STATISTICS AND DATA SCIENCE

EDITORS

Prin. Dr. M. M. Rajmane
Mrs. S. V. Mahajan
Dr. Mrs. S. P. Patil

Bonus Publication, Kolhapur

STATISTICS AND DATA SCIENCE

EDITORS

Prin. Dr. M. M. Rajmane
Mrs. S. V. Mahajan
Dr. Mrs. S. P. Patil

Prarup Publication, Kolhapur

Statistics and Data Science

First Edition

February 2023

Editors –

Principal (Dr.) M. M. Rajmane

Mrs. S. V. Mahajan

Dr. Mrs. S. P. Patil

Publisher –

Prarup Publication, Kolhapur.

ISBN –

Type Setting –

Dr. Mrs. S. P. Patil

Dept. of Statistics, S. G. M. College, Karad.

ISBN - 978-81-956739-9-5

STUDY OF CLASSIFICATION TECHNIQUES (LOGISTIC REGRESSION, SUPPORT VECTOR MACHINE AND LINEAR DISCRIMINANT ANALYSIS) IN PREDICTION OF PREVALENCE OF HEART DISEASE

[1]R. D. Patil , [2]Dr.R.R. Kumbhar, [3]Dr. S.V. Kakade

[1]Assistant Professor, S.G.M. College Karad,

[2]Principal, Vivekananda college Kolhapur,

[3]Professor, Krishna Institute of Medical Science Demmed to be an University karad

Email: [1]patilrd1996@gmail.com ,

[2]rrkumbhar@yahoo.co.in ,

[3]satishvkakade@yahoo.co.in

Abstract:

In this study we have used the publicly available Cleveland Dataset and compare it through various classification techniques such as Logistic regression, Support vector machine, Linear Discriminant analysis for prediction of the prevalence of heart diseases. From these different models we evaluated their accuracies in predicting a heart disease. We have claimed that logistic regression and Linear Discriminant Analysis give more accurate results than support vector machines for predicting heart disease. We also used F1 score, AUC curves, precision and recall as evaluative measures. Here our aim is to provide a benchmark and improve earlier ones in the field of heart disease diagnostics with the help of different classification techniques.

Key words: Heart disease, Logistic regression(LR),Linear Discriminant Analysis(LDA),classification, Cleveland Heart Disease.

Introduction:

In this study we focused on heart disease as it leads to death. Heart disease is very complex to determine due to many problems of health such as cholesterol, chest pain, and blood pressure. Here we have used the heart disease dataset of University of California (UCI). Various investigations of prediction are made by many researches on this dataset. Through this we make comparative study of Logistic regression, Support Vector Machine and Linear discriminant analysis classifiers in the case of classification of heart disease.

ISBN - 978-81-956739-9-5

Materials and methods:

Dataset Details:

Here we have used the Cleveland Heart Disease Dataset from the UCI repository in our study which is publicly available. The data consisting of 303 instances with only 14 attributes.

The Cleveland dataset has 14 attributes as follows:

Data types with values are as follows:

- Age, in years
- Sex: female , male
- Chest Pain type (a) typical angina (angina), (b) atypical angina (abnang),
  (c) non-anginal pain (notang),
  (d) asymptomatic (asympt).
  These are denoted by numbers 1 to 4
- Trestbps: Patient's resting blood pressure in mm Hg at the time of admission to the hospital
- Chol: serum cholesterol in mg/dl
- Fbs: Boolean measure indicating whether fasting blood sugar is greater than 120 mg/dl: (1 = True; 0 = false)
- Restecg: electrocardiographic results during rest
- Thalach: maximum heart rate achieved
- Exang: Booelan measure indicating whether exercise inducing angina has occurred
- Oldpeak: ST depression induced by exercise relative to rest
- Slope: the slope of the ST segment for peak exercise
- Ca: number of major vessels (0 - 3) colour by fluoroscopy
- Thal: the heart status (normal, fixed defect, reversible defect)
- The class attributes: value is either healthy or heart disease (sick type: 1, 2, 3, and 4).

For our research purposes, we convert class attribute as binary by indicating a heart disease by 1 and healthy by 0.

