Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

On Predicting the Survival of Cancer Patients Using Statistical and Machine Learning Algorithms

R. Jaisankar[1], D. Victorseelan[2]
[1] Professor, Department of Statistics,
Bharathiar University, Coimbatore, India
r_jaisankar@rediffmail.com

[2]Research Scholar, Department of Statistics,
Bharathiar University, Coimbatore, India
victorseelan@gmail.com

Article Info

Page Number: 1007 - 1021

Publication Issue:

Vol 71 No. 4 (2022)

Article History

***Article Received**: 25 March 2022*

***Revised**: 30 April 2022*

***Accepted**: 15 June 2022*

***Publication**: 19 August 2022*

Abstract

Machine learning algorithms and Statistical Methods are used potentially nowadays in various biological problems, in particular, predicting patients' survival from their data on various parameters.However, the choice between the selection of statistical procedures or Machine learning procedures depends on various factors like the nature of the problem, main objectives and the data frame etc. The present work is aimed to compare statistical and machine learning methods with reference to regression for prediction. The data taken is regarding Heart failure among cardiac patients. The methods of comparison include the Logistic Regression (LR) and the Support Vector machine (SVM).

Methods: In this paper, the analyses have been performed based on a dataset of 299 heart failure patients with the problem of heart failure collected in 2019. Statistical prediction Methods and Machine learning classifiers are applied both to predict the patient's survival and identify rank the features corresponding to the most important risk factors. The results obtained by applying the Statistical prediction Methods and machine learning algorithms are compared in terms of their accuracy.

Keywords: Logistic Regression, Machine learning, Support Vector Machine (SVM), Random Forest Classification.

Introduction

In Cardiovascular diseases, Heart failure (HF) is a condition that occurs when the heart cannot pump enough blood to meet the needs of the body which may be fatal. As per the records of Heart disease and stroke statistics, USA, nearly 18.6 million people died of cardiovascular disease around the world in 2019.

Medical records of HF patients quantify the symptoms, body features, and clinical laboratory test values and other parameters of life style which are used for diagnosis and also to perform bio-statistical analyses aimed at highlighting patterns, correlations and identifying associated risk factors in general.The prediction of events in cardiovascular diseases is still critical for medical industry because it allows them to develop strategic treatment/awareness programs that will help to decrease the number affected. In such contexts, it is possible to apply either statistical or machine learning procedures for prediction. Both of these approaches may have some specific special features and as such one may compare them with reference to their prediction accuracy.The present work aims on an experimental comparison study of the statistical procedure Logistic regression and the machine Learning Algorithm SVM for predicting the Heart failure.

A sample of 299 patients with the cardiovascular problems iscollected in 2019from different hospitals in overall TamilNadu, India. Statistical prediction Methods and Machine learning classifiers are applied both to predict the patient’s survival and to identify rank the features corresponding to the most important risk factors. The power of the models is measured by correct prediction rate. In the current scenario, it seems that machine learningmethods are taking advantage over the Statistical Methods, but however in reality it is not always true. Some statistical concepts are inherited by the machine learning methods make them potential.A brief description of Logistic Regression Analysis andSupport Vector Machine that were used for the analysis of this study is given below.

Literature Review

The application of Machine Learning(ML) algorithms for solving problems in medical and clinical studies is quite old since from 2010. Many studies have been carried out not only using ML algorithms but also to make comparisons between Machine Learning algorithms and statistical procedures for fulfilling the specific objectives like the identification of factors, prediction and classification, clustering and decision making etc.

Oyewola et al. [2017] have made such a comparative study on Logistic regression (LR), linear discriminant analysis (LDA) with Machine Learning algorithms like Random forest (RF), Support vector machine (SVM) classificationsand quadratic discriminant analysis (QDA), for making breast cancer biopsy predictions with a mammographic diagnosis.

Bernal et al. (2017) adopted Statistical and Machine learning techniques such as logistic regression, neural networks and decision trees to predict the disease of the patients and shown that the parameter configuration plays a fundamental role in the models performance.

