STATEMENT OF WORK

|  |  |
| --- | --- |
| Topics | |
| Indexing | |
|  | |
| Project Heads | |
|  | |
| Background, Problem Statement, Objectives | |
|  | |
| Project Milestone | |
|  | |
| Test Plan | |
|  | |
| Technical Solution | |
|  | Solution Design – Model Architecture |
|  | |
|  | Data Requirements |
|  | |
|  | Data Layout |
|  | |
|  | Exploratory Data Analysis - Data Pre-processing & Report |
|  | |
|  | Feature Engineering |
|  |  |
|  | Conclusion |
|  | |
|  | |
| Links & filenames | |
|  | |

|  |
| --- |
| PROJECT NAME |
| CORONARY HEART DISEASE RISK CLASSIFICATION |
| PROJECT ESTIMATION |
| 45 DAYS |
| PROJECT TEAM |
| *Lawrence Wanderi Mwangi*  Student ID – 100836815  *Shrutika Raut*  Student ID – 100844617  *Tejas Devani*  Student ID – 100846988  *Vineth Rajendran*  Student ID – 100835622 |

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

PROJECT MILESTONES

Timeline

Description automatically generated

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).**

All the models would be evaluated using k-Fold Cross Validation for accuracy and Confusion Matrix for performance. Configuration of k would be decided while model training for the best accuracy.

SOLUTION DESIGN

Model Algorithms

Model Evaluation

Training Data

Dataset

Deployment

Modelling

Data Preparation

Test Data

System Architecture

Data Preparation

EDA

Pre-processing

Feature Extraction & Selection

Data Preparation Process

Model Algorithm

LOGISTIC REGRESSION MODEL

Model

Training Data

TenYearCHD - NO

X2

Class 1

W3

W2

Class 0

W1

X1

X3

TenYearCHD - YES

Diagram

Description automatically generated

Model Algorithm

SVM

Model

Training Data

**Support Vectors**

TenYearCHD - NO

**Maximum Margin**

**Maximum Margin**

**Positive Hyperplane**

TenYearCHD - YES

**Negative Hyperplane**

Class 1

Class 0

Support Vector Machine

Diagram

Description automatically generated

TenYearCHD - YES

TenYearCHD - NO

Model

Training Data

Model Algorithm

Random Forest

Random Forest Algorithm

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 LAYOUT

* **Demographic:**

1. Sex – Nominal
2. Age – Continuous

* **Behavioural:**
  1. Current Smoker – Nominal
  2. Cigarette per Day – Continuous (Since one can even have half a cig per day)
* **Medical History:**

1. Blood Pressure Medication – Nominal
2. Prevalent Stroke – Nominal
3. Prevalent Hypertension – Nominal
4. Diabetes – Nominal

* **Medical Current:**
  1. Total Col – Continuous
  2. Systolic Blood Pressure – Continuous
  3. Diastolic Blood Pressure – Continuous
  4. Body Mass Index – Continuous
  5. Heart Rate – Continuous (treated as continuous since there are many possible values.
  6. Glucose – Continuous
* **Target Variable:**
  1. CHD (10-year risk of coronary heart disease) – Nominal

EXPLORATORY DATA ANALYSIS

Dataset undergone through the preliminary processes of,

* Data sampling
* Data profiling
* Dataset statistics
* Variable overview
* Attribute Interactions
* Correlations
* Handling Missing Values by Numerical Imputation
* Checking outliers for most correlated feature

Based on Data Pre-processing, EDA report generated is as follows:

## **Main Data Features**

From the visual statistics of the data generated, the attributes with the highest correlations to CHD are:

1. Age – positively co-related
2. Prevalent Hypertension status – Higher risk if hypertensive
3. Systole Blood Pressure – positively co-related

**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
  + Patients who have Diabetes
  + Patients with Blood Pressure Medications
  + Patients with Prevalent Stroke

Hence, model may be overly optimistic or pessimistic in regard to weighting such variables.

## **Target Class Imbalance**

There is a big imbalance of the target attribute (10-year CHD status). There are only **644** positive cases vs **3594** negative cases. Therefore, the test, validation and train data should be split with some thought and cross-validation of the final models should be done to get a correct accuracy figure.

Feature Engineering

Feature engineering involves modifying, deleting, and combining the data features. It prepares dataset to be compatible with the algorithm. It’s the first step towards improving the performance of machine learning algorithm.

Based on EDA report,

* Some features are not balanced
  + Patients who have Diabetes
  + Patients with Blood Pressure Medications
  + Patients with Prevalent Stroke

But they have minimal correlation with the target variable to be handled.

* The project’s raw dataset has around 645 missing values as per EDA report, approximately 15% of the complete data. This would have a potential impact on the accuracy of the model.

There are different feature engineering approaches to handle such limitations.

Considering the project’s dataset, best practice is to carry out Numerical Imputation. As the entire data is numerical with small value range. Also, except target variable none of the attributes depend on each other.

Numerical Imputation involves calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others.

Conclusions

## **Real Data vs Synthesized Data**

The size of the dataset (4000 records) should be adequate for the models to be employed i.e., Logistic regression and SVM. This is real-life data with an unknown probability distribution and should only be mixed with synthesized data when required to do so.

LINKS & FILENAMES

|  |  |  |
| --- | --- | --- |
| GitHub Repository | CoronaryHeartDiseaseRiskClassification | [*https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification*](https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification) |
| SOW 1 | CoronaryHeartDiseaseRiskClassification-SOW1.docx | [*https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/problem\_discovery*](https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/problem_discovery) |
| SOW2 | CoronaryHeartDiseaseRiskClassification-SOW2.docx | [*https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/data\_acquisition*](https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/data_acquisition) |
| Dataset | Dataset.csv | [*https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/data\_acquisition*](https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/data_acquisition) |
| EDA Report | CHD-EDA-Report.html | [*https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/data\_acquisition*](https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/data_acquisition) |
| After Feature-Engineering Report | CHD-PostProcessing-Report.html | [*https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/data\_acquisition*](https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/data_acquisition) |
| SOW3 | CoronaryHeartDiseaseRiskClassification-SOW3.docx | <https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/ml_modeling> |
| Data Assessment | CHD\_AI\_Algo.ipynb | <https://github.com/AIDI-1002-Group-Project-Durham/CoronaryHeartDiseaseRiskClassification/tree/main/ml_modeling> |