Heart disease prediction using Machine Learning techniques

P.Mohana Vamsi   Shardul Thakare   Ashish Mishra   Suryansh Raj   B.Sangeetha

Dr. Tanveer Ahmed

Abstract

- ***Huge amount of patient related data is maintained in the health industry. This stored data can be useful for predicting future diseases.***
- ***The complex task of finding heart diseases requires lots of experience and knowledge. Some of the many ways of predicting it are stress test, ECG and heart MRI etc.***
- ***Our model uses 13 parameters for predicting heart disease that include blood pressure, chest pain, cholesterol level, heart rate,etc.***
- ***The main aim this model is to provide analysis of various machine learning models like SVM, k-NN, Decision tree, Random Forest, XG-Boost, Cat-Boost, Light GBM, Artificial Neural Networks (ANN) and using stacking to create models for efficient prediction of Heart diseases.***

Introduction

The heart is one of the most important organs in the human body. Life is dependent on the heart functioning well. Each year, around 17.9 million people die due to heart diseases. People with cardiovascular diseases or those who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes) need early detection and management using counselling and medicines, this is why prediction is so important when it comes to these diseases. We plan to achieve that by using machine learning algorithms.

Figure 1: Complexity in Diagnosis of Heart Diseases

Proposed Method

As you can see in this chart, SVM, LightGBM and XGBoost were used in ensemble using the averaging method to get our model.

- SVM: SVM refers to Support Vector Machines. An n-dimensional space is formed by the numerical input variables(x). On trying the various kernels, RBF was found to be the most accurate one, and the one that gave the least log loss.
- LightGBM: Light GBM is a gradient boosting framework based on decision tree algorithm, which we have used for classification here.

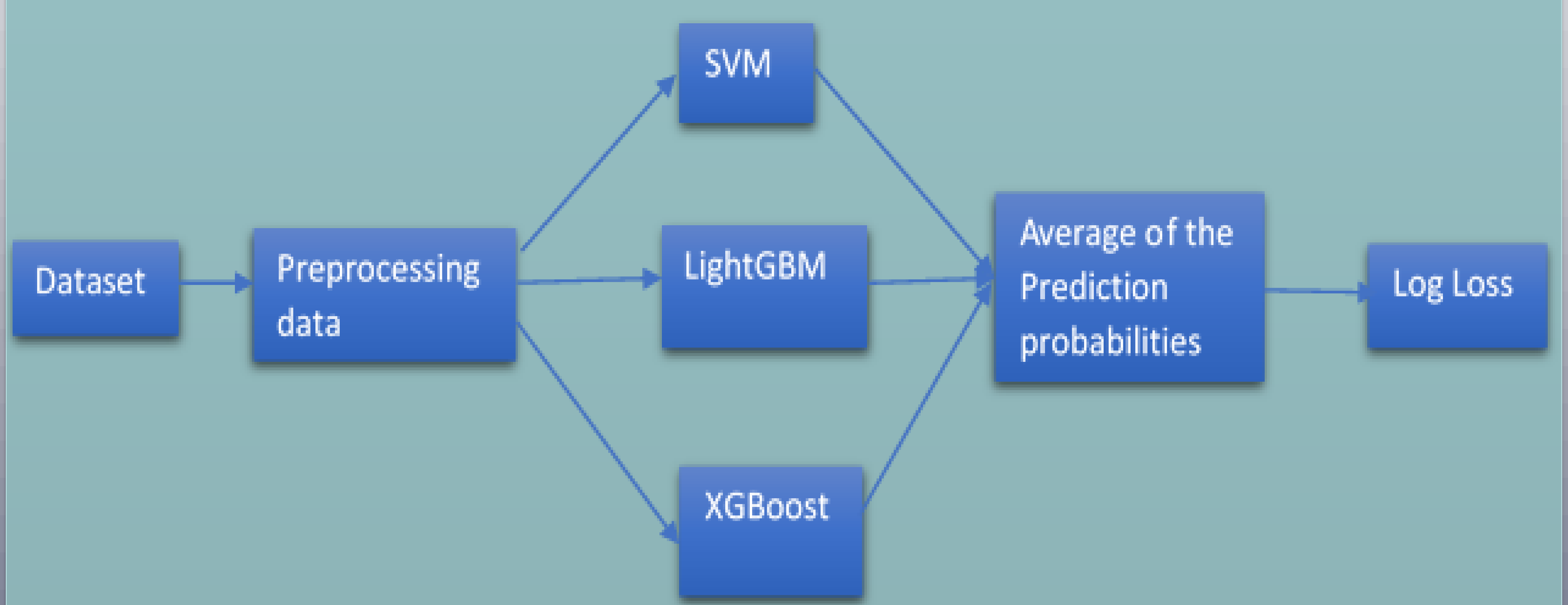

Fig 2: Architecture of the model

Experimental Results and Discussion

- The dataset has 13 essential features, and 180 instances.There features are not dependent on each other.
- The features include:Patient_id,Slope of the peak exercise ST segment,Thal,Resting Blood Pressure,Chest pain type,Number of major vessels colored by fluoroscopy,Fasting Blood Sugar,Resting Ekg Results,Serum Cholesterol in mg/dl,ST depression induced by exercise relative to rest,Sex,Age,Max Heart Rate achieved,Exercise induced angina.

