**STATEMENT OF WORK**

**PROJECT NAME:** CORONARY HEART DISEASE RISK CLASSIFICATION

**PROJECT ESTIMATION:** 45 DAYS

**PROJECT TEAM:**

|  |  |
| --- | --- |
| **STUDENT ID** | **TEAM MEMBER** |
| 100836815 | Lawrence Wanderi Mwangi |
| 100844617 | Shrutika Raut |
| 100846988 | Tejas Devani |
| 100835622 | Vineth Rajendran |

**BACKGROUND:**

According to WHO, nearly 12 million deaths occur worldwide, every year due to heart diseases. One of the biggest challenges facing integrated health systems and risk-bearing entities, is figuring out who is at high risk. Identifying the high-risk patients at early stage would aid in opting healthy lifestyle, strategies for greater care coordination, and reduce complications. However, the great diversity of case definitions, diversity among stakeholders, availability of source data, access to technology and analytical manpower, all complicate the refinement and use of a high-risk patient identification.

**PROJECT STATEMENT:**

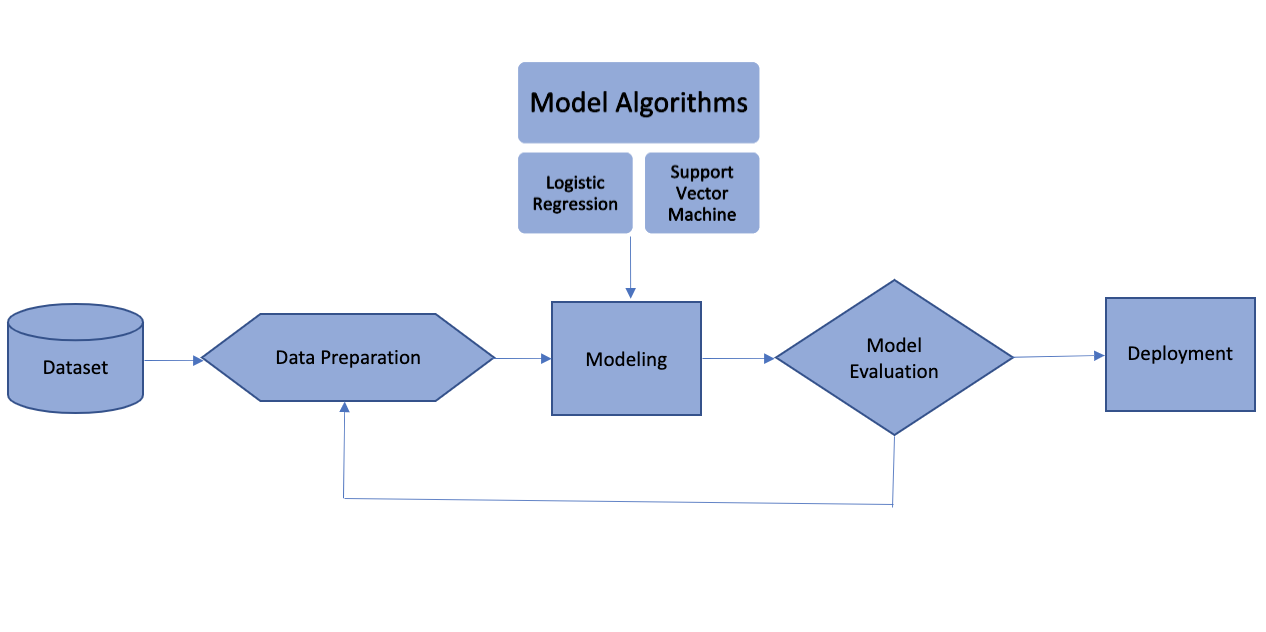
Predict if the patient has a risk of developing coronary heart disease over a 10-year period. The prediction is based on patient’s demographic, behavioural and medical parameters.

