Prediction for Percutaneous Coronary Intervention Using Machine Learning Techniques

Abstract:

Percutaneous coronary intervention (PCI) refers to a procedures used to open clogged coronary arteries (those that deliver blood to the heart). By restoring blood flow, the treatment can improve symptoms of blocked arteries, such as chest pain or acute heart attack. Clogged coronary arteries is identified as a global health problem. To ensure healthy life it is essential to predict PCI in time and prior to applying required treatment. Since manual detection of diseases costs a large amount of time and money. In this paper we are proposing a machine learning approach to early prediction of PCI. After necessary pre-processing, the dataset was trained on with a range of different machine learning algorithms including that of KNN, Naive Bayes, Logistic Regression, Support vector machine and achieved an accuracy of over 98% when applied on the test dataset.

Index Terms—Imbalance Dataset, Percutaneous coronary intervention, Machine learning, Supervised learning, SMOTE.

1. Introduction:

PCI, formerly known as Angioplasty. It is a non-surgical procedure. PCI is used primarily to open a blocked coronary artery and restore arterial blood flow to heart tissue, without requiring open-heart surgery. In patients with a restricted or blocked coronary artery, PCI may be the best option to re-establish blood flow as well as prevent angina (chest pain), myocardial infarctions (heart attacks) and death. PCI is the urgent use of people with acute heart attack.

Machine learning techniques have recently established itself to be immensely useful in the field of medical diagnosis. These techniques require data to train their predictive models but the models intuitively has a tendency of biased prediction towards the majority class if the dataset is imbalanced. In our work, we used Mendeley dataset (imbalanced dataset) to train machine learning classifiers in order to predict PCI. The dataset contained majority class of negative data compared to positive data. This means many patients who need to PCI will be classified as no need to PCI resulting in life endangering consequences considering the associated risks of Heart attack. Predicting false negative is increasing the risk of a life. To achieve our goal, we applied synthetic minority over-sampling technique (SMOTE) and came up with an unbiased predictive model that can predict Percutaneous coronary intervention(PCI) with higher reliability.

2. Literature Review:

This paper [6] addresses they predict of preoperational time for patients with the acute coronary syndrome. Using data, they tried to predict time before the coronary stent operation (PCI) with regression methods. During this preprocessing, they divided health records into three clusters with k-means method and compared the results of cluster's prediction for five different classification methods. The results show that it is possible to classify initial data with the accuracy of 68.31 % on average.

A big data approach [5] that utilizes advanced machine learning algorithms identifies new associations among risk factors and provides high accuracy for the prediction of in-hospital mortality in patients undergoing percutaneous coronary intervention. A boosted ensemble algorithm (AdaBoost) had optimal discrimination with AUC of 0.927 compared with AUC of 0.913 for XGBoost, AUC of 0.892 for Random Forest, and AUC of 0.908 for logistic regression.

This paper [7] proposed enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques. They registered data on patient characteristics, clinical presentation, treatments, and outcomes associated with PCI at more than 1,500 participating sites across the United States. They divided into different training (70%) and test (30%) sets using 100 different random splits, and the performance of the models was evaluated internally in the test sets. They use logistic regression and gradient descent boosting. The model used 13 variables and achieved a significantly better area under the receiver operating characteristic curve (AUC) of 0.752 versus 0.711, a significantly better Brier score of 0.0617 versus 0.0636.

3. Proposed work

The main idea of this work is to create a Percutaneous coronary intervention (PCI) or coronary angioplasty detection model using machine learning algorithms that can be helpful for doctors to decide whether a patient needs to do PCI surgery or not. The data for this task is collected from Mendeley Data [2]. Jupyter Notebook, an open-source web application has been used to apply different machine learning algorithms to train our model.

4. Dataset

The dataset is originally from a research article titled "The association of chest pain duration and other historical features with major adverse cardiac events" published In "American Journal of Emergency Medicine", [3]. There are records of 1002 patients in the dataset and 41 features, where our target variable is Event PCI. Out of this 1002 record after deleting the null entries and entries without impact and noisy value we get 796 records to work on.

