Evaluating the Performance ofk-NN, Random Forest and Logistic Regressionsfor Predicting the Cardiac Disease

Dr.WINSTON DUNN
Department of Computer Engineering
SSBT's College of Engineering and Technology
Jalgaon M.S. India

.Dr.BUCIO PITTY
Department of Computer Engineering
SSBT's College of Engineering and Technology
Jalgaon M.S. India

Abstract

In medical science, making predictions based on past history and a prescription of the patient is used to provide the preventive health care andimprove the diagnosis for patient well-being.Availability of the past history of a particular patient is one of the central challenges in medical sciences. This is required for manual preventives, which are error-prone, tiresome and time consuming. Regression analysis in machine learning is used to determine the relationship between a single dependent variable from more independent variables and yield predictions from combinations of the independent variables. In Machine learning there are plenty of models used for regression analysis, in this paper, we have used and evaluated the performances of k-NN, random forest and logistic regressions for modeling and predicating the heart attacks in the patients. For the test runs for all regression models, we have used cardiac disease past history of 300 under treatment patients harvested from the UCI Machine Learning Repository. We have obtained the highest accuracy of 92.50% for predicting cardiac disease using logistics regression.

Keywords:k-Nearest Neighbor, Random Forest, Logistic Regression, Preventive Health Care and Feature Selection.

1. Introduction

The use of healthcare methods to prevent diseases is known as preventive healthcare (Hugh & Leavell, 1979). Primal, primary, secondary, and tertiary preventive measures are the cornerstones of disease prevention (Mokdad , Marks, Stroup, & Gerberding, 2004). Millions of individuals die from avoidable causes each year. A 2004 study revealed that avoidable exposures and behaviors were to blame for almost half of all deaths in the United States in 2000. Cardiovascular disease, chronic respiratory disease, accidental traumas, diabetes, and several viral disorders were the main contributing factors. The same study suggests that poor diet and a sedentary lifestyle cause 400,000 deaths annually in the United States. In 2011, the World Health Organization (WHO)

estimated that 55 million people died worldwide, with non-communicable diseases such cancer, diabetes, chronic cardiovascular diseases, and lung diseases accounting for two thirds of these deaths (The top 10 causes of death, 2020). Given the rise in chronic illness prevalence and related fatalities worldwide, preventive healthcare is extremely crucial. There are numerous approaches to illness prevention. One of them is stopping teen smoking by providing information (Isensee & Hanewinkel, 2018).

Regression analysis is a group of statistical procedures used in statistical modeling to determine the relationships between a dependent variable and one or more independent variables (often referred to as "predictors," "covariates," "explanatory variables," or "features") (Patil S. , 2023). In linear regression, the most typical type of regression analysis, the line (or a more complicated linear combination) that most closely matches the data in terms of a given mathematical criterion is found. Two conceptually separate uses of regression analysis predominate (Brzezinski & Knafl, 2019) (Ran, Zheng, & Wang, 2018). First, there is a significant overlap between the usage of regression analysis and machine learning in the areas of prediction and forecasting. Second, regression analysis can be used to infer causal links between the independent and dependent variables in specific circumstances. Regressions by themselves, it should be noted, only illuminate connections between a dependent variable and a group of independent variables in a given dataset.

In Machine learning there are plenty of models used for regression analysis, in this paper, we have used and evaluated the performances of k-NN, random forest and logistic regressions for modeling and predicating the heart attacks in the patients (Hilbe) (Zou, Hu, Tian, & Shen, 2019). Detailed discussions on regression analysis for heart diseases with prior art, detailed philosophy, and empirical results are discussed in the following sections (Yang & Li, 2019)(Jaiswal & Samikannu, 2017) (Jangale & Patil, 2022) (Bhaise & Patil, 2021) (Patil & Patil, 2020).

2. Prior Art

Many studies have been conducted to investigate the use of regressions for the predictions of disease.Most of the studieshave no difficulty in predictionsbut type of disease remains a challenging problem. For the classification of cardiac disease, FirdaAninditaLatifah et al. (Latifah & Slamet, 2020) presented a comparison analysis of two machine learning models, logistical regression and random forest. The study used the 3656-record Framingham dataset. The model was accurate 85% of the time.