The metric used for the competition is logarithmic loss.

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

This is the formula for log loss. $y`$ is the probability that $y=1$. Logarithmic loss provides a steep penalty for predictions that are both confident and wrong.

- The goal is to minimize the log loss. The testing procedure involved us passing the data through the model, noting down the accuracy and log loss and repeating it for the various models.

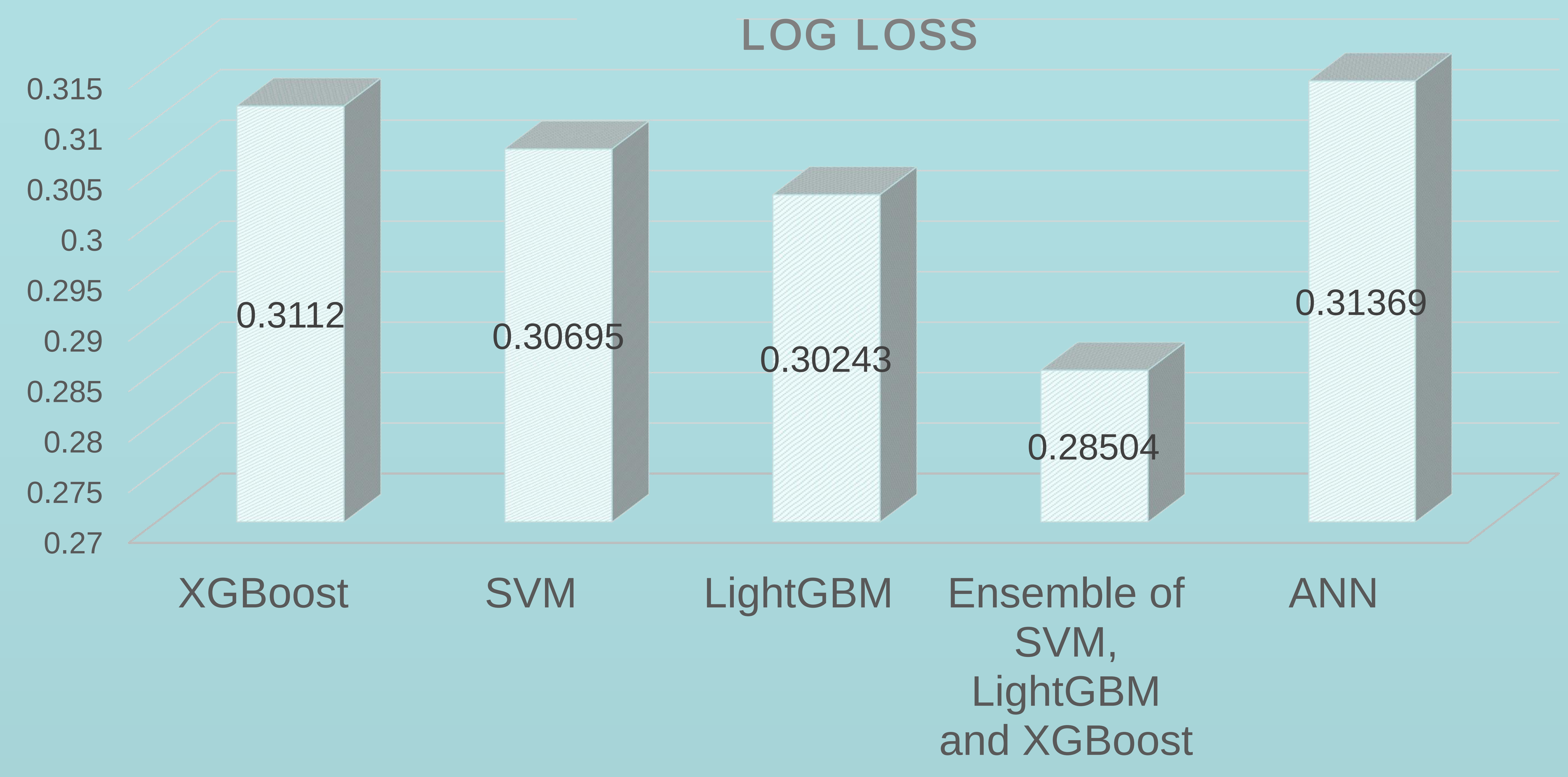

Fig 3: Log Loss results

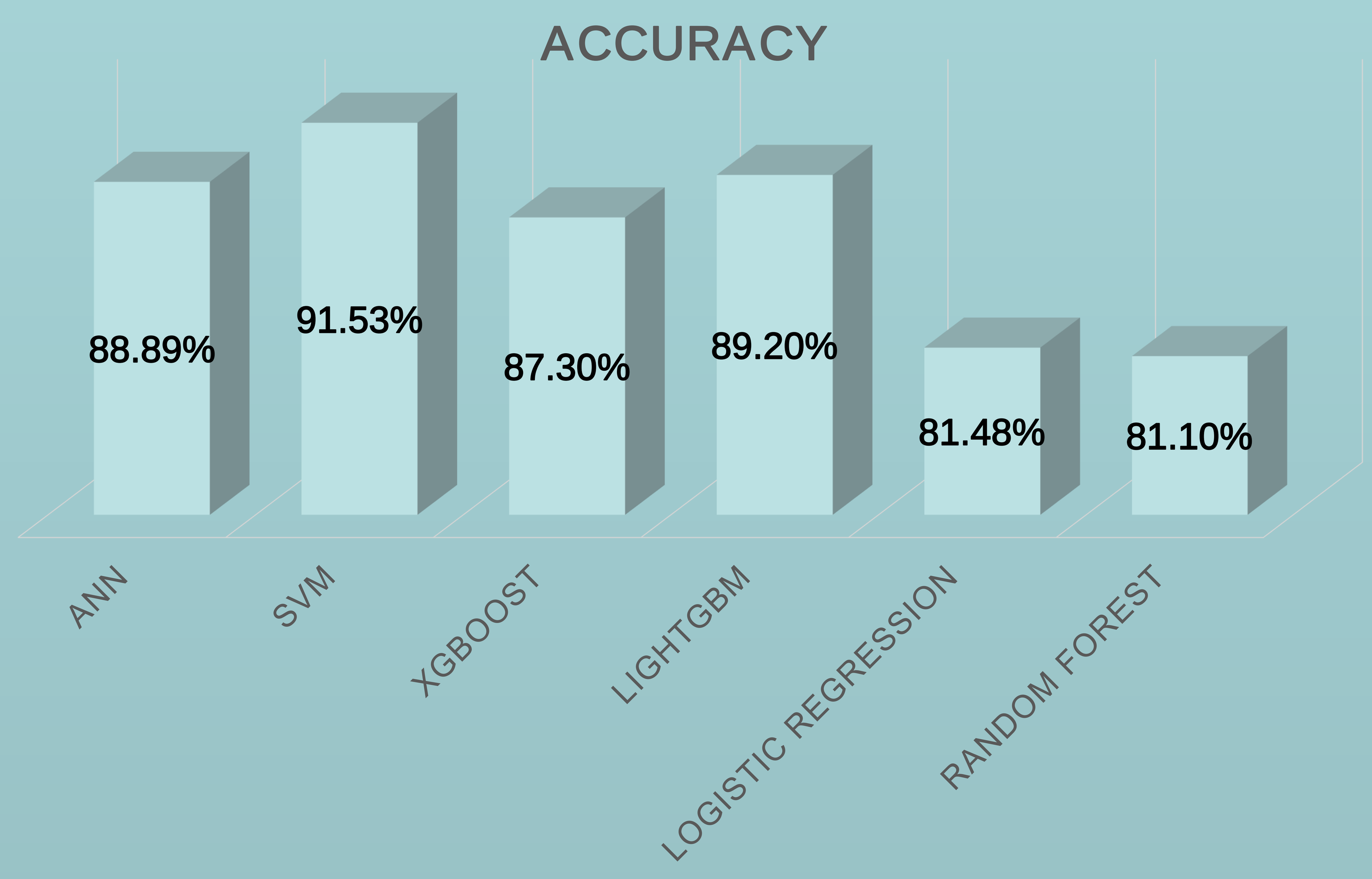

Fig 4:Accuracy results

These two graphs show us the best metrics that we have gotten from our models. Log loss is the defining metric but we included the top 6 performers on basis of accuracy as well.

Conclusions

This experiment provides the deep insight into machine learning techniques for classification of heart diseases. After applying numerous models, and defining log loss as the evaluation metric, we have found the ensemble of SVM, LightGBM and XGBoost to be the best model for achieving the metric laid down. Averaging the probabilities out was the best way to get the least log loss even if the individual models didn't give the best log loss by themselves.

References

1. H. I. Elshazly, A. M. Elkorany, and A. E. Hassanien, "Lymph diseases diagnosis approach based on support vector machines with different kernel functions," Computer Engineering & Systems 9th International Conference (ICCES),Cairo, pp. 198–203, 2014
2. https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc
3. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

Team No. - 15