International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Supervised Linear Estimator Modeling (SLEMH) for Health Monitoring

Amandeep Kaur, Anuj Kumar Gupta

Abstract: ***In this research work, the E-Health monitoring system has been developed using fifteen health indicators. These fifteen features were selected by following a Recursive Feature Elimination with Cross-Validation method. The dataset was labeled as per medical limits and segregated into three classes (normal, borderline and onset of unhealthy state). A rigorous process was followed at each step to find out which linear estimator and model is suitable for classifying health condition of persons. Five regression estimators were evaluated and it was found that logistic regression and linear discriminant analysis methods are providing highest accuracy and lowest error for classifying three health states of a patient.***

Keywords: E-Health

I. INTRODUCTION

Medical sensors [1] are part of so-called IoT revolution [2] [3]. Outpatient care and remote health monitoring are being impacted by the innovations happening in the field of medical sensors. The confluence of multiple technologies [1] and logical reasoning of the industry to bring the healthcare cost down [4] is driving the need to develop new kinds of frameworks, algorithms, and strategies to analysis the health care results. The biggest benefit of medical sensors is that it is helping in exponential growth of preventive medicine [5]. This is because it is far more, easier to detect the onset of a health problem now as compared to the previous era [6]. Wearable biosensors [7], physiological sensors [8] are making easy to collect and process health data , especially data related to the life style diseases [9]. In fact, mankind has come long way from non-pharmacological treatments [10] to the current use of artificial intelligence [11] in detecting complex medical conditions .Now, it is possible to characterize the health issue using sensor data and take appropriate action in advance . The process of monitoring health [12] has become simple and easy to implement due to innovation in statistics and hardware. of The current algorithms that help in detecting complex health issues are undergoing lot of changes but, the fundamentally the mathematics of most of the detection [13] and diagnostic algorithms is based on regression. With time new medical modalities are been discovered and there is a need to do next iteration of research work. In this research work, an exploration of the regression methods that can act as classifiers is done.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Amandeep Kaur*, Research Scholar ,Deptt of Computer Science & Engg Ikgptu Kapurthala, Punjab. Akdhaliwal361@gmail.com
Anuj Kumar Gupta,Professor,Deptt of Computer Science & Engg, CGC, LANDRAN, PUNJAB.Anuj21@hotmail.com

It is expected that a suitable model that can detect onset of the life style health issues will be constructed with high accuracy and low error rate.

II. REVIEW

The current editions of the journals show an underlying trend of discussions of researchers on the use of regression and classification methods. This is because, to solve real world problems, understanding of the nature of the task is the key for selecting the right algorithm to use. Many authors [14] put regression and supervised learning concepts under one category. This is simply because of these concepts use similar methods of dealing with the dataset. Both these methods use the concept of training data to make predictions. In health care, regression analysis has been done to find mathematical relationships between the various factors based on which a health problem can be identified. It has been used as tool to find the root cause of epidemics [15], diseases, and other health related conditions. The regression methods can help to find the strength of impact of multiple variables on the growth of a medical condition. Linear regression [16] is most widely used method in health care industry for conducting visual analysis of the health parameters [14]. In effort to overcome the issues with the linear regression methods such as sensitivity to the outliers and issues related to collinearity etc., Many authors [17] [18] take help of methods such as step up regression (forward) or step down (backward) selection of variables that take part in the regression and prediction process. In many medical cases, many health parameters are not fully observable or are partially observable. In such cases, the regression methods such as logistic regression may be useful. Typically, logistic regression is useful in detecting or predicting events that are binary in nature such as “healthy “or “unhealthy “. But, this method can also be successfully be applied multiple class problems as well with good level of accuracy. To solve the problem of multi-collinearity, ridge regression [19] and lasso (least absolute shrinkage and selection operator) estimators have been also been used to solve the medical problems. The health parameters such as weigh and basic metabolic rate (BMR) are normally highly correlated to each other. Such pairs of variables do not add more value to the estimator but, in fact increase the overall load in running the algorithm. The current survey of the other methods shows that researchers are now focusing on statistical learning models. Models that learn from the data patterns [20] and execute tasks . Some of these methods are better known by term machine learning models. Most of these models are based on the fundamentals of regression and probability.