In order to classify cardiovascular disease, Zameer Khan et al. (Khan, Mishra, Sharma, & Sharma, 2020) presented an empirical research of many machine learning methods, including logistic regression, KNN Classifier, RF, SVM, Decision Tree, and Gauss Nave Bayesian. used data from UCI. The accuracy of the logistic regression was 85% of the time.

On the basis of the 4238 records in the Framingham dataset, ThanujaNishadi A S et al. (Nishadi, 2022) suggested a logistic regression model for classifying heart disease. The accuracy of the logistic regression was 86.66%.

Logistic regression, decision trees, logistic regression support vector machines, etc., were proposed by Saba Bashir et al. (Bashir, Khan, Khan, Anjum, & Bashir, 2019). The study used the UCI dataset, and the accuracy of the logistic regression and logistic regression support vector machine was 82% and 84%, respectively.

3. Motivation

To prevent the disease's compounding effect in low-income or developing nations, the medical industry's main task at the moment is to forecast life-threatening cardio vascular disease as its high mortality rate contributes to nearly 20 million deaths all over the world. Early diagnosis helps to treat the disease in timely manner to prevent mortality using a more affordable and reliable way. Early detection enhances quality of life while also lowering costs.

4. Methodology

The study of the heart disease UCI dataset is carried out in the suggested system utilizing appropriate data collecting, preprocessing by cleaning the data, and then using selection of all the characteristics which have high correlation with the goal function. The k-NN, random forest and logistics regression model are then trained and evaluated to determine whether or not it could detect cardiac disease.

The process for developing a k-NN, random forest and logistic regression model for classifying and predicting the heart diseases is shown in Fig. 1.

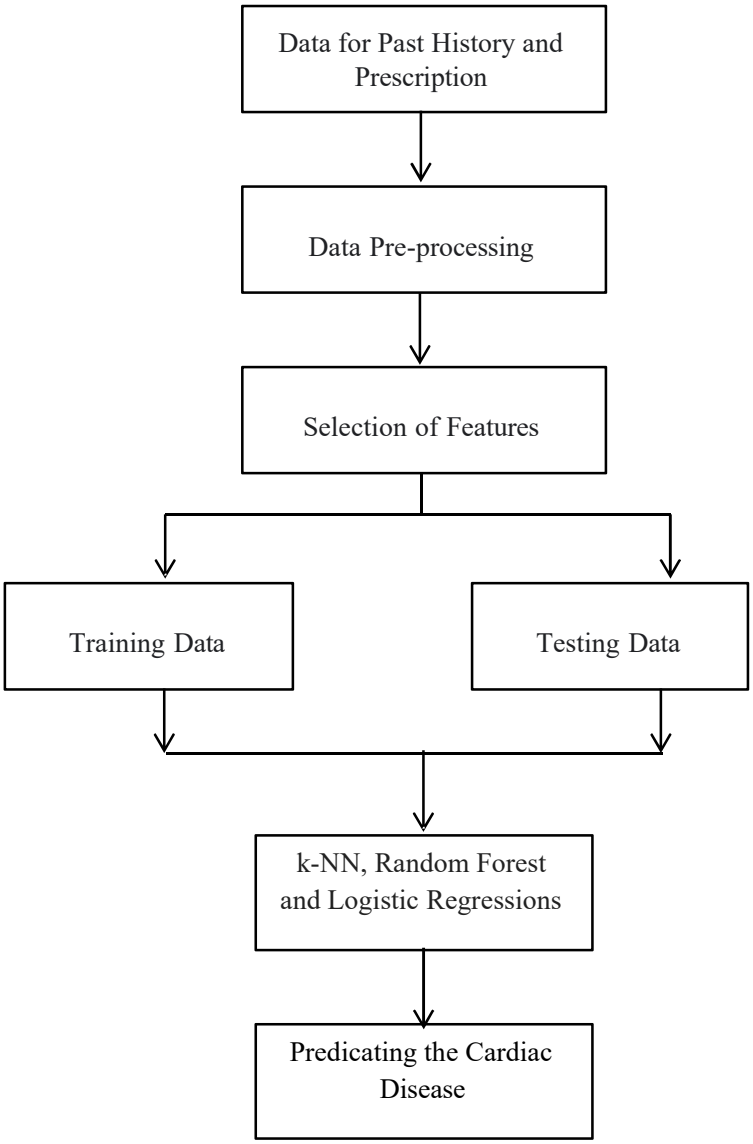

Figure 1. Regression Analysis for Predicating the Cardiac Disease

4.1 Procedure for Predicating the Cardiac Disease

Input: Feature selected data

Output: Best classification and prediction

For i ← 1 to k

For each training & testing data instance d_i in the dataset

Set the target value for the k-NN, Random Forest and Logistic Regressions

Initialize the weight of instance

Finalize the data with class values

Classification of label decision

Assign (class label:1) if P (yes|d) >0.5, otherwise (class label:2)

