning

* One week for complete product to be available for demonstration.