**PROJECT OBJECTIVES:**

The objective of this project is to develop a solution model that would predict overall risk of heart disease. The key algorithm to build the model would be Logistic Regression. The deliverable includes more accuracy assurance by integrating one more algorithm, Support Vector Machine.

* Study data and attributes & perform exploratory data analysis
* Evaluate and prototype models
* Train model with data
* Test model with data
* Perform quality assurance by validating accuracy
* Project deployment

**SOLUTION DESIGN:**



**REQUIREMENTS:**

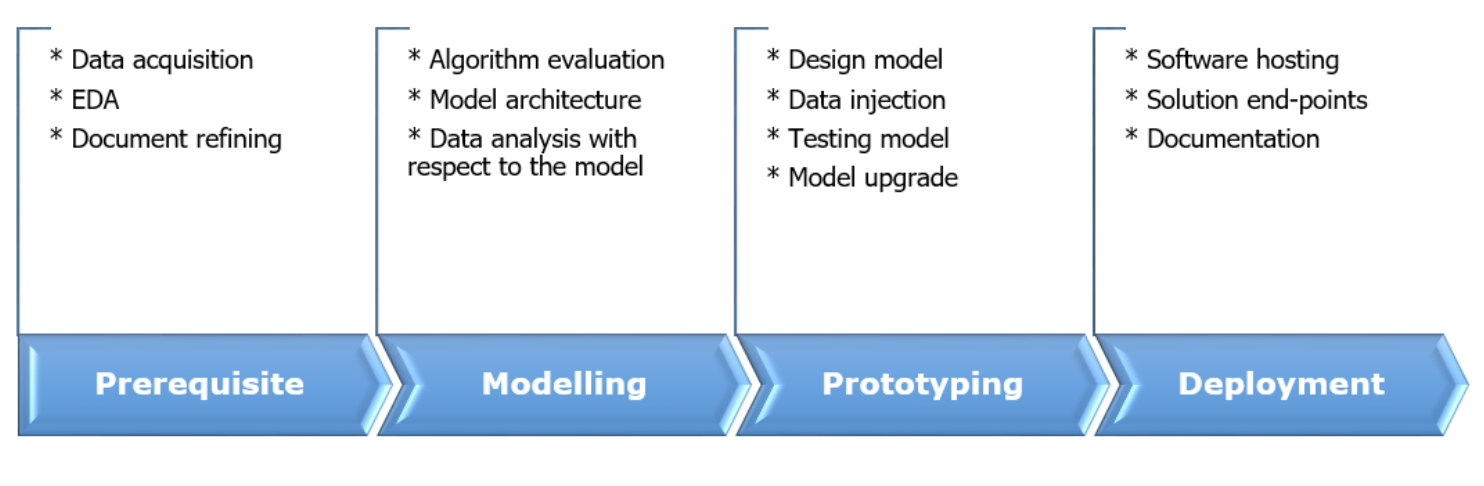
**Data Source:**

The dataset has been generated from a cardiovascular study on residents of the town of Framingham, Massachusetts. It includes over 4000 records with 15 independent attributes and 1 dependent attribute (Coronary Heart Disease Risk).

**Data Limitations:**

* The data has been derived from a unique demographic and is doubtful whether it will be accurate if generalized in other distinctly different populations.
* The dataset is not balanced for some variables such as in patients who have diabetes. Hence, model may be overly optimistic or pessimistic in regard to weighting such variables.

**PROJECT MILESTONES:**



**TEST PLAN & QUALITY ASSURANCE:**

The test process will involve comparing and contrast two methods of predicting the target variable. Along with **Logistic Regression** the supporting algorithm would be **SVM (Support Vector Machine).**

The dataset would be divided into 2 subsets i.e., training and test data in the ratio of **2:1.** That is, 67% of the data will be used for model training while the rest of it shall be used for model testing. Both algorithms will be used for the same train and test datasets, so as to have an apples-to-apples accuracy comparison between them.

Quality assurance will be done using T-Test hypothesis for relevancy of parameters.