4.2 Features Selected for Predicating the Cardiac Disease

We have selected the following features from the past history and prescription of the patient

1. age: age in years
2. sex: (male/ female)
3. cp: chest pain types
   - 0: Typical angina: chest pain related decrease blood supply to the heart
   - 1: A typical angina: chest pain not related to heart
   - 2: Non-anginal pain: typically esophageal spasms (non heart related)
   - 3: Aysmptomatic: chest pain not showing signs of disease
4. Trest_bps: resting blood pressure (in mm Hg on admission to the hospital): anything above 130-140 is typically cause for concern
5. Chol: serum cholesterol in mg/dl
   a. serum = LDL + HDL + .2 * triglycerides
   b. above 200 is cause for concern
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
   a. '>126' mg/dL signals diabetes
7. restecgresting electrocardiographic results:
   - 0: Nothing to note
   - 1: ST-T Wave abnormality
     - can range from mild symptoms to servere problems
     - signals non-normal heart beat
   - 2: Possible or definite left ventricular hypertrophy
     i. Enlarged heart's main pumping chamber
8. Thalach: maximum heart rate achieved
9. Exang: exercise induced angina (1 = yes; 0 = no)
10. Oldpeak: ST depression induced by exercise relative to rest
    a. looks at stress of heart during excercise unhealthy heart will stress more
11. slope: the slope of the peak exercise ST segment
    - 0: Upsloping: better heart rate with exercise (uncommon)
    - 1: Flatsloping: minimal change (typical healthy heart)
    - 2: Downslopins: signs of unhealthy heart
12. Ca: number of major vessels (0-3) colored by flourosopy
    a. colored vessel means the doctor can see the blood passing through
    b. the more blood movement the better (no clots)
13. thal: thalium stress result
    a. 1,3: normal
    b: fixed defect (used to be defect but ok now)

c: reversable defect (no proper blood movement when excercising)

Prediction: have disease or not (1=yes, 0=no)

4.3 Regression Analysis for Predicating the Cardiac Disease

Acquisition of Data for Past History and Prescription: The cardiac disease data harvested from the UCI. It contains 13 features and 300 patients' records under treatment.

Data Pre-processing: A cardiovascular dataset is initially loaded and all records are then subjected to data cleansing and missing value searches. The dataset provides all necessary data. The dataset's attributes are multiclass variable in nature and doubly classified.

Selection of Features: The patient record is individually identifiable by two dataset features, sex and age, which are selected from 13 dataset variables and assign individual ids. The remaining features are all related to medicine. The medical data are essential indicators for predicting heart disease.

4.4 Classification and Predicating the Cardiac Disease:

The k-Nearest Neighbors Regressions: It is a non-parametric, supervised learning regression, which uses proximity to make classifications or predictions about the grouping of an individual data point required in predicating the cardiac disease. It approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood.

Random Forest Regression: It is supervised ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction for cardiac disease.

Logistic Regressions: It is also used to estimate the relationship between a dependent variable and one or more independent variables; it is also used to make a prediction about a categorical variable versus a continuous one for cardiac disease.

5. Experimental Setup, Test Bed and Evaluation Strategy

The proposed system is implemented using python ML like sklearn, pandas, and matplotlib in a Python 3.7 environment. To run the code, use googlecolaboratory and tested on data-bed which is contains cardiac disease data of total 300 patients harvested from UCI data (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). This dataset is used to test the performance of the experimentation conducted on k-NN, random forest and logistic regressions.

To evaluate the performance of proposed cardiac disease prediction system, we have divided the training and testing data on 4 folds of 300 records shown in table 1,

Table 1: Split Percentage of Training and Test Set.