The medical industry [13] is extensively using these methods and some researchers under value the power of regression methods that can be used for regression as well for classification jobs. Most of these researchers focus on constructing models that work on categorical data rather than on continuous dataset series [21]. In many cases, it has been found that there is a need to do binning or process by which a continuous time series dataset is required to converted into categorical dataset. For example, if there is a need to identify a group of patients that are prone to some disease. The proneness of the disease cannot be measured in absolute terms, some may be less prone and there might be others who are more prone a particular disease. The degree of proneness is required to be defined and the dataset is required to undergo transformation from continuous dataset into categorical dataset. For , this process many researchers are using unsupervised methods such Kmeans [22], c-fuzzy clustering [23] methods and some are following "if then else" rules to segregate the dataset . In the field of supervised learning models, linear classifiers [14] such as perceptron learning models , Linear discriminant model, logistic regression [24] , lasso regression [25] [26] and probability models such as Naïve Bayes [27] are widely used in medical classification tasks . Methods such as Neural networks [28] , Support vector machine [29], etc., are popular when the classification problems are typically multi-class and the nature of dataset is non-linear . The logic of finding appropriate input variables or predictors with respect to a label is usually done with help of methods that makes the dataset orthogonal and linearly separable. Regression [30] remains the core method of doing the causal analysis (a process that determines what will be input and output of the machine learning algorithms) and stochastic gradient descent (SGD) [31] ,[32] and linear SVM are in also in use for solving medical and health care classification problems.

A. *Problem Formulation*

Health monitoring using sensor data requires the construction of statistical models that can accurately predict the trends in health dataset. The statistical models based on regression can accurately predict the trend between the dependent and predictor variables. For example,

OLS Regression Results

Dep. Variable:	VF	R-squared:	0.795
Model:	OLS	Adj. R-squared:	0.793
Method:	Least Squares	F-statistic:	451.3
Date:	Tue, 10 Sep 2019	Prob (F-statistic):	2.39e-275
Time:	01:21:17	Log-Likelihood:	-2991.2
No. Observations:	823	AIC:	5998.
Df Residuals:	815	BIC:	6036.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.025e-14	0.321	6.31e-14	1.000	-0.630	0.630
BA	0.0038	0.002	2.128	0.034	0.000	0.007
BDA	0.0911	0.128	0.713	0.476	-0.160	0.342
BMI	0.1332	0.012	11.182	0.000	0.110	0.157
WT	-0.0264	0.039	-0.679	0.497	-0.103	0.050
BPSys	0.3701	0.028	13.198	0.000	0.315	0.425
BPdia	0.0367	0.016	2.341	0.019	0.006	0.068
SM	0.0574	0.011	5.003	0.000	0.035	0.080

Omnibus:	35.712	Durbin-Watson:	1.183
Prob(Omnibus):	0.000	Jarque-Bera (JB):	44.356
Skew:	0.436	Prob(JB):	2.33e-10
Kurtosis:	3.729	Cond. No.	241.

Fig.1. Regression Analysis for predicting Health Parameter Behavior

Based on the Ordinary Least Square (OLS) [33] regression model Fig.1 it can be concluded that the variable 'VF ' can act as a function of other variables for predicting values . The variables BA, WT, BPSys, do have much effect on the values of other health indicators . The changes in the values of Visceral Fat and do the cause of change in values of BA, BDA, BMI, SM, BPSys and vice versa. This can be understood using Fish-bone Ishikawa diagram for better understanding.

