Characterization of Binary Machine Learning Classifier for Robust Heart Disease Prediction

Aishwarya Soman Nair, Ankit Sharma, Ishita, Parag Vijayvargiya, Hiren Kumar Thakkar, *Member, IEEE*

Abstract: *Cardiovascular diseases (CVDs) – disease of the heart or blood vessels commonly referred to as heart disease is one of the most common cause of death of a number of people worldwide. So before curing heart disease it is necessary to detect if the person is suffering from heart disease or not. Machine learning provides a platform to detect if heart disease is present or not. Numerous health organizations or societies keep information of the patients. In machine learning language this information also known as features acts as a dataset which is necessary to carry out the analysis. This problem is supervised machine learning algorithm so various algorithms can be used which are discussed in this paper. The primary objective of this research paper is to carry out analysis of various machine learning models such as Logistic Regression, Naïve Bayes, Decision trees, random forests, K nearest Neighbors.*

Keywords: *Machine learning, Heart disease prediction, supervised learning, optimization*

I. INTRODUCTION

Heart- the most important body organ and most sensitive, its smooth functioning is important for a better and healthy living [1]. But a number of factors affects its smooth functioning such as change in blood pressure, cholesterol level and many more which leads to a number of fatal diseases. So its detection is important so that patient gets cure on time otherwise it would be life threatening.

Additionally, along with detection, correct and proper detection is important because getting wrong or false decision will be fatal. So it becomes challenging task to detect the probability of having heart disease but it becomes relatively easy using machine learning algorithms.

So selecting relevant features and working on them is important so as to get good accuracy and prediction. Some of the features with terms related to heart are resting blood pressure, chest pain types, thal, resting EKG results etc. And some of the common features are sex and age [2]. Along with these features the patient id and corresponding feature values are also provided. With the help of these features and some of the machine learning algorithms for a supervised learning problem like Decision Trees, Logistic Regression, Naïve Bayes the probability of having a heart disease was found. All these algorithms or models were used to get the most accurate results or least log loss.

II. RELATED WORK

Nowadays, efficient and reliable and inexpensive mobile healthcare systems are becoming basic need of people worldwide [3]. In recent years, many experiments were conducted to provide people with this type of technology and assist people with their up-to-date healthcare information. Machine learning based approaches are used in cardiac health monitoring which are quite reliable as compared to other approaches such as IoT based and others. Most of the papers have implemented several data mining techniques

such as Logistic Regression (LR), Naïve Bayes (NB), Decision Trees (DT), Random Forests (RF), K-Nearest Neighbors (KNN) showing different levels of accuracy on multiple datasets of patients from around the world. In one of the recent research paper: On the design of an efficient cardiac health monitoring system through combined analysis of ECG and SCG signals [4] with the help of the above mentioned signals cardiac health is monitored. Many more different efforts were also made such as *Lifeguard* [5], which is one of the earliest approaches, that monitors heart rate, respiration rate, temperature, blood pressure and electrocardiogram. Lifeguard uses signals that are sent to the base station via Bluetooth-enabled cellphone and that activates the buzzer alarm wherever abnormality is detected. In [6], ECG based body networks also raises alarm when a heart attack is detected. But sometimes we get false alarm too due to bad quality signals, so in that case we use [7] which is a false arrhythmia alarm reduction framework designed using machine learning. Sometimes monitoring of a person becomes challenging especially when that person is an athlete or is engaged in activities like cycling. So to monitor such cases a special *Smart Helmet* is used that measures ECG signals and respiratory signals in an uninterrupted manner [8].

III. METHODOLOGY

Our methodology for the final model involves taking the logistic regression as our model to find our results.

The main idea behind logistic regression is that it is a binary classifier that is it classifies data into two different classes 0 and 1. The two classes are separated by a boundary also known as decision boundary.

Before selecting the appropriate model for our problem many more different techniques were applied on the training and testing data such as preprocessing which includes normalization, feature extraction and one hot encoding and another technique is hyper parameter tuning. As the dataset provided was already separated into test set and training set so splitting of data wasn't needed.

(a) Normalization

Normalization was performed on the real valued features that are max heart rate, age, oldpeak_eq_st_depression, serum_cholesterol_mg_per_dl, resting_blood_pressure of both training and test sets. Normalization is changing the values in the dataset to a common scale. Normalization was carried out using MinMaxScaler to get the value of each feature between 0 and 1.

(b) Feature Extraction

Then another preprocessing technique was used- feature extraction. Under feature extraction two methods were tested Filter method (Pearson Correlation) and Recursive Feature Elimination (RFE). In filter method we got four important or relevant features namely thal, chest_pain_type, num_major_vessels, and exercise_induced_angina. In RFE we got seven important or relevant features which are slope_of_peak_exercise_st_segment, thal, chest_pain_type, num_major_vessels, fasting_blood_sugar_gt_120_mg_per_dl, sex, exercise_induced_angina.

(c) One Hot Encoding

After the feature extraction, another preprocessing technique called one hot encoding was used on the categorical data of both training and test datasets in which the categorical value represents the numerical value of the entry in the dataset. So one hot encoding was performed on the following features slope_of_peak_exercise_st_segment, thal, chest_pain_type, num_major_vessels, and fasting_blood_sugar_gt_120_gm_per_dl.

Finally different combinations of dataset like normalized data with one hot encoding, RFE with one hot encoding, Filter method with one hot encoding, were tested on different models or algorithms and classifiers and log loss was noted. It was observed that logistic regression model gives the best results out of the other tested models in terms of least log loss and max accuracy.

