**Heart Disease Prediction Analysis**

**Machine Learning Model Proposal**

**Objective:**

Heart Disease is the leading cause of death for men and women in the United States. It is said that women are at a greater risk for Heart Disease than men. About 1 in every 5 female deaths is due to heart disease([www.cdc.gov](http://www.cdc.gov)). The diagnosis of heart disease can be challenging especially in women since they present atypical symptoms, as opposed to men. According to <https://pubmed.ncbi.nlm.nih.gov/26210899/>, “this, in association with a consistent underestimation of their risk for ischemic heart disease, leads to underdiagnosis and undertreatment in women.”

An automated system based on Machine Learning Models, that predicts the probability of heart disease in patients, based on medical and lifestyle parameters could be effective in predicting and preventing the disease and increased accuracy of diagnosis. This system would be beneficial to patients at risk, to hospitals and health systems in providing improved quality of care as well as to insurance companies.

**Data:**

The data extracted from the database is a join of 4 tables – Patient Demographics, Patient Habits, Pre-existing conditions and Parameters. The columns of the data frame created from the extracted data are :

* Female Patient ID – Unique ID of each patient
* Age
* Education
* Whether the patient is a smoker currently
* How many cigarettes per day
* Whether they take BP Meds
* Whether they have a stroke
* Whether there is prevalent Hypertension
* Whether the patient has Diabetes
* The Total Cholesterol level
* Systolic Blood Pressure reading
* Diastolic BP reading
* BMI
* Heart rate
* Glucose level
* Ten year CHD: Whether they are likely to develop heart disease in the next 10 years.

**Data Preprocessing:**

The following actions were performed as part of data preprocessing:

1. The raw data from our data source had data for both male and female patients. This was then cleaned as part of the ETL process to only retain data for female patients in order to address the needs of this project.
2. Null Values were identified and dropped.
3. The number of rows and columns in the data were determined
4. The datatypes were determined to ensure that they were valid and there were no “object” datatypes.
5. Data was checked for existence of duplicate rows.
6. The number of unique values under each column was determined.
7. The Categorical and Numeric features were identified as follows:

|  |  |
| --- | --- |
| Categorical Features | Numeric Features |
| Education | Age |
| Current Smoker | Cigarettes per Day |
| BP Medication | Total Cholestrol |
| Prevalent Stroke | Systolic BP |
| Prevalent Hypertension | Diastolic BP |
| Prevalent Diabetes | BMI |
|  | Heart rate |
|  | Glucose |

Categorical features predominantly indicated whether a condition was present or not (except for Education). Numeric features were mostly actual parameters (except for Age).

1. The Value counts for each unique value under the categorical columns were determined.

**Preliminary Feature Engineering:**

As part of preliminary feature engineering, data distribution and relationships were examined for each feature.

1. Boxplots were used to determine outliers. Extreme outliers in the Total Cholesterol and Systolic BP were removed.
2. A Statistical analysis was performed on the data frame using the describe() function.
3. Data Distribution Analysis was performed on each of the categorical features using Count Plots. It was found that the data for BP\_Meds, Diabetes and Prevalent\_Stroke were highly imbalanced. Also there was a large difference in the number of records for the target variable - Ten\_year\_CHD.
4. Data Distribution for Numeric Values was analyzed using distplots. It was found that the distribution for Cigarettes per Day was very uneven. Total Cholestrol, Systolic BP, Diastolic BP,Heart Rate, BMI and Glucose were more or less normally distributed.
5. In order to analyze the relationships between the different features, a Correlation heat map was generated. It was observed that both Education and Current\_Smoker had negative correlation with the target variable.

**Preliminary Feature Selection:**

1. Patient\_female\_ID was removed from the Dataset since it was a unique ID and would not have an impact on the target variable.
2. Education and Current\_Smoker were removed because they had a negative correlation with the target variable.

**Train- Test Split:**

1. Since the numeric variables were at a higher scale than the categorical variables, feature scaling was done using the Standard Scaler method.
2. Data was split into training and testing sets using the train\_test\_split method in a 3:1 ratio.

**Proposed** **Machine Learning Model:**

From the dataset, it is evident that the Ten year CHD is the target column which represents if a patient is likely to develop Coronary Heart Disease within the next 10 years or not. Most of the other columns are features that determine what the result would be in the Target column.

Given the pattern of the dataset, it has been proposed that a Classification model be used as our Machine Learning Model in order to predict the likelihood of Heart Disease in a patient. To begin with a Logistic regression model will be used. Further classification models such as Support vector Model, Random Forest Classifier and Decision Tree Classifier may also be used to get the best accuracy score.