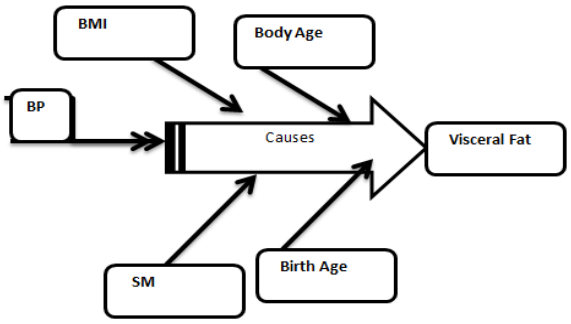

Fig.2. Causal Analysis of E-Health Data

Fig.2.shows which parameters are impacting the VF Variable. The value of VF can be calculated on the basis of all other seven parameters using the following regression equation .Using the prescribed medical limits of visceral fat , the onset of the health issue can be detected using equation no 1

$$y = cax1 + cbx2 + ccx3 + cdx4 + cex5 + cfx6 + cgx7 + chx8 + cix9 + cjx10 + ckx11 + cl12 + z \quad \ldots\ldots. \quad (1)$$

where y is the dependent variable (VF) and x1, x2 …..x12 are independent variables (WT,BMI, BPSys,BPDia,BA, BDA) associated with y. The ca , cb, …….. cl are coefficients work out by the OLS regression model. But , when the relationship between the various variables is complex and the nature of dataset is discrete and categorical [34], then it is hard to build one trend models. Then, there is a need to build a trend prediction model for each discrete class or category. In fact, the problem no longer remains a regression problem [35]. It becomes a classification problem.

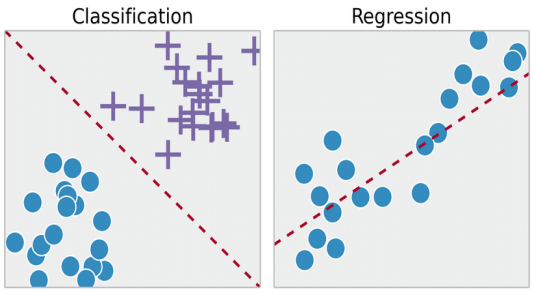

Fig.3.Difference Between Regression and Classification Problems

This research work revolves around building a supervised learning model or estimators that can help to identify and classify the data into three types of cases: normal cases, borderline, and the onset of health issues case. By definition, the normal cases are those cases that are in compliance with the medical limits of the health parameters.

The borderlines cases are those ones that show early signs of health problems problem and all cases that have crossed the border line medical limit are those ones that have confirmed the onset of life style diseases. The problem revolves round building an automatic system to detect all these three types of conditions.

III. METHODOLOGY

In this section, steps are explained with the help of flow diagrams, graphs and statistical procedures for developing a health monitoring model that try to build logical flow for detecting the onset of some health issue. The section begins with the explanation of health dataset and continues the explanation of all the proceeding steps.

A. Data Characteristics

The data related to the vital health parameters was collected using an Omron Body Composition, Blood Pressure and Sugar readings apparatus The data was collected over a period of three years (2016-2018) and the total data points collected were 2823. But, to maintain and to confirm to that data quality standards 1525 data points were used for this study. All rows that have missing or incomplete data entry were eliminated and the clean dataset was made. The data primarily consists of 12 health issue indicators. These parameters include Birth (BA) and body age (BDA), Height (H) ,Gender (G) , Weight (W) ,Body mass index (BMI), Body (BF) and visceral Fat (VF), Skeleton muscle (SM),Resting metabolism (RM) ,Waist (W), Blood pressure (BP) , Pulse rate (P) and Sugar readings (Sugar Fast) before and after breakfast (Sugar PP)