Fold No.	Training Set	Test Set
Fold_1	60%	40%
Fold_2	70%	30%
Fold_3	80%	20%
Fold_4	90%	10%

We have recorded the following measures and used them in estimating the classical automated metrics:

1. The number of cardiac disease correctly predicted (TP)
2. The number of cardiac disease missed during prediction (FN)
3. The number of healthy patients wrongly predicted as cardiac possibility(FP)
4. The number of healthy patients correctly eliminated (TN)

5.1 Classical Metrics Estimated

Exactness of Cardiac Disease Prediction: It is the percentage of predicted cardiac disease that is actually correct. It is estimated using Equation 1.

$$\text{Exactness of Disambiguation} = \left(\frac{TP}{TP+FP}\right) \quad 100 \qquad (1)$$

Sensitivity of Cardiac Disease Prediction: It represents the percentage of relevant (correct) cardiac disease that is correctly predicted. It is estimated using Equation 2.

$$\text{Sensitivity of Disambiguation} = \left(\frac{TP}{TP+FN}\right) \text{ X } 100 \qquad (2)$$

Specificity of Cardiac Disease Prediction: It represents the percentage of not-relevant cardiac disease records that is correctly eliminated and recognized as not-relevant. It is estimated using Equation 3.

$$\text{Specificity of Disambiguation} = \left(\frac{TN}{TN+FP}\right) \text{ X } 100 \qquad (3)$$

Accuracy of Cardiac Disease Prediction: It is the percentage of cardiac disease that are correctly predicted and not-relevant cardiac disease records that are correctly eliminated. It represents the ability of cardiac disease prediction; it is estimated using Equations 4.

$$= \left(\frac{TP+TN}{\text{Total Results}} \right) X\ 100 \quad (4)$$

6. Results and Discussion

The k-NN, random forest and logistical regression are tested in 4-folds with UCI dataset and their accuracy as shown in the table 2. The highest accuracy of 0.93 is obtained by logistical regression for split ratio of training and testing is 90:10, it is lowest of 0.81 for k-NN, while it is moderate of 0.84 for random forest. The increasing accuracy of all the models by increasing the training data.

Table 2: Performance Evaluation using Classical Metrics

Regression Technique	Exactness in %	Sensitivity in %	Specificity in %	Accuracy in %
k-NN	81.82	90.00	66.67	81.25
Random Forest	83.93	92.16	68.97	83.75
Logistic Regression	92.86	96.30	84.62	92.50

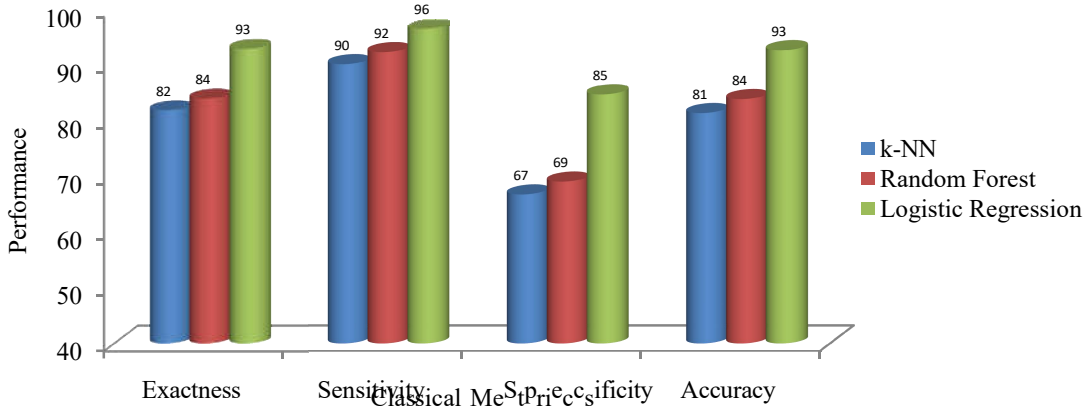

Figure 2: Performance Evaluation using Classical Metrics

Figure 2 shows the performance evaluation of k-NN, random forest and logistic regressions using exactness, sensitivity, specificity and accuracy. From the empirical evaluation it is observed that logistic regressions has highest exactness, sensitivity, specificity and consequently accuracy as compare to k-NN and random forest, it is because logistic regression is a parametric model, in which pre-assumed parameters are exists are used to perform variable screening or feature selection and strongly supports for linear solutions, while k-NN is a parametric in nature and lazy in execution at the time of prediction it use to selects the required features, so k-NN has lower specificity and

consequently has lower in accuracy, as random forest is also non-parametric so it has moderate exactness, sensitivity, specificity and accuracy.