Agarap (2018) made investigations on the Wisconsin Diagnostic Breast Cancer and compared Machine learning (ML) algorithms like Multilayer Perceptron (MLP), GRU-SVM Nearest Neighbor (NN) search, Softmax Regression, and Support Vector Machine (SVM) by measuring their classification test accuracy, and their sensitivity and specificity values and shown that the MLP algorithm stands out among the implemented algorithms with a test accuracy of nearly 99.04%.Similar type of study has been carried out by Westerdijk [2018] for the prediction of breast cancercells with an ensemble models.

Puja Gupta and Shruti Garg [2020]have performed a study for predicting breast cancer. They have used various machine learning tools with proper hyper parametric change and have shown that such a modification would bring most accurate results with minimum loss. Deep learning using Adam Gradient Descent Learning is used for achieving the accuracy.

A Comparison of various machine learning models used for Coronary heart Disease Prediction was done by Sunil Kr. Tiwari and Suresh Kumar Garg [2021]. It is observed that k-NN method is a better prediction algorithm with better accuracy and precision value than Support vector Model, Decision Tree, Random forest and Logistic Regression models. A brief review of the methods taken for comparison is given below.

Logistic Regression

Logistic Regression is a specialized regression method developed when the response variable is categorical. In this method in order to make continuity of the response variable, logit transformation is made on it.Logistic Regression (LR) prevails as the most important statistical and datamining techniques employed by statisticians and researchers for the analysis andclassification of binary and proportional response data sets.Logistic Regression can naturally provide probabilities andextend to multi-class classification problems. The general form of the logistic regression equation is described below.

Suppose that there is a categorical dependent variable Y having G distinct values is to be regressed on a set of p independent variables, say, $X_1, X_2, \dots, X_p$. Let the values of Y be$1,2, \dots, G$.

Then the logistic regression model is given by the following *G* equations.

$$In\left(\frac{p_g}{p_1}\right) = In\left(\frac{p_g}{p_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p$$

Here, p_g is the probability that an individual with values $X_1, X_2, \dots, X_p$ is in event g. That is,

$$p_g = Pr(Y = g \mid X)$$

Usually X_1 is taken to be equal to 1 for including the intercept though it is not necessary. $p_1, p_2, \dots, p_g$represent the prior probabilities of event membership. If these are assumed to be equal, then $In\left(\frac{p_g}{p_1}\right)$ would be equal to zero and hence dropped. If the priors are not assumed to be equal, then therewould be a change the values of the intercepts in the logistic regression equation. The event one is called the reference value. The regression coefficients$\beta_{11}, \beta_{12}, \dots, \beta_{1p}$for this reference value are set to zero. One can arbitrarily choose this reference value. Usually, it is taken as the most frequent value or a control value to which the other events are to be compared. This makes *G*-1 equations in the logistic regression model. Then the β's are to be estimated from the observed data. Komarek and Moore [2004] were the first to shown thatthe Logistic Regression is highly potential to classify large data sets, and that it can outperform the Support Vector Machines (SVM),which is considered a state-of-the-art algorithm.

Support Vector Machine (SVM)- Radial Kernel

The Support Vector Machine (SVM) is one of the Machine Learning algorithmsfirstproposed by Vapnik[1995] and has since attracted a highdegree of interest by the research, originally designed for binary classification, are large margin classifiersthat try to separate instances of different classes with the maximum margin hyperplane. The margin is defined as the minimum distance from instances of different classes to the classification hyperplane.

Support Vector Machine (SVM), a machine learning algorithm,which is categorized under supervised learning method,is used for both classification and regression and it is best suited for classification.

Take a linear classifier$Y = sign(D^T X + B)$, the hinge loss can be used to evaluate the fitness to the data:

$$\sum_{ii=1}^{m} max\{0, 1 - Y_i(D^T X + B)\}.$$

Then,

$$\frac{|D^T X + Bb|}{\|D\|}$$

is the Euclidean distance from an instance, sayX_i to the hyperplane $D^T X + B$.

If $|D^T X_i + B|$ is assumed to be grater than or equal to 1 for all instances, the minimum distance to the hyperplane is $\| D \|^{-1}$. Hence, the problem of SVM is to maximize$\| D \|^{-1}$. That is to optimize,

$$(D^*, B^*) = arg\,min \frac{\|D\|^2}{2} + C \sum_{i=1}^{m} \varepsilon_i$$

Subject to the constraints

$$Y_i(D^T X_i + B) \geq 1 - \varepsilon_i \ (\forall i = 1,2, \ldots., m)$$

$$\varepsilon_i \geq 0 \ (\forall i = 1,2, \ldots., m),$$

where C is a parameter and ε_i's are slack variables introduced to enable the learner to deal with data that could not be perfectly separated, such as data with noise.