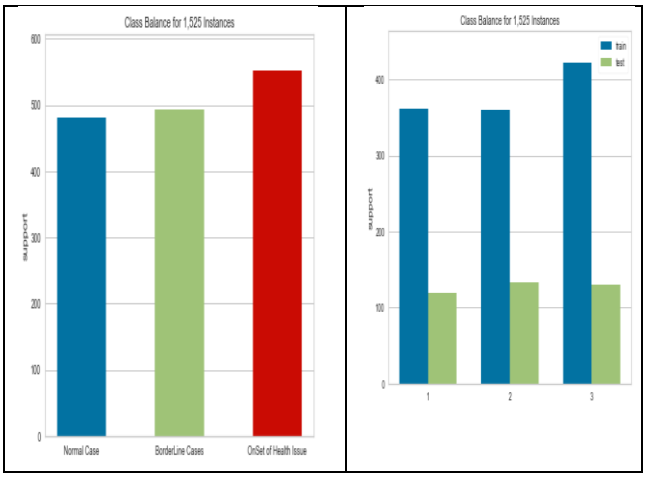

Fig.4. (a) Class balance analysis of the full dataset. (b) Class balance analysis in the training and testing phases.

It is always desired, that the ratio of the instance of each class dataset should be balanced [36] [37]. The dataset consists of three categories (encoded as 1, 2, and 3). The first category consist of those group of people who are healthy and the second group consists of people who are in medical terms are borderline cases. The third category cases are those who have developed signs and symptoms of unhealthy health indicators. These are those people, who have medically proven issues related to lifestyle diseases. It can observe from Fig.4 (a) and (b) that the difference between the class ratios is between the normal class and borderline class is insignificant. And the difference in the third class and the second /first class are small. Hence, there is no need to do class imbalance treatment. The ratio remains almost the same when the algorithms split the dataset into training and testing datasets. It is apparent that, in the process of building a linear estimator as a classifier, the imbalance class distribution will not impact the accuracy and no significant bias will be playing a role.

B. Supervised Linear Estimator Modeling for E-Health (SLEMH)

In this section, the process of automating the identification of instances with respect to the three classes (normal, borderline, onset of healthy issue) is discussed. Five linear estimators have been selected to model the classification process. The selection of these five models is based on fact; each of these handles the dataset differently and find the trends, borders between the classes differently. Each of these algorithms has a different way to deal with fitting of the data and process to deal with under fitting and overfitting.The evaluation of all these algorithms will help us to build an appropriate health monitoring system.

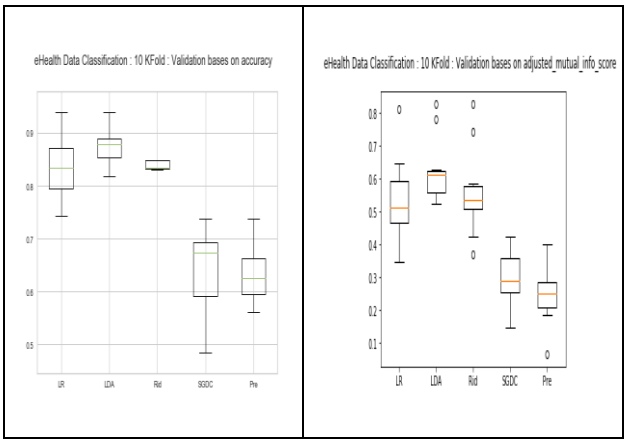

Fig.5. K-Fold Based Model Selection

Table-I: K-Fold Results for Model Selection.

	Model Selection		
S.No	Algorithm	Accuracy / Standard deviation	Adjusted Mutual Info /Standard deviation
1	MLR	0.845758 (0.054941)	0.561382 (0.120634)
2	LDA	0.879883 (0.040543)	0.628958 (0.093330)
3	Ridge	0.839280 (0.061256)	0.551529 (0.127746)
4	SGDC	0.630839 (0.143979)	0.296170 (0.143538)
5	Preceptor	0.630839 (0.081733)	0.244083 (0.083225)

It is clear from the boxplot and Table-I, that the five linear estimators /models have been put under selection procedure for building a health monitoring system that can detect three states of health. The first model is multinomial logistic regression (MLR or simply LR), the second one is Linear Discriminant Analysis (LDA) [38] [39].