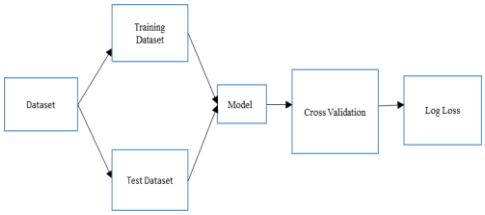

Fig 1: Model Architecture

In Fig 1, model architecture is shown that is how we reached the end of this model or the procedure we followed to get our results.

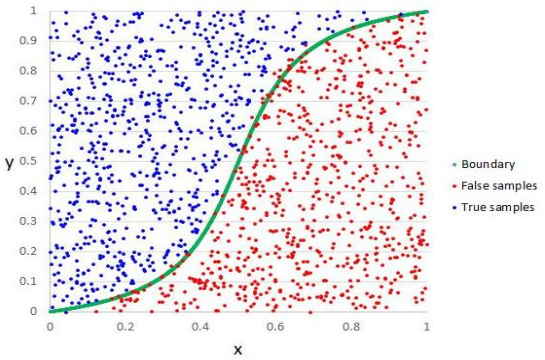

Fig 2: Logistic Regression

As you can see in Fig 2, logistic regression divides the data into two different classes or samples that is true samples and false samples separated by a boundary and same goes for our problem.

IV. EXPERIMENTAL RESULTS

The dataset used in this problem is the dataset provided to us during the time of competition [2].

The training dataset consists of 13 essential features and 180 instances and the test dataset also consists of 13 essential features but 90 instances. Both the datasets have an additional column of patient_id and the dataset for the training labels or outputs is also provided.

The metric used for the evaluation is Logarithmic Loss or log loss or cross entropy loss as shown in Eq 1:

$$\text{Log loss} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad (1)$$

Where $\hat{y}_i$ represents probability of y=1

The main goal was to reduce this log loss which implies a good accuracy. The method involved was passing the data through the model, noting the log loss and repeat this method for different models.

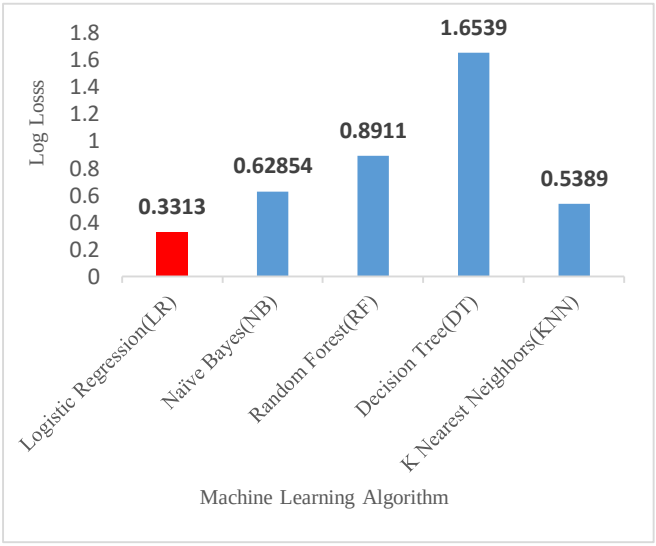

Fig 3: Graphical Representation of log loss

The graph in Fig 3 shows the log loss values obtained from the various models. As shown, Logistic regression gives least log loss among these and decision trees maximum.

BEST	CURRENT RANK	# COMPETITORS	SUBS. TODAY
0.33130	158	2688	0 / 3

Fig 4: The competition rank and final log loss.

V. CONCLUSION

The primary emphasis in this research is on the use of various data mining techniques and most importantly on Logistic Regression and preprocessing of data. After applying numerous models we found that logistic regression with normalized and one hot encoded data gives the best result.

The other models tried were Decision trees, Random Forests, Naïve Bayes, K-Nearest Neighbors. Support vector machines (SVM) was already discarded as it gave absolutely garbage results.

VI. REFERENCES

1. https://www.world-heart-federation.org/resources/cardiovascular-diseases-cvds-global-facts-figures/
2. http://archive.ics.uci.edu/ml/datasets/statlog+(heart)
3. Annus, P., Samieipour, A., Rist, M., Ruiso, I., Krivoshei, A., Land, R. ... & Min, M. (2013). Wearable Data Acquisition System of Multimodal Physiological Signals for Personal Health Care. *Studies in health technology and informatics*, *189*, 107-112.
4. Sahoo, P., Thakkar, H., & Lee, M. Y. (2017). A cardiac early warning system with multi-channel SCG and ECG monitoring for mobile health. *Sensors*, *17*(4), 711.
5. Mundt, C. W., Montgomery, K. N., Udoh, U. E., Barker, V. N., Thonier, G. C., Tellier, A. M., ... & Ruoss, S. J. (2005). A multi parameter wearable physiologic monitoring system for space and terrestrial applications. *IEEE Transactions on Information Technology in Biomedicine*, *9*(3), 382-391.
6. Wolgast, G.; Ehrenborg, C.; Israelsson, A.; Helander, J.; Johansson, E.; Manefjord, H. Wireless body area network for heart attack detection [Education Corner]. IEEE Antennas Propag. Mag. 2016, 58, 84–92.
7. Tanantong, T., Nantajeewarawat, E., & Thiemjarus, S. (2015). False alarm reduction in BSN-based cardiac monitoring using signal quality and activity type information. *Sensors*, *15*(2), 3952-3974.
8. Von Rosenberg, W., Chanwimalueang, T., Goverdovsky, V., Looney, D., Sharp, D., & Mandic, D. P. (2016). Smart helmet: Wearable multichannel ECG and EEG. *IEEE journal of translational engineering in health and medicine*, *4*.
9. Sahoo, P., Thakkar, H., Lin, W. Y., Chang, P. C., & Lee, M. Y. (2018). On the design of an efficient cardiac health monitoring system through combined analysis of ecg and scg signals. *Sensors*, *18*(2), 379.