The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

Methodology

The total number of records collected was from 299 heart failure patients. At first, the dataset is checked to clean to make it free from noise, outliers, missing values and redundancy using statistical techniques like box plot. The main objective of this study to predict the patient’s survival and identify rank the features corresponding to the most important risk factors with a help of Statistical prediction Methods and Machine learning classifiers and compared in terms of their accuracy.

The prediction results of both Logistic Regression and SVM methods were based on the same datasets that were used for verification and calibration. This provides objectivity by comparing only the performance of each method. The success rate of classification is determined by the ratio of correctly classified recordings to the total number of recordings in that set. The factors taken in this study which are expected to have influence on heart failure are as under:

Age (years)(X_1), Sex(X_2), Smoking status(X_3), Time-follow-up period (days) (X_4), Anemia(X_5), High blood pressure(X_6), Creatinine phosphokinase(mcg/L) (X_7),Diabetes(X_8), Ejection fraction of blood (X_9),Platelets (kiloplatelets /mL) (X_{10}),Serum creatinine (mg/dL) (X_{11}), Serum sodium (mEq /L) (X_{12}), and the event of death(Y).

Checking for the Absence of multicollinearity:

In order to get a better regression model and estimates, the multicollinearity that may exist in the data has to be ruled out. A correlation plot is used to identify the multicollinearity, which is presented below. It is clearly seen that none of the independent variables has highly correlated with others and hence it is found that there is no evidence of multicollinearity.

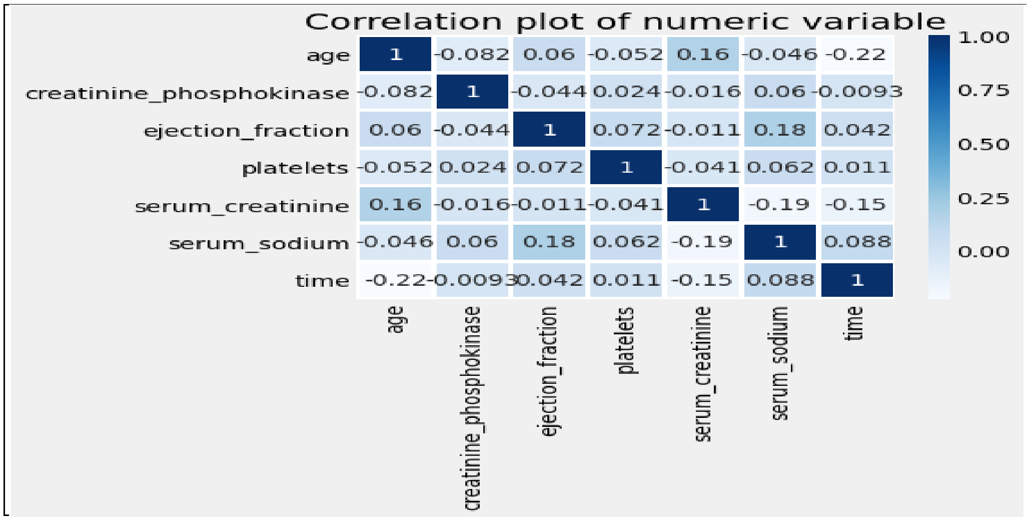
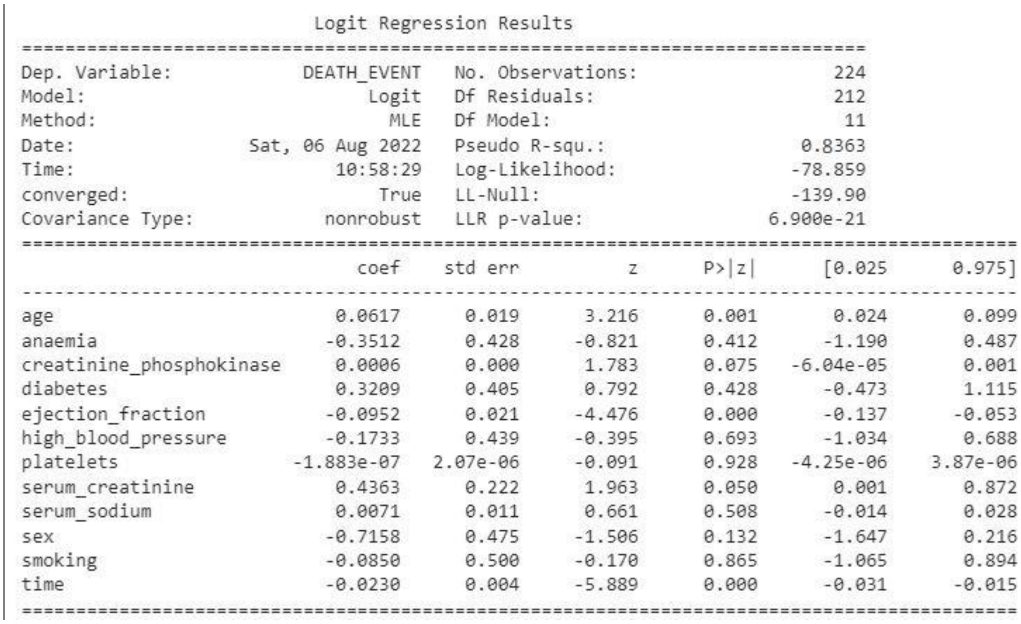