In the third serial number the Ridge regression algorithm

[40] is mentioned. The fourth algorithm used under evaluation is the Stochastic Gradient decent model or SGDC and the fifth one is Preceptor Linear model.

Interpretation

It can be observed from the boxplots and Table-I that the LDA algorithm is the best performer. This is evidence from the values of the two metrics i.e. accuracy and adjusted mutual information. For both these metrics, ten rounds of evaluations were done and average using interquartile method (IQR) is computed. It can also be observed that the standard deviation values of the LDA are also quite low for both the metrics. This means, that the LDA algorithm is numerically consistent with its task of classification. From the performance of LDA, it can be further inferred that there is an average degree of separation between the three classes due to which it reached an accuracy level of 87.9%.

The LR algorithm is second best in terms of accuracy as well in terms of adjusted mutual information metric. It is clear that the results of LDA and LR are competing with other, while other algorithms have not been able to achieve a good level of accuracy with tight standard deviation in results. This clearly shows that the algorithms SGDC, Ridge, Preceptor are not suitable for the task of doing classification and estimation of respective classes of health status .Hence, in the coming sections, the best performers are compared. The next section investigates if it is possible to improve the accuracy of the LDA and LR algorithm with the help of feature engineering.

Feature Engineering [41]

In this section, which features that can help to build the most accurate estimator or machine learning model are investigated by using 10 K-fold mechanisms [19]. The advantage of 10 K-fold mechanisms is that an average of ten round of evaluation for all predictors is done and biased results are avoided. Elimination of features and predictors is done on the basis of mean importance computed using a performance metric. Lower ranked features are removed either in a backward or forward method. In context of our the research work, Recursive Feature Elimination with Cross-Validation (RFECV)[42] has been used and the outcomes are graphically shown in Fig. 6(a) and (b).

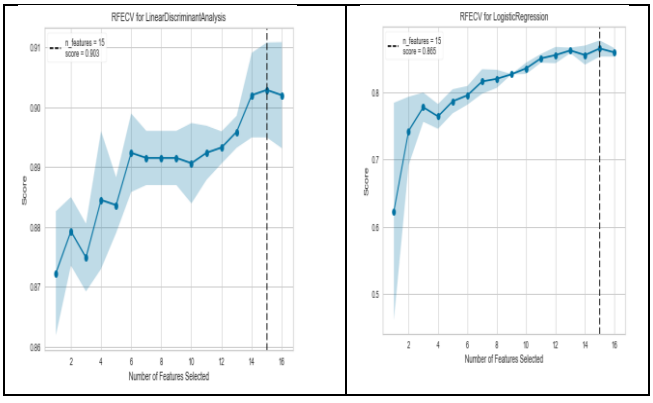

Fig.6.
(a) RFECV for Linear Discriminant Analysis
(b) **RFECV for Linear Regression**

Interpretation

It is can be observed from Figure 4 (a) and Figure 4 (b) that the RFECV algorithm produces comparative results in terms of feature selection. The behavior of LDA shows that its accuracy reached maximum only when the total number of feature selected reached 15 numbers. A similar outcome can be observed when the same algorithm is applied to logistic Regression. In fact, in both cases, the accuracy remains low when the number of features are less than 15.The second observation that can be made from these graphs that LDA algorithm is performing better as compared to the logistic regression right from the time, it was given just one predictor and its performance improves till the number of features becomes 15 and it finally dips a bit. From these observations, it can be concluded that all the features are important to expect one i.e. Sugar Fasting.

It can also be observed that there is a 2.1 % improvement of accuracy of the LDA algorithm by eliminating just one feature. During, the model selection phase with full feature set the accuracy of the LDA algorithm was highest among the five algorithms but it was 87.90% and now it is 90.003. And similar outcome can also be observed in case of LR. The accuracy of LR also improves by 2%.

IV. RESULT AND DISCUSSION

In this section, a comparative investigation on the aspect of the accuracy of both algorithms is done. This is done on the basis of micro and macro recall and precision values. The values of f1-score and area under curve are also computed so that tradeoff between the recall and precision can be understood for finally selecting the best liner estimator for the tasks of classification. The first, sub-section of this section give explanation of the performance metrics used from freezing the optimized model.

