Predictive Cardiology: Machine Learning Insights

1. **Abstract**

Heart disease is a major death causing disease around the world. Early detection of heart disease can help many patients to start treatment before it is not fatal. Usage of Machine learning in health sector is significantly increasing and much useful. I present my study of using various machine learning algorithms to predict heart disease in the patient. Our findings can help researchers to improve their accuracy. Our aim in this paper is to explore the potential of the machine learning algorithms for heart disease prediction.

1. **Introduction**

According to the World Health Organization, Heart diseases account for 31% of global deaths each year. Machine learning algorithms have great potential in predicting the heart disease. By analyzing large amount of data, they find the patterns among the data and predict the new instance by following that pattern.

In this research paper, I explore various machine learning algorithms for the prediction of heart disease and compare the performance of the algorithms and use the most accurate algorithm for predicting new instances or new patient details whether they have heart disease or not. In this process we try to increase the accuracy or performance of the algorithm by processing the data like replacing null values, deleting duplicate values, removing outliners etc. based on the data distribution.

These methods can significantly vary the performance of the algorithm. Our study will contribute to the significant increase of the machine learning techniques in the heart disease prediction.

1. **Previous Related Works:**

Many previous studies were done to predict heart disease using machine learning algorithms. A paper contain heart disease prediction using machine learning techniques prepared a EHPDS model with a data set coating patient details like age, blood pressure, chest pain etc. using logistic regression and KNN algorithm got an accuracy of 87.5%. Many haven’t tried the ensemble learning methods which can increase the performance. Most studies have used clinical data such as age, sex, blood pressure, alcohol consumption etc. There is no standardized collection of data. Some have used advanced data such as electrocardiogram signals, echocardiography and coronary angiography. There are also some limitations in the previous studies like using only some kind of algorithms, the dataset they have used and processing of the data. The limitations made me to do further study in this section.

1. **Proposed Methodology:**

* **Dataset:**

I used a dataset from kaggle, which consists of 1025 instances and 13 features (age, sex, trestbps, chol , fbs, restecg , thalach, exang, oldpeak, slope, ca, thal, target)