Figure (1): Correlation Plot

Statistical prediction Method - Logistic Regression

The Figure (2) of interest is the model summary. This table provides the R^2 value and p value, which can be used to determine how well a regression model fits the data, which is the proportion of variance in the dependent variable that can be explained bythe independent variables. The R^2 value of 0.8363indicates that our predictors explain 83.6% of the variability of our dependent variable (death event).



```
                           Logit Regression Results
==============================================================================
Dep. Variable:            DEATH_EVENT   No. Observations:                  224
Model:                          Logit   Df Residuals:                      212
Method:                           MLE   Df Model:                           11
Date:                Sat, 06 Aug 2022   Pseudo R-squ.:                  0.8363
Time:                        10:58:29   Log-Likelihood:                -78.859
converged:                       True   LL-Null:                       -139.90
Covariance Type:            nonrobust   LLR p-value:                 6.900e-21
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
age                          0.0617      0.019      3.216      0.001       0.024       0.099
anaemia                     -0.3512      0.428     -0.821      0.412      -1.190       0.487
creatinine_phosphokinase     0.0006      0.000      1.783      0.075   -6.04e-05       0.001
diabetes                     0.3209      0.405      0.792      0.428      -0.473       1.115
ejection_fraction           -0.0952      0.021     -4.476      0.000      -0.137      -0.053
high_blood_pressure         -0.1733      0.439     -0.395      0.693      -1.034       0.688
platelets                -1.883e-07   2.07e-06     -0.091      0.928   -4.25e-06    3.87e-06
serum_creatinine             0.4363      0.222      1.963      0.050       0.001       0.872
serum_sodium                 0.0071      0.011      0.661      0.508      -0.014       0.028
sex                         -0.7158      0.475     -1.506      0.132      -1.647       0.216
smoking                     -0.0850      0.500     -0.170      0.865      -1.065       0.894
time                        -0.0230      0.004     -5.889      0.000      -0.031      -0.015
============================================================================================
```

Figure (2): Model Summary of Logit Regression

The fitted Logistic Regression Line is:

Y (odds ratio of death event) = Age (0.0617) – Sex(0.7158) –Smoking Status(0.0850)- Time(0.0230) - Anemia (0.3512) - High blood pressure(0.1733) + Creatinine phosphokinase (0.0006) + Diabetes(0.3209) - Ejection fraction (0.0952) – Platelets(0.000) + Serum creatinine(0.4363) - Serum sodium(0.0071)

It is seen that the variables Age, Creatinine phosphokinase and Ejection fraction, are significantly influencing the event whereasSex, time, smoking, anemia, diabetes, high blood pressure, platelets, serum creatinine and serum sodium are not significant.

The adequacy of the Logistic Regression model and its fitness are tested using the following measures.