Comparison based on Recall and Precision

The value of the recall and precision gives the information on the level of accuracy at class level and at the level of full dataset level. In this section a comparative view of algorithms is shown after running ten rounds of evaluation on different sets of training and testing datasets.

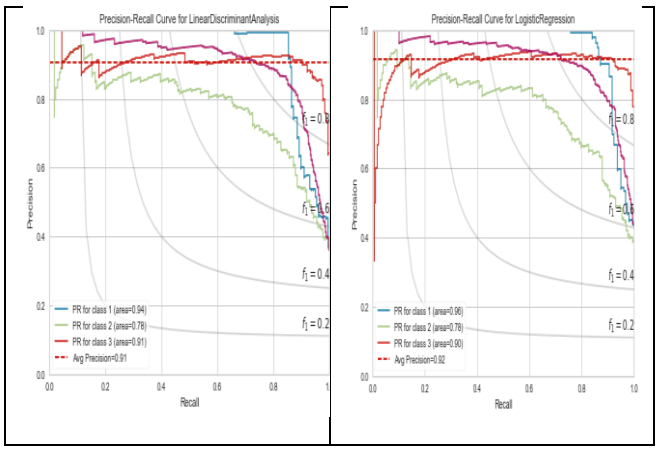

Fig.7.
(a)Precision-Recall Curves (LDA)
(b) Precision-Recall Curves (LR)

Interpretation

1. A system with high recall value and low precision values may lead to a result that a lot of wrongs predicted labels. But, in the context of this research, it can be observed that average precision value of LDA and LR is 0.91 and 0.92 respectively and recall values are lower than the precision values. But the values are quite high. This shows that both the algorithm have higher ratio of correctly predicted instances of each class.
2. The metric micro-precision is useful when the proposition of the instances of the classes is different and not equally balanced. The LDA graphs show that micro precision value for the normal class is 0.90 and for second class it is 0.78 and the class 's micro precision is 0.91. And comparative results can be observed in the case of LR. In-fact, the average precision of the LR is 1% higher than the LDA.
3. Until now, the LDA algorithm has been performing better than the LR based on accuracy, adjusted mutual information and other metrics both a detailed study of precision and recall graphs shows that LR has the almost same level of accuracy, values of area under curve (AUC) and f-score. Hence, for a deeper understanding of the performance of both algorithms, more experimentation and evaluations were done. The details are discussed in the next section.

Comparison based Learning Curves

The rate at which an estimator or classifier acquires the ability to predict and classifier is important. Especially, in cases when there are multiple candidates for becoming the most appropriate classifiers. As mentioned in the last section that the performance of both algorithms is highly comparable, this section evaluates both the contenders' algorithms for their repeatability, numerical stability and consistency of their results. Hence, for both algorithms learning phase and training phase are put under evaluations. This section, however, gives information on the comparative analysis between the LDA and LR performance during their learning phases.