* **Data Pre- processing:**

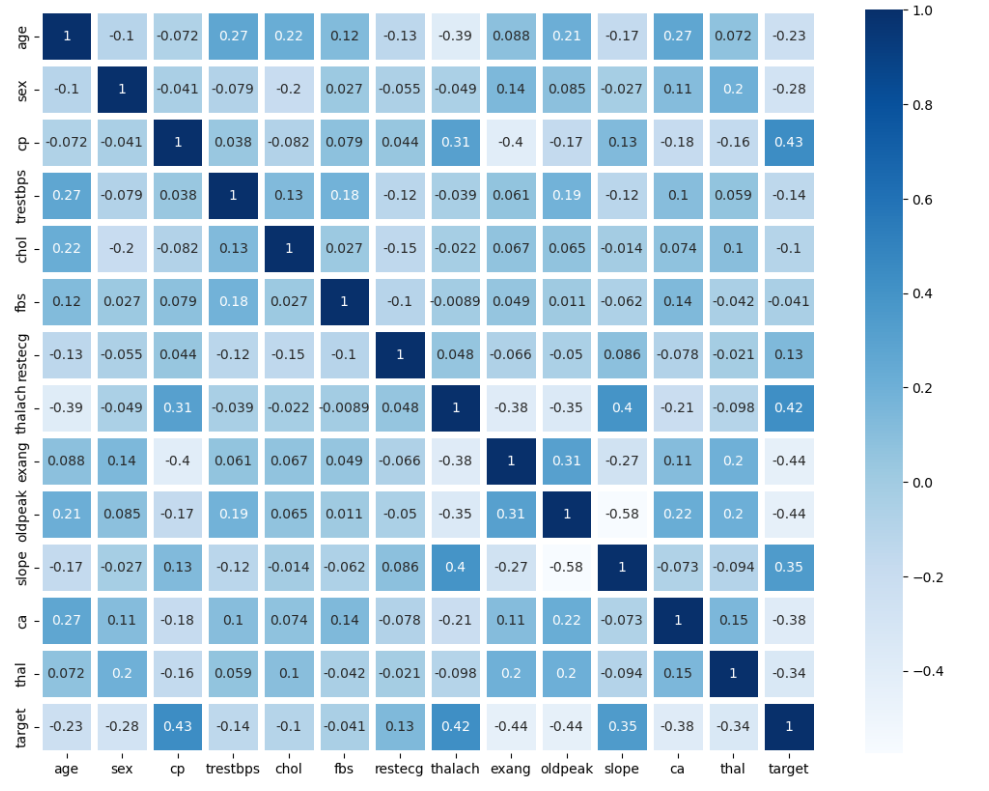
1. **Important Features:**

This technique sorts the input features based on the scores implied by the importance of each feature. In this dataset ca (Fluoroscopy colored major vessels) has the highest score which implies it is the most important feature among input features and FBS ( Fasting blood sugar) has the least score i.e., least important.

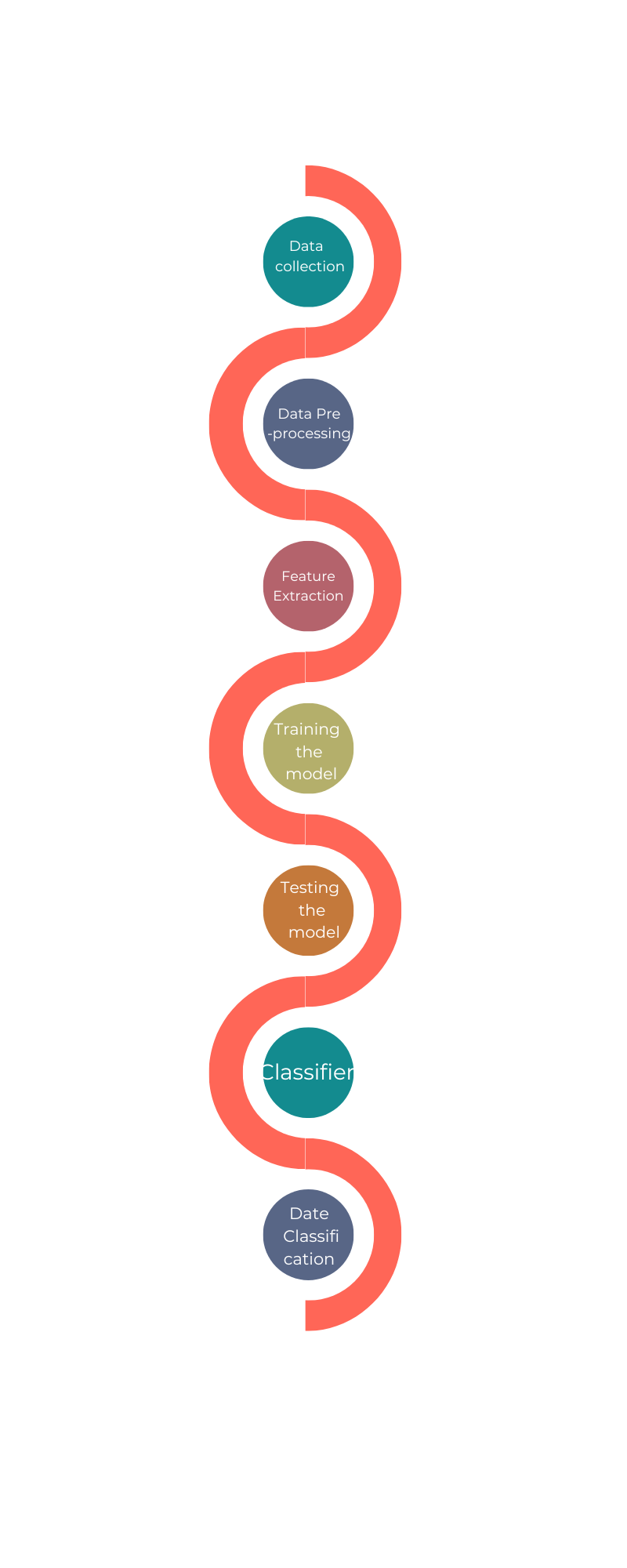
A graph with blue and white stripes

Description automatically generated

1. **Exploratory data Analysis:**
   1. Among the 1025 entries 529 patients have heart disease and 499 patients don’t.
   2. Some features have missing attributes. Replacing the missing values with the mean of the attribute.
   3. Many features in the dataset are correlated some are positively correlated and some negatively