AUC (Area under Curve)

The Receiver Operating Characteristic curves (ROC) in logistic regression are frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests. ROC curve is a plot of sensitivity against one minus specificity as the value of cut point is increased from 0 to1 is used for determining the best cutoff value for predicting whether a new observation is a "non-occurrence" (0) or "occurrence" (1). AUC stands for "Area under the ROC Curve." That is, and measures how well a model is able to distinguish between the classes. ROC value for our Logistic regression analysis is 0.861 which shows a higher prediction probability for classifying a point belong to the positive class.

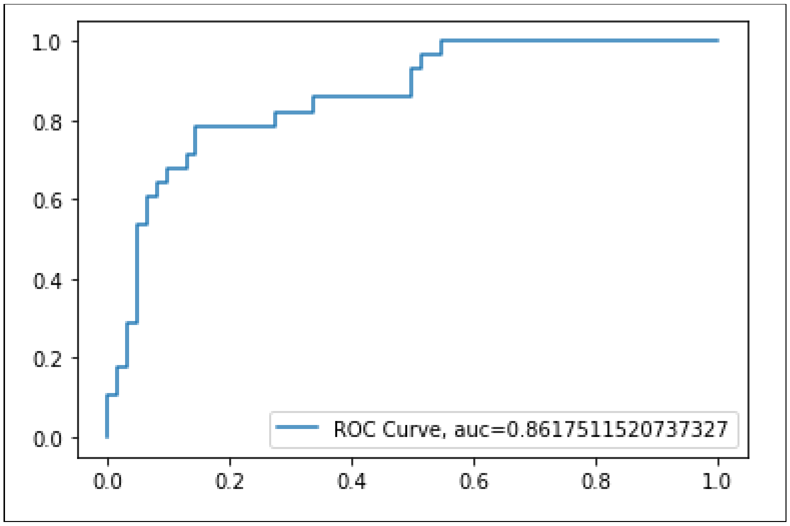


Figure (3): ROC Curve

Classification Accuracy (CA)

Accuracy is one metric for evaluating classification models is the number of correct predictions to the total number of input samples. That is,

$$Accuracy = Number of Correct Predictions \,/\, Total Number of Predictions$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$Accuracy = (TP+TN) \,/\, (TP+TN+FP+FN)$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

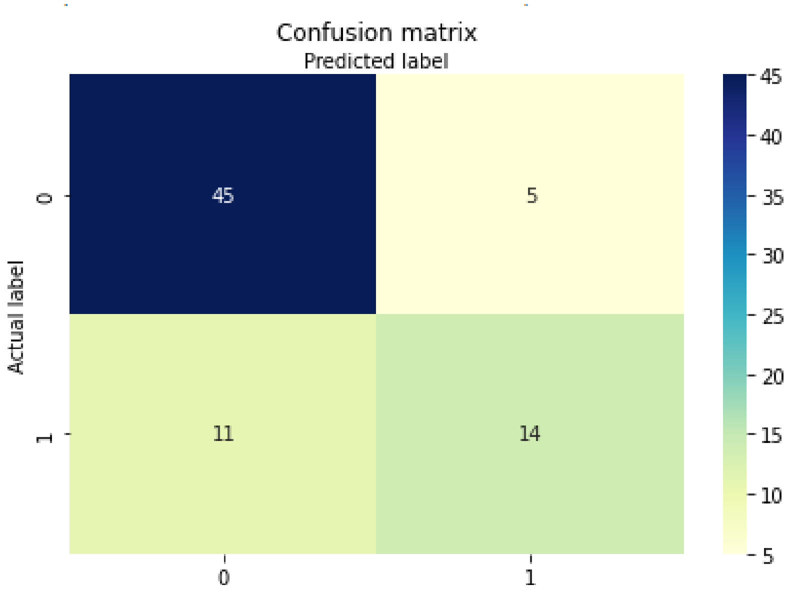
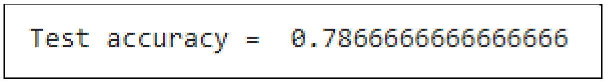