7. Conclusion and Future Work

Predicting cardiovascular illness using the patient's existing data is one of the key areas of the medical profession. Cardiovascular disease can be predicted both as being present and not present.

In this paper, we have used k-NN, random forest and logistic regressions and developed a machine learning model capable of predicting whether or not someone has cardiac disease based on their clinical parameters.All the experimental results and evidences proved the fact that using regression we can predict the cardiac threat and can save one's life. It is possible to increase efficiency and accuracy even more by either increasing the training records, tuning the hyper parameters by randomized search and grid search in modelling the regression (Patil S. , 2023)techniques.

References

1. Hugh, R., & Leavell, E. (1979). The science and art of preventing disease, prolonging life, and promoting physical and mental health and efficiency. *Robert E. Krieger Publishing Company.*
2. Latifah, F. A., & Slamet, I. (2020). Comparison of heart disease classification with logistic regression algorithm and random forest algorithm. *Proceedings of the AIP Conference*, (pp. 2296-2304).
3. (2020). *The top 10 causes of death.*
4. Bashir, S., Khan, Z. S., Khan, H., Anjum, A., & Bashir, K. (2019). Improving heart disease prediction using feature selection approaches. *Proceedings of the 16th International Bhurban3 Conference on Applied Sciences and Technology (IBCAST)* .
5. Bhaise, T., & Patil, S. (2021). Automatic Question Paper Generation System Using Shuffling Algorithm. *International Journal of Innovative Research in Computer and Communication Engineering*, 949-8954.
6. Brzezinski, J. R., & Knafl, G. J. (2019). Logistic regression modeling for context-based classification. *enth International Workshop on Database and Expert Systems Applications. DEXA 99,* (pp. 755-759,). Florence, Italy: IEEE.
7. Hilbe, J. M. (n.d.). *Practical Guide to Logistic Regression.* CRC Press.
8. Isensee, B., & Hanewinkel, R. (2018). *Be Smart - Don't Star.*

9. Jaiswal, J. K., & Samikannu, R. (2017). Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. *2017 World Congress on Computing and Communication Technologies (WCCCT)* (pp. 65-68). Tiruchirappalli, India,: IEEE.

10. Jangale, N., & Patil, S. (2022). Use of Class Dependent Features in K-NN Classifier of Encrypted Data. *International Journal of Scientific Research in Science, Engineering and Technology*, 44-50.

11. Khan, Z., Mishra, D. K., Sharma, V., & Sharma, A. (2020). Empirical study of various classification techniques for heart disease prediction. *Proceedings of the IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, (pp. 57-62).

12. Mokdad , A., Marks, J., Stroup, D., & Gerberding, J. (2004). Actual causes of death in the United States, 2000. *JAMA*.

13. Nishadi, A. (2022). predicting heart diseases in logistic regression of machine learning algorithms by python jupyterlab. *nternational journal of advanced research and publications* .

14. Patil, M., & Patil, S. (2016). Wet and Dry Fingerprint Enhancement by using Multi Resolution Techniques. *2016 International Confence on Global Trends in Signal Processing, Infomation Computing and Communication*, (pp. 188-193).

15. Patil, S. (2023). IMPROVING THE PERFORMANCE OF WET, DRY, FLAKY, ITCHY AND PEELING. *Industrial Engineering Journal, 12*(5).

16. Patil, Y., & Patil, S. S. (2020). Improving Performance of Elastic K-means Clustering using Similarity Measures". *International Journal of Research in Applied Science & Engineering Technology*, 250-255.

17. Ran, J., Zheng, T., & Wang, W. (2018). Logistic Regression Analysis on Learning Behavior and Learning Effect Based on SPOC Data. *2018 13th International Conference on Computer Science & Education (ICCSE)* (pp. 1-5,). Colombo, Sri Lanka: IEEE.

18. Yang, Z., & Li, D. (2019). Application of Logistic Regression with Filter in Data Classification. *2019 Chinese Control Conference (CCC), Guangzhou, China,*, (pp. 3755-3759). China.

19. Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019). Logistic Regression Model Optimization and Case Analysis. *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, (pp. 135-139,). Dalian, China.