Team ID: PTID-CDS-JAN22-1265

Project ID: PRCP-1016-HeartDieseasePred

Members:

1. Atharvdesh Pande

2. Nazeef Shaikh

3. Lukmanul Hakeem M

4. Sahil

5. Sumit Jangale

Project Analysis:

In this project we are trying to build a machine learning classification model for predicting whether a patient have heart disease or not. Data set used has 180 records and 14 input features including dependent feature. We have build model for this problem using scikit-learn’s Gaussian Naïve Bayes algorithm. Model has 0.94 recall and 0.83 accuracy scores on test set. This tells that our model can classify heart disease patients as correctly, have only 0.06 % misclassification.

Workflow:

We first go through domain studies and domain analysis. We try for getting insights from data by doing univariate and multivariate analysis. Then check data’s correctness by checking missing values and outlier presence.

There was outlier presence in three numerical features. Since three features are skewed we use IQR method to locate the outliers and replace it with median values. Before doing cross check with algorithms, we use ‘SMOTE’ to balance dataset and ‘Standard Scaler’ to scale all feature within similar range.

For cross checking by using cross validation we use scikit-learns classification algorithms. We found Gaussian Naïve Bayes with recall score 0.825 is best suited for our problem. After hyper-parameter tuning, model got 0.812 recall score on training set. Final evaluation on test dataset, model got 0.94 recall and 0.83 accuracy score.

Files:

* ‘main.ipynb’: main notebook file where modeling done.
* ‘description.docx’: contain brief description about Business problem and features in dataset.
* ‘values.csv’: file contains input features and values.
* ‘labels.csv’: file containing output values.