Hence in this research problem, problem of the multi-class classification is converted to binary classification problem. In this study, after selection of parameters through cross-validation accuracy was tested on the test data. This is done for keeping a sufficient amount of data from biasing the models and thus giving a completely fresh perspective for testing.

ISBN - 978-81-956739-9-5

Since, by using cross-validation each observation in data have an equal chance of training also once the chance of testing.

Results & Discussion:

Table No.1.Performance evaluation of logistic regression by using 5 fold cross-validation.

	Accuracy	AUC	Recall	Prec.	F1
0	0.8605	0.8882	0.8750	0.8750	0.8750
1	0.9070	0.9761	0.8696	0.9524	0.9091
2	0.8333	0.9085	0.8261	0.8636	0.8444
3	0.8095	0.8719	0.7826	0.8571	0.8182
4	0.8571	0.9245	0.9130	0.8400	0.8750
Mean	0.8535	0.9138	0.8533	0.8776	0.8643
SD	0.0324	0.0359	0.0448	0.0391	0.0308

Table No.2. Performance evaluation of Support Vector Machine by using 5 folds cross-validation

	Accuracy	AUC	Recall	Prec.	F1
0	0.5581	0.5965	1.0000	0.5581	0.7164
1	0.5814	0.6022	0.9200	0.5897	0.7188
2	0.5714	0.6366	0.9583	0.5750	0.7188
3	0.5714	0.5718	1.0000	0.5714	0.7273
4	0.6190	0.5660	1.0000	0.6000	0.7500
Mean	0.5803	0.5946	0.9757	0.5789	0.7262
SD	0.0207	0.0252	0.0322	0.0146	0.0124

ISBN - 978-81-956739-9-5

Table No.3.Performance evaluation of linear discriminant(LDA) analysis by using 5 fold cross-validation.

	0.8605	0.9057	0.9167	0.8462	0.8800
	0.8837	0.9696	0.8696	0.9091	0.8889
	0.8571	0.8902	0.8696	0.8696	0.8696
	0.7857	0.8741	0.8261	0.7917	0.8085
	0.8571	0.9016	0.9130	0.8400	0.8750
	0.8488	0.9082	0.8790	0.8513	0.8644
	0.0331	0.0326	0.0333	0.0384	0.0287

The mean accuracy for logistic regression is 85.35%, for SVM is 58.03%, and for LDA is 84.88%. Also Area under curve (AUC) for LR, SVM, LDA are 92.45%, 0.5946%, 90.82% respectively.

Conclusion:

As results from the analysis done by us, Logistic Regression and the Linear Discriminant Analysis approach has higher accuracy as well AUC is greater than 0.90. So, Logistic Regression and Linear Discriminant Analysis models have higher Area under curve (AUC) so they have, higher discriminating power of classification while the Support Vector Machine (SVM) has low accuracy. Also SVM model has low Area under curve (AUC) so; it has lower discriminating power than Logistic Regression and Linear Discriminant Analysis. Hence they give better performance than SVM for Prediction of Heart disease.

ISBN - 978-81-956739-9-5

References:

1. Harshit Jindal et al (2021) Heart disease prediction using machine learning algorithms. IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072.
2. Boser, B.E., Guyon, I.M. & Vapnik, V.N.(1992): "A training algorithm for optimal margin classifiers"; *ACM Press, Pittsburgh, PA, pp144–152.*
3. Chan Y.H.(2005): "Discriminant analysis"; *Singapore Medical Journal 46:54-62*
4. Metehan Makinacı(2007): "Support Vector Machine Approach for Classification of Cancerous Prostate Regions"; international journal Medical sciences,Vol.1(7)pp470-473
5. Maja Pohar, Mateja Blas, and Sandra Turk(2004): "Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study"; *Metodološki zvezki.;Vol.1(1)pp 143- 161*
6. Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande(2015): Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease International Journal of Machine Learning and Computing, Vol. 5
7. Klaus D Werneck .(1994) On the Application of Discriminant Analysis in Medical
8. Diagnosis.Springer,pp.267-279