Figure (4): ROC Curve

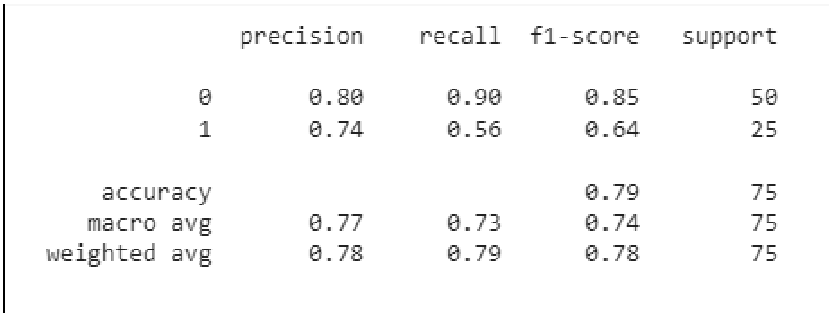


```
Test accuracy =  0.7866666666666666
```

The Accuracy Value of the fitted Logistic Regression equation is0.786, which means that the total number of predictions that the model gets right is about 78%.

Precision

The Precision measure is the fraction of correct classifications within all elements classified as such and is calculated by: $TruePositive$ / ($TruePositive$ +$FalsePositive$). The observed Precision value is 0.77 which means that 77% of positive identifications are actually correct.



```
              precision    recall  f1-score   support

           0       0.80      0.90      0.85        50
           1       0.74      0.56      0.64        25

    accuracy                           0.79        75
   macro avg       0.77      0.73      0.74        75
weighted avg       0.78      0.79      0.78        75
```

Figure (5): Accuracy Measures

Recall : Recall is a measure of the classifier's completeness; the ability of a classifier to correctly find all positive instances. For each class, it is defined as the ratio of true positives to the sum of true positives and false negatives.

Recall Measure is computed by using the formula:

$$TruePositive/(TruePositive + FalseNegative\ )$$

Precision and Recall measure accuracy value closer to 1 is the best fitted model and closer to 0 is not a good fitted model. The observed value of the Recall measure is 0.73.

F1 Score

The F1 score is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify samples into 'positive' or 'negative'. That is, a good F1 score represents lower levels of both false positives and false negatives. The F1 score is the harmonic mean of Precision and Recall. An F1 score nearer to 1 indicates the model is nearly perfect.

The F1 score is calculated by using the following formula:

$$F1=2*((Precision*Recall)/(Precision+Recall))$$
$$=TP/(TP+1/2\ *(FP+FN))$$

The observed value of the F1 score value of the present case is0.74.

Machine Learning Survival prediction classifiers - SVM - Linear Kernel

This part of the analysis focuses on the binary prediction of the survival of the patients in the follow-up period, using SVM – linear kernel model. The dataset was split into 60% (179 randomly (selected patients) for the training set, 20% (60 randomlyselected patients) for the validation set, and 20% (theremaining 60 patients) for the test set. The model has been fitted using Python coding and the adequacy of the SVM - Linear Kernel model and its fitness are tested using the following measures.

Confusion Matrix and Classification Report Table:

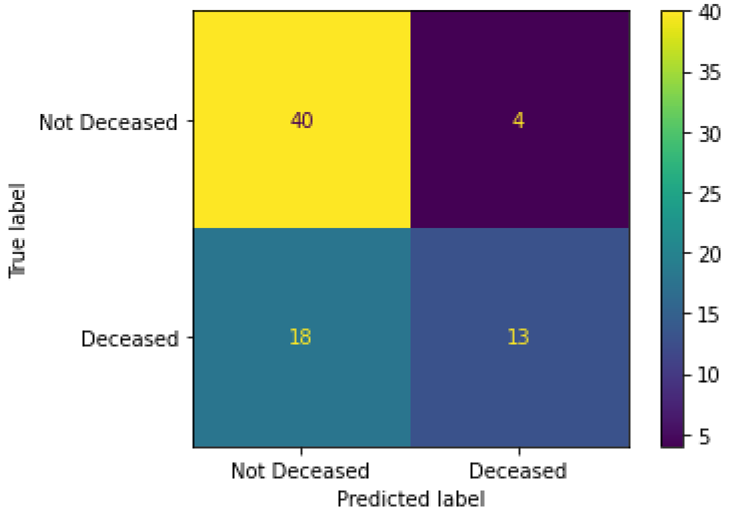


Figure (6): Confusion Matrix

AUC Value:

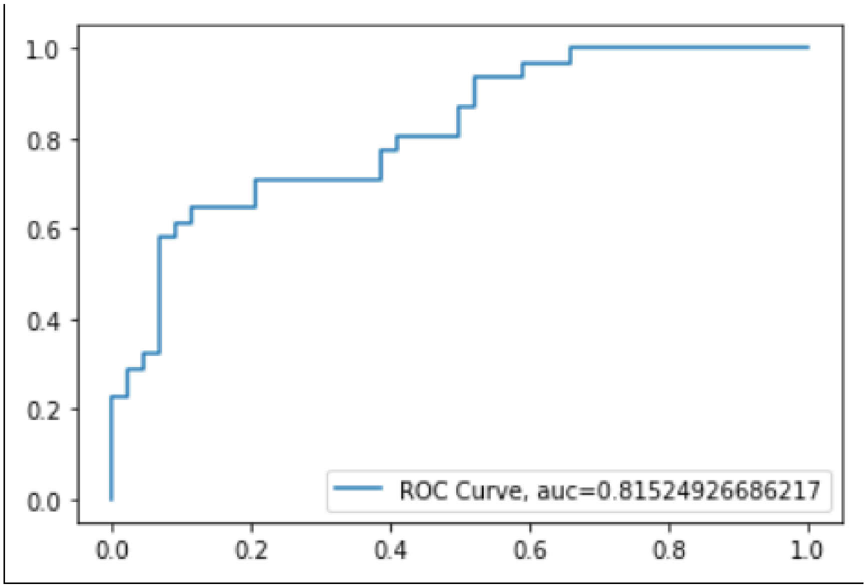


Figure (7): AUC Curve

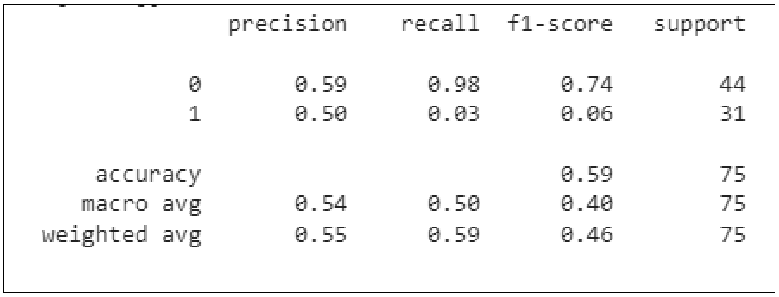
The Fig (7) showed that, obtaining ROC AUC of SVM is 0.815

$$Accuracy=TP+TN\ /\ TP+TN+FP+FN$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

The accuracy level of SVM is 0.815.

Classification Accuracy (CA), F1 Score, Precision and recall value:



```
              precision    recall  f1-score   support

           0       0.59      0.98      0.74        44
           1       0.50      0.03      0.06        31

    accuracy                           0.59        75
   macro avg       0.54      0.50      0.40        75
weighted avg       0.55      0.59      0.46        75
```

Figure (8): Classification Accuracy

From Fig (8), it found that the mean value of F1 score is 0.40, Precision value is 0.54and recall value is 0.50.

Comparison: Logistics Regression vsSVM - Linear Kernel

The target of this research work is to study the effectiveness of Logistic Regression and Support Vector Machine - Radial Kernelin prediction of Heart failure.To compare the performance of the Machine Learning Algorithm with the Logistic Regression approach, the correct prediction rate of predictive accuracies in the Machine Learning Algorithm with the Logistic Regression models arelisted below.

Model	Correct prediction Rate %
Logistic Regression	79%
SVM - Linear Kernel	59%

Figure (9): Comparison Table

It is found that the predictive accuracy rate of Logistic Regression is 79% and the prediction rate of Support Vector Machine - Radial Kernel is 59%. Clearly, the Logistic Regression demonstrates a superior ability to predict the patients affected by cardiovascular disease - Heart failure.It is seen that Logistic regression having high accuracy as well as highest area under curve with acceptable precision and recall when comparing with SVM Method.

Performance indices	AUC	CA	F1 Score	Precision	Recall
LR	**0.861**	**0.786**	**0.74**	**0.77**	**0.73**
SVM	**0.815**	**0.59**	**0.40**	**0.54**	**0.50**

Figure (10): Performance indices

TheFig (10) compares the Logistic regression and SVM model, all the measures of LR are closer to 1compared to SVM and these models are best fitted models using LR method and SVM algorithm.