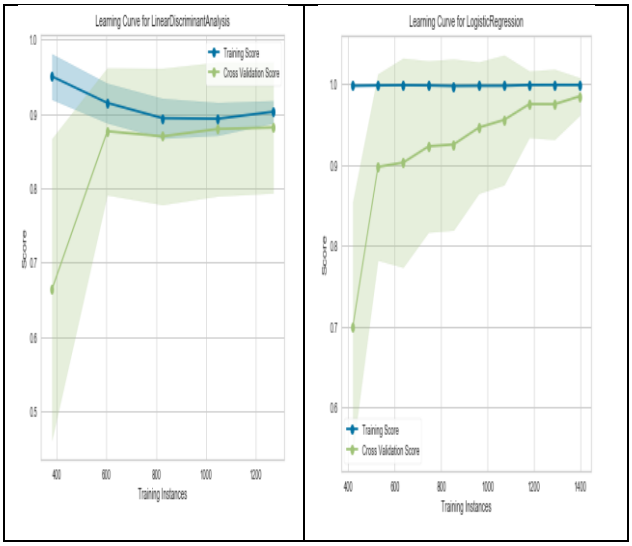

Fig.8.
(a) LDA Learning phase curve
(b) LR Learning phase curves

Interpretation

1. It can be seen from Fig.8.(b), that the training score of LR remains near to the value of 1 . Initially, the LR's training score is almost close 1, but as the number of instances of the dataset are increased the training score marginally drops to 0.99%. The cross validation score , however is quite low. The cross validation score in case of LR begins with value of 0.7 and then the curve converges close to the training score. This means the algorithms are getting generalized with increase in number of training instances. The gap between the training score curve and the validation curve is reduced when maximum level of accuracy is achieved.
2. In the case of LDA, it can seem that initially, the cross – validation score starts with values of 0.70, which is similar to the LDA .But, as multiple evaluations happen the LDA 's performance jumps to 0.88% and finally it reaches maximum level of 0.91. This value is lower as compared to the LR.
3. The gap between the training score curve and cross validation curve in case of LR algorithm is quite high initially as compared to the initial stage of LDA. This means that the LDA is a bit more stable in its performance but LR is more accurate.

Comparison based Validation Curves

The learning curve compassion between the LDA and LR show that LR is more accurate than the LDA. In this section, the testing phases of the algorithms are compared so that a logical conclusion can made about the final selection of the algorithm. The objective validation of the algorithms is based on 10 K Fold validation process. The LDA algorithm is evaluated using 'number of components' metric and LR is evaluated on the basis of 'C' parameter.

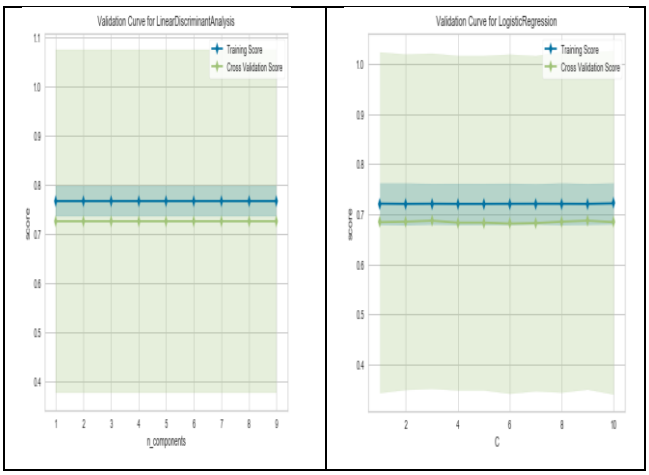

Fig.9. Validation Curves

Interpretation

It is clear from both the graphs that in the testing phase the LDA is giving 1% better performance as compared to the LR. This means the difference between the performances of both algorithms is not much. Secondly, it can also be observed that both algorithms are having similar behavior in terms of training score and cross –validation score. The bar graphs Fig.10 show how the cross validation scores are changing as evaluation run till 10 rounds.

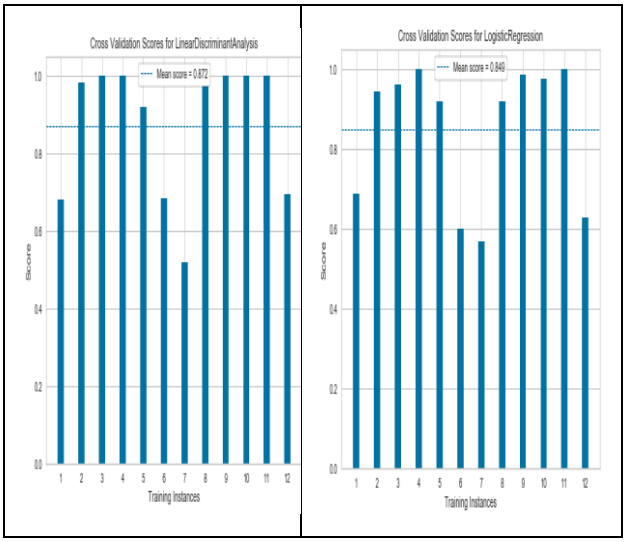

Fig.10. Cross Validation Scores in the testing phase.

