**Project Proposal: Heart Failure Prediction**

**1) Team Members:**

* Azam Jah Al Ahmed
* Tejashri Chavan

**2) Problem-Opportunity Statement:**

Cardiovascular-related diseases are the primary cause of death globally and heart failure (HF) is a frequent ultimate manifestation of many heart pathologies. Timely interventions through early diagnosis and control of EHF (**End-Stage Heart Failure**) condition would greatly improve results. The increasing popularity of machine learning algorithms allows clinicians to develop predictive models that can be used to predict whether or not an individual will develop heart failure using many medical and lifestyle factors. Objective: To develop a predictive model to help healthcare providers determine high-risk (HR) patients who need tailored preventive care.

**2. Business Objective:**

The main business purpose of this analysis Is to create an accurate prediction machine learning model identifying if a patient has the risk of Heart Failure due to their sustainable features predicated on medical and lifestyle necessities. Healthcare providers can use the model to decide which patients need urgent attention and how to best allocate medical resources. Furthermore, it has the potential to develop tailored treatments for those at risk.

**3. Predictive Modelling Task:**

The task at hand is to build a model that can predict if a patient will go on to develop heart failure (Yes/No). The target variable (will the patient suffer from heart failure?) is binary, so we have a classification problem.

**4. Dataset Details:**

**Dataset Name**: Heart Failure Prediction Dataset

**Source**: The dataset has been obtained from Kaggle and is publicly available.

**URL:** [Heart Failure Prediction Dataset (kaggle.com)](https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data)

**Number of Observations:** The dataset contains approximately 918 observations.

**5. Data Usage:**

**Training Data**: Approximately 643 (70% of the total observations) will be used for training the model.

**Validation Data:** Approximately 275 (30% of the total observations) will be used for validating the model’s performance.

**6. Dependent Variable (Target):**

**Heart Disease:** This is a binary variable that indicates whether the patient has been diagnosed with heart disease (1 = Yes, 0 = No). We aim to predict this outcome.

**7. Independent Variables (Features):**

The dataset contains various medical and lifestyle variables that can impact heart disease risk. The following are the independent variables in the dataset:

**Age:** Age of the patient.

**Sex:** Gender of the patient (1 = Male, 0 = Female).

**ChestPainType:** Type of chest pain (4 categories: **TA:** Typical Angina, **ATA:** Atypical Angina, **NAP:** Non-Anginal Pain, **ASY:** Asymptomatic)

**RestingBP:** Resting blood pressure (in mm Hg).

**Cholesterol:** Serum cholesterol level (in mg/dl).

**FastingBS:** Fasting blood sugar (> 120 mg/dl) (1 = True, 0 = False).

**RestingECG:** Resting electrocardiogram results (Normal, ST, LVH).

**MaxHR:** Maximum heart rate achieved during exercise.

**ExerciseAngina:** Exercise-induced angina (1 = Yes, 0 = No).

**Oldpeak:** ST depression induced by exercise relative to rest.

**ST\_Slope:** The slope of the peak exercise ST segment (Up, Flat, Down).

These variables, reflecting patient demographics, symptoms, and clinical test results, are expected to provide valuable insights for heart disease prediction.

**8. Data Mining Techniques:**

For this project, the following data mining techniques will be utilized:

**Decision Trees**

**Logistic Regression**

**Neural Networks**

**9. Data Mining Software:**

In addition to KNIME, the following could be used for data analysis:

**Python**

**Excel**

**PowerBI**