Conclusions:

This discovery has the potential to impact on clinical practice, becoming a new supporting tool for physicians when predicting if patient will get heart failure or not. Compared with the SVM algorithm, Logistic Regression modelhas proven to be more accurate in Prediction of survival of patients and had higher overall performance indices.

Further studies of this model may be considered with the effect ofa more detailed databases that includes complications and clinicalexamination findings that will evolve into an effective adjunctiveclinical decision-making tool.

Limitation:

As a limitation of the present study, only a lesser size of the dataset (299 patients) is used.A larger dataset may yield more reliable results. Additional information about the physical features of the patients (height, weight, body mass index, etc.) and their occupational history would have been useful to detect additional risk factors for cardiovascular health diseases.

References:

[01] Bernal, J.L.; Cummins, S.; Gasparrini, A. Interrupted time series regression for the evaluation of public healthinterventions: A tutorial. Int. J. Epidemiol. 2017, 46, 348–355.

[02]Oyewola, D.; Hakimi, D.; Adeboye, K.; Shehu, M. Using five machine learning for breast cancer biopsypredictions based on mammographic diagnosis. Int. J. Eng. Technol. IJET 2017, 2, 142–145. [CrossRef]

[03] Agarap, A.F.M. On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset. In Proceedings of the 2nd International Conference on Machine Learning and SoftComputing, Phuoc Island, Vietnam, 2–4 February 2018; pp. 5–9.

[04]Westerdijk, L. Predicting Malignant Tumor Cells in Breasts; Master Business Analytics; VrijeUniversiteitAmsterdam: Amsterdam, The Netherlands, 2018.

[05] A. Agresti. An Introduction to Categorical Data Analysis.Wiley-Interscience, 2007.

[06] T. Hastie, R. Tibshirani, and J. Friedman.The Elements of Statistical Learning.Springer Verlag, 2 edition, 2009.

[07] J. M. Hilbe. Logistic Regression Models.Chapman & Hall/CRC, 2009.

[08] D. G. Kleinbaum, L. L. Kupper, A. Nizam, and K. E. Muller. Applied RegressionAnalysis and Multivariable Methods. Duxbury Press, 4 edition, 2007.

[09]P. Karsmakers, K. Pelckmans, and J. A. K. Suykens. Multi-class kernel logisticregression: a fixed-size implementation. International Joint Conference on NeuralNetworks, pages 1756–1761, 2007.

[10] D. W. Hosmer and S. Lemeshow. Applied Logistic Regression.Wiley, second edition,2000.

[11] P. Komarek. Logistic Regression for Data Mining and High-DimensionalClassification.PhD thesis, Carnegie Mellon University, 2004.

[12] V. Vapnik. The Nature of Statistical Learning.Springer, NY, 1995.

[13] Maher Maalouf. Logistic regression in data analysis: An overview, International Journal of Data Analysis Techniques and Strategies 3(3):281-299, July 2011.

[14] Yoonkyung Lee. Support Vector Machines for Classification: A Statistical Portrait, Methods in molecular biology (Clifton, N.J.) 620:347-68. January 2010

[15] World Health Organization, World Heart Day. https://www.who.int/cardiovascular_diseases/world-heart-day/en/. Accessed 7 May 2019.

[16] P. Arumugam,Prediction, Cross Validation and Classification in the Presence COVID-19 of Indian States and Union Territories using Machine Learning Algorithms, International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume 10 issue 1, May 2021.

[17] Moguerza, J. & Muñoz, A. (2006), Vector machines with applications, Statistical Science 21(3), 322–336.

[18] Tibshirani, R. & Friedman, J. (2008), The Elements of Statistical Learning, Springer, California.

[19] Verplancke, T., Van Looy, S., Benoit, D., Vansteelandt, S., Depuydt, P., De Turck,F. and Decruyenaere, J. (2008), 'Support vector machine versus logistic regressionmodeling for prediction of hospital mortality in critically ill patients withhaematological malignancies', BMC Medical Informatics and Decision Making8, 56–64.

[20] K.S. Keerthika& S. Parthiban. (2021), Nat. Volatiles & Essent. Oils, 8(5), 3641-3649.