From all the experiments and evaluations, it is clear that algorithms LR and LDA are most suitable for the task of classification on the basis of the current dataset. Both the algorithms have comparable performance in terms of accuracy (>0.90%) with 15 features.

V. CONCLUSION AND FUTURE SCOPE

It can apparent that the both the logistic regression estimator and linear discriminant analysis are good in the classification tasks. But, the LDA estimator is 1% more accurate than LR. The recall , precision and f1-scores value comparison also points out that the difference between the performance of both algorithm is not much but LDA seems to more stable in terms of repeatability of their results. The focus of this work has been to understand how linear regression modeling and estimation can help in classification tasks. For future scope, this work can be extended by doing explorative study on non-linear algorithms.

REFERENCES

1. [A. Pantelopoulos and N. Bourbakis, "A survey on wearable biosensor systems for health monitoring," in *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS'08 - "Personalized Healthcare through Technology,"* 2008.
2. A. Kaur and A. Jasuja, "Health monitoring based on IoT using Raspberry PI," in *Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2017*, 2017.
3. H. Fouad and H. Farouk, "Heart rate sensor node analysis for designing internet of things telemedicine embedded system," *Cogent Eng.*, vol. 4, no. 1, 2017.
4. G. Pardeshi and V. Kakrani, "Mobile based Primary Health Care System for Rural India," *Int. J. Nurs. Educ.*, vol. 3, no. 1, pp. 61–68, 2011.
5. B. L. Verma, S. K. Ray, and R. N. Srivastava, "Mathematical models and their applications in medicine and health," *Heal. Popul. Perspect. Issues*, 1981.
6. K. Mahato, A. Srivastava, and P. Chandra, "Paper based diagnostics for personalized health care: Emerging technologies and commercial aspects," *Biosensors and Bioelectronics*. 2017.
7. H. Baali, H. Djelouat, A. Amira, and F. Bensaali, "Empowering Technology Enabled Care Using IoT and Smart Devices: A Review," *IEEE Sensors Journal*. 2018.
8. G. Yang *et al.*, "A Health-IoT platform based on the integration of intelligent packaging, unobtrusive bio-sensor, and intelligent medicine box," *IEEE Trans. Ind. Informatics*, vol. 10, no. 4, pp. 2180–2191, 2014.
9. S. Palanivel Rajan, R. Sukanesh, and S. Vijayprasath, "Design and development of mobile based smart tele-health care system for remote patients," *Eur. J. Sci. Res.*, 2012.
10. E. Begic, Z. Begic, A. Dobraca, and E. Hasanbegovic, "Productive Cough in Children and Adolescents - View from Primary Health Care System.," *Med. Arch. (Sarajevo, Bosnia Herzegovina)*, vol. 71, no. 1, pp. 66–68, Feb. 2017.
11. J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: a review," *Artif. Intell. Rev.*, pp. 1–44, Jan. 2018.
12. G. Kaur, D. Sharma, and V. Kaur, "Telemedicine in Transient Phase: Emergence of M-Health Care Services," *Indian J. Sci. Technol.*, vol. 9, no. 15, May 2016.
13. A. Suresh and R. K. R. Varatharajan, "Health care data analysis using evolutionary algorithm," *J. Supercomput.*, 2018.
14. A. P. Grieve, "Medical Statistics," in *The Textbook of Pharmaceutical Medicine*, 2013.
15. V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, "The modeling of global epidemics: Stochastic dynamics and predictability," *Bull. Math. Biol.*, 2006.
16. T. N. Beran and C. Violato, "Structural equation modeling in medical research: A primer," *BMC Res. Notes*, 