Summary of Dataset	
Number of total instances	1002
Number of instances used	796
Number of Event PCI True	759
Number of Event PCI False	37
Number of total Features	42
Number of Feature used	29

Table 1: Summary of Dataset

5. Attribute Selection

Initially our dataset had many unnecessary redundant attributes. After dropping those features we worked on 29 features

List of selected features given below

Selected Features	Description
Race	Race of the Patient
Length min	Length of the chest pain in minutes
How bad was CP? 0-10	How bad was the chest pain on the scale of 0-10
CP radiate Back?	Did the chest pain spread to the back of the patient
CP radiate L Shoulder/Arm	Did the chest pain spread to the left hand of the patient
CP radiate R shoulder/arm	Did the chest pain spread to the right hand of the patient
CP radiate neck/throat	Did the chest pain spread to the neck of the patient
CP radiate Abd	Did the chest pain spread to the abdomen of the patient
Nausea	Did the patient experience nausea while having chest pain
Vomiting	Did the patient vomit while having chest pain
Sweating	Did the patient experience sweating while having chest pain
Lightheadedness	Did the patient experience lightheadedness while having chest pain
Cough	Did the patient experience coughing while having chest pain
Quality CP Pressure	Chest pain described as “pressure”
Quality CP Tightness	Chest pain described as “tightness”

Quality CP Sharp/stabbing	Chest pain described as sharp or “stabbing”
Quality CP Pins/Needles	Chest pain described as “pins and needles”
Having heart attack?	Did the patient feel they were having a heart attack
Compare No Previous MI	Did the patient have heart attack previously
Compare Similar previous	Was the chest pain similar to previous heart attack
Compare different	Was the chest pain completely different from previous heart attack
Worse w/ activity?	Did the chest pain worsen with strenuous activities
Age	Age of the patient
Gender	Gender of the patient
Trop High	Was the Troponin level high
Acute MI	Did the patient experience Acute MI
Major Cardiac Event 6 wks	Did any major cardiac event happen after 6 weeks
Event MI	Did the patient experience MI
Event PCI	Did the patient experience PCI surgery

Table 2: Description of Dataset

6. Splitting Dataset

In our work, the dataset size was initially 1002 but after deleting null value entries and noisy value records our dataset size became 796. The dataset was then divided into two parts: training and test sets where training data contains 636 records (80% of the 796) and test data contains 160 records (20% of the 796). In our training data we see that out of 636 only 28 records are Event PCI True and rest 608 records are False so we are faced with imbalanced data, in order to make it balance we used Synthetic Minority Oversampling Technique (SMOTE) [4] to synthetically generate records of Event PCI True to balance our training data. After using SMOTE our training data contains 608 records of Event PCI False and 608 records of Event PCI True.

7. Methodology

Figure 1 illustrates the methodology we employed in our study. Firstly, we obtained the dataset and then, we preprocessed it. Then, each classifier was evaluated. Our dataset consisted of a total of 1002 instances and 41 features.

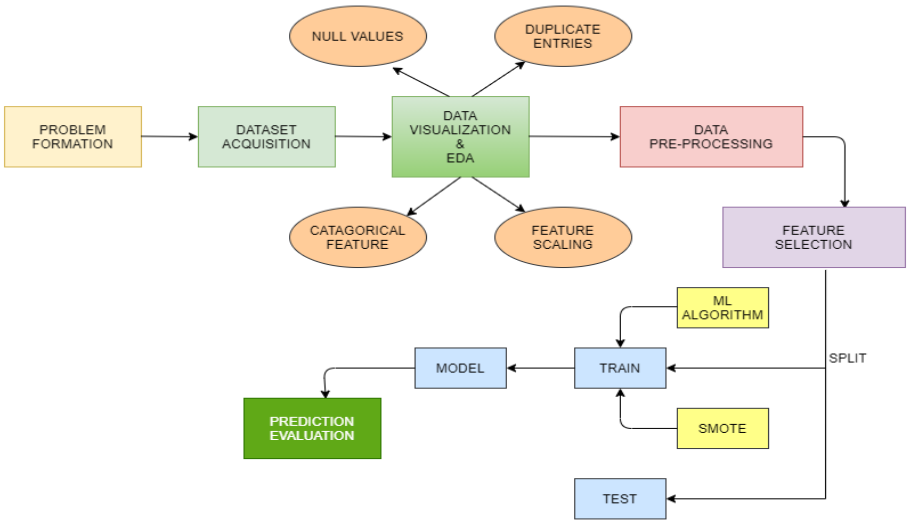

Fig 1. Flowchart of the Methodology

The dataset contains a total of 1002 instances and 41 features. There were some missing values in the dataset. Figure 2 illustrates the distribution of some of the attributes.

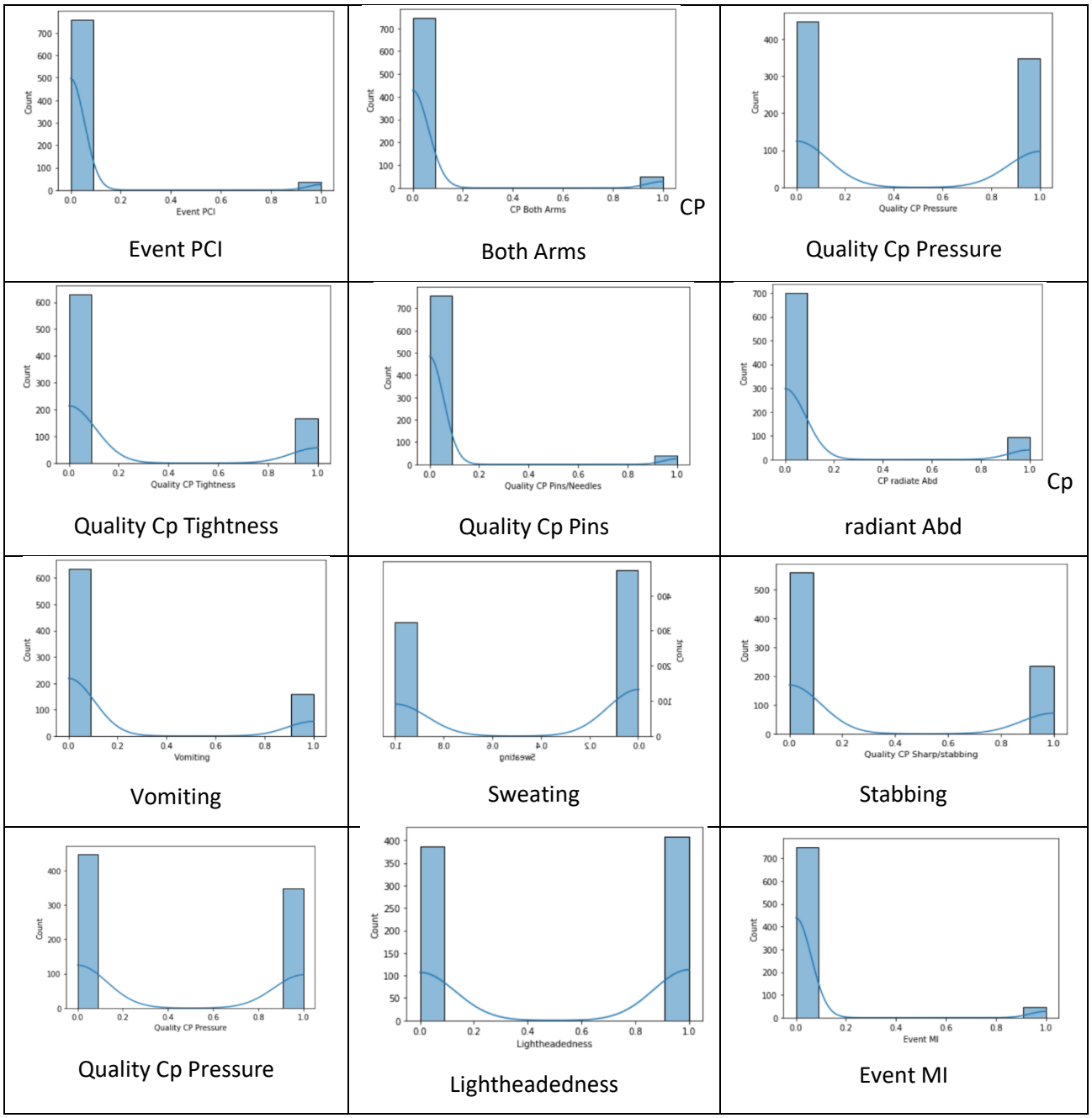

Fig. 2: Histogram representation of distributions according to the features

8. Machine learning algorithm implementation

Five different classifiers were used in this work.

8.1 K-Nearest Neighbor

K-nearest neighbor (KNN) determines the distances of each instance from the query instance. Following this, the classifier identifies the k nearest instances. Each of the k instances vote for the output. Using the voting technique, the predicted result of the query instance is determined and returned. We used Euclidean distance metric as a means of computing the distance.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

For this dataset, we found that k = 10 provides the best result.

8.2 Decision Tree classifier

The main motive of a Decision Tree is to develop a model that predicts classes or values of target variables using a decision tree, drawn from the training set of the data. This approach divides the data set into divisions by starting with the most appropriate attribute. The splitting procedure is repeated until all of the divisions are homogenous.

$$\mathrm{Gini}_{\mathrm{index}} = 1 - \sum_{\mathrm{i=1}}^{\mathrm{c}} \mathrm{p_i(t)}^2$$

Where, $pi(t)$ is the probability of class i belonging to node t.

8.3 Support Vector Machine

This is a supervised ML algorithm which inspects data for classification and regression problems. This algorithm is vastly used because of the superior accuracy it usually yields. In our research, the highest accuracy was obtained using Support Vector Machine.

8.4 Logistic Regression

Logistic regression is familiar for solving binary classification problem. It can, be extended to solve multi-classification problem. Logistic regression uses the concept of linear regression and apply sigmoid function on the top of it. In our research, Logistic Regression Classifier, also yielded superior results.

8.5 Multinomial Naive Bayes

Multinomial Naive Bayes algorithm is a collection of many algorithms, all of which share one common principle, which is, each feature being classified, is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of another feature. It is a probabilistic learning method.

9. RESULTS AND DISCUSSION:

Training and Test dataset contains 636 and 160 instances respectively and 28 attributes were chosen. Table I shows the accuracy of five classification algorithms after performing 10- fold cross validation on training data (80% of dataset)and test data (20% of dataset) where best five attributes were selected.

The performance of the classifier was assessed using five evaluation metrics which are described below.

9.1 Accuracy

The number of times the correct estimates made is referred to as accuracy. Accuracy can be calculated using the following equation.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Here,

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False negatives

9.2 Recall

Recall can be used to measure how well the model detects true positives. A high recall is an indicator that the model has done well identifying true positives. On the contrary, if the recall value is low, it indicates that the model encountered a lot of false negatives.

$$Recall = \frac{TP}{TP+FN}$$

9.3 Precision

Precision (also known as positive predictive value) refers to how well the model predicts positive values out of all positive values predicted by the model.

$$Precision = \frac{TP}{TP+FP}$$

9.4 Area under the ROC curve

The ROC curve depicts the relationship between the algorithm's sensitivity and specificity. The area under this graph is used as a performance metric as well.

9.5 F1-score

The harmonic mean of precision and recall is the F1-score. It can be calculated using the equation

$$F1 = 2\ \frac{Precision}{Precision+Recall}$$

Table 3: ACCURACY ON TRAINING AND TEST DATASET

Algorithms	*Accuracy On Training Set*	*Accuracy On Testing Set*
SVM	0.968%	0.980%
KNN(K=10)	0.971%	0.944%
Logistic Regression	0.984%	0.979%
Naive Bayes	0.957%	0.948%
Decision Tree	0.971%	0.973%

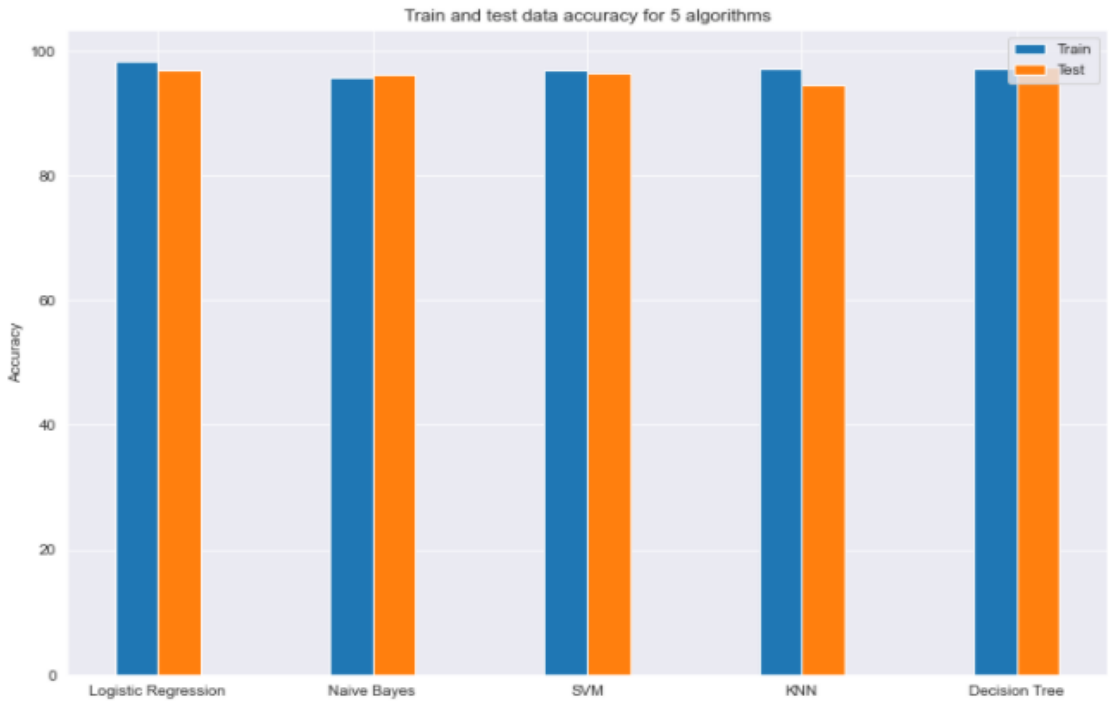

Fig. 3: Comparison graph among algorithm

The comparison between the accuracy of the five classification algorithms are represented in Figure 3. Besides accuracy, other performance measures like Precision value (Positive Predictive Value), Recall value (Sensitivity), F-Measure and AUC (Area under ROC) are also evaluated to compare among the five algorithms and it reveals in Table 4, SVM algorithm outperforms all other algorithms for predicting the result.

Table 4: Obtained results after applying the algorithms

Algorithms	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*AUC*
SVM	0.980%	1.0+/-0.4%	0.960+/-0.1%	0.980+/-0.2%	0.99%
KNN(K=10)	0.944%	1.00+/-0.1%	0.943+/-0.2%	0.680+/0.5%	0.80%
Logistic Regression	0.979%	0.987+/-0.2%	0.980+/-0.2%	0.983+/-0.2%	0.98%
Naive Bayes	0.948%	0.986+/-0.3%	0.974+/-0.2%	0.980+/-0.2%	0.97%
Decision Tree	0.973%	0.944+/-0.6%	1.00+/-0.6%	0.971+/-0.6%	0.82%

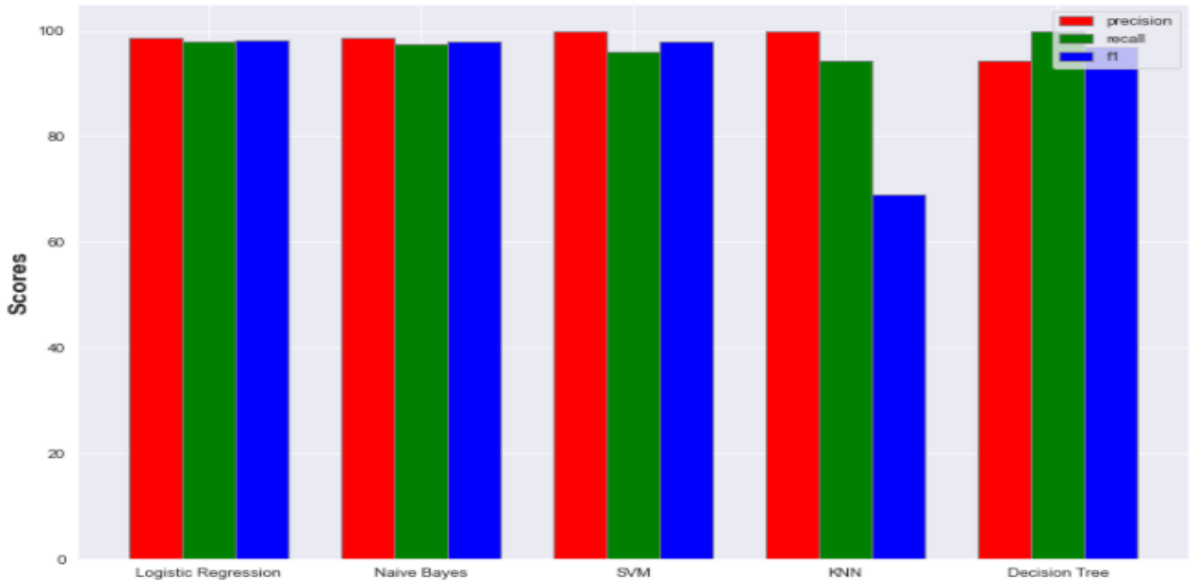

Fig. 4: Comparison graph among precision, recall, f1

Figure 5 shows the comparison graph among all the classifiers after implementing with K-fold cross validation.

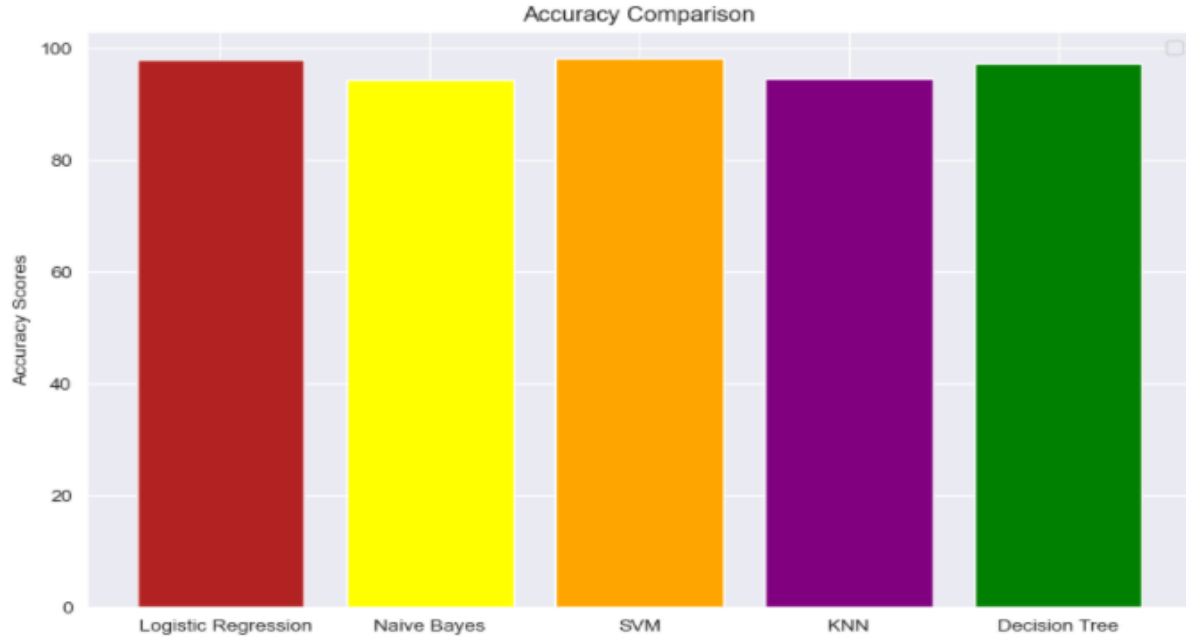

Fig. 5: Comparison graph among the classifiers

10. Conclusion

Percutaneous coronary intervention (PCI) is a treatment for persons experiencing myocardial ischemia (inadequate blood flow to the heart) or myocardial infarction (heart attack). This paper presents a machine learning approach to detect whether a person suffering from chest pain needs to do the PCI treatment or not cause chest pain is one of the symptoms of heart attack. A comparison between five machine learning algorithms (including that of KNN, Decision tree, Logistic regression, Naive Bayes, SVM). The algorithms preformed with varying degrees of accuracy. It was found that SVM performed the best with 98% accuracy on test data. Our data was initially imbalanced because there were less patient with PCI treatment. In order to reduce the biased behavior towards false negative we used SMOTE.

For future work, we aim to collect dataset with higher instances so that the proposed model can be further enriched to provide results with more confidence Further extension of the work also includes the performance evaluation of complex classifiers such as deep neural network (DNN) and other deep learning techniques. The research idea can further be used in other types of disease predictions where data insufficiency is a major challenge.

Reference:

1. T. Zitek, "Data for: The Association of Chest Pain Duration and other Historical Features with Major Adverse Cardiac Events," Mendeley Data, 27 April 2021.
2. https://data.mendeley.com/datasets/xzvtr7csdv/1.
3. E. C. M. A. G.-I. J. W. M. Tony Zitek MD, "The association of chest pain duration and other historical features with major adverse cardiac events," The American Journal of Emergency Medicine, vol. 38, no. 7, pp. 1377-1383, July 2020.
4. K. W. B. L. O. H. W. P. K. N. V. ChawlaN, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of artificial intelligence research, vol. 16, pp. 321-357, 2002.
5. https://www.ahajournals.org/doi/full/10.1161/JAHA.118.011160

6. https://www.sciencedirect.com/science/article/pii/S1877050916326886?fbclid=IwAR0iUEHObaA-3uAwDreYkhiCYNwdL-ilp1A1hxSA96aGTJp5CcEP6i4Y2A0