* 1. Outliners, which are different from the remaining dataset that affects the performance of the model due to skewed data distribution. In this present dataset there are few outliners. Removing the outliners from the dataset can help us.
  2. Separating features and target, testing and training data from the dataset used to train and test the model
  3. Standardizing the features values, making the mean to zero and variance to one prevent adding weights to the feature with range values



**3. Applying Machine Learning Algorithms:**

* **Support Vector Machine:**

In SVM, the dataset is represented as points in a multidimensional space, where each point represents a feature of the dataset. SVM creates a boundary or hyperplane by selecting a subset of points, known as support vectors, that are nearest to the decision boundary. The distance between the support vectors and the decision boundary should be maximized, which ensures that the boundary has the maximum margin possible.

Using SVM we got testing accuracy of 80.36% and testing accuracy is 92.34%

**A screenshot of a computer screen

Description automatically generated**

* **Decision Tree Classifier:**

A Decision tree is constructed based on the importance of each feature using the information gain criteria and the remaining data is split into subsets based on that feature. The feature with highest information gain will be the root node. Based on that root node further tree is constructed till the stopping criterion is met. A new instance is classified based on the path of the tree.

Using decision tree Classifier, training accuracy is 100% and testing accuracy is 76.79%

A screenshot of a computer

Description automatically generated

* **Logistic Regression:**

Logistic Regression model predict the new instance by calculating the probability of the instance belongs to class or not. It uses maximum likelihood estimation method. This model uses sigmoid function to predict the new instance i.e. 0 or 1. This model is like linear regression except it uses sigmoid function instead of linear line

Using Linear Regression, training accuracy is 83.93% and testing accuracy is 85.59%

A screenshot of a computer program

Description automatically generated

* **Random Forest Classifier:**

The random forest classifier algorithm selects some random subsets from the training data and creates a decision tree for those subsets. Each decision tree is created using a different subset that is chosen randomly. This randomness helps to reduce the variance and over fitting of the model.

Using Random Forest Algorithm, training accuracy is 92.79% and testing accuracy is 87.5%

A screenshot of a computer

Description automatically generated

* **ADA Boost Classifier:**

ADA boost Classifier users boosting technique by training various weak learners using Decision Tree Classifier and aggregate them to make strong learner. All weak models are created parallel using the subsets of the main dataset. Ada boost aggregates all the weak learners using weighted average.

Using ADA Boost Classifier, training accuracy is 79.28% and testing accuracy is 78.57%

A screenshot of a computer screen

Description automatically generated

**Result and Analysis:**

We tried Support vector Machine, Decision Tree Classifier, Logistic Regression, Random Forest and ADA Boost Technique to detect heart disease. Comparison of these Machine Learning techniques is given below.

|  |  |  |
| --- | --- | --- |
| Algorithm | Training Accuracy | Testing Accuracy |
| SVM | 92.34% | 80.36% |
| Decision Tree | 100% | 76.79% |
| Logistic Regression | 83.93% | 85.59% |
| Random Forest | 92.79% | 87.5% |
| ADA Boost Technique | 79.28% | 78.57% |

1. **Conclusion:**

Through this project we implemented four machine learning algorithms on the dataset containing 1025 instances and 13 various features including Support Vector Machine, Decision trees, Logistic Regression and Random Forest to predict which patient is suffering with heart disease. Among these four algorithms Random Forest outperforms the remaining with 87.5% of testing accuracy and Decision Tree performs least with 76.79% of testing accuracy.

1. **References**

|  |
| --- |
|  |
| 1. Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. *SN Computer Science*, *1*, 1-6. 2. Singh, Archana, and Rakesh Kumar. "Heart disease prediction using machine learning algorithms." *2020 international conference on electrical and electronics engineering (ICE3)*. IEEE, 2020. 3. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554. 4. Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684-687. 5. Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, *7*(1), 129-137. 6. Yadav, A., Gediya, L., & Kazi, A. (2021). Heart disease prediction using machine learning. *International Research Journal of Engineering and Technology (IRJET*, *8*(09). 7. Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing. |
|  |  |

|  |
| --- |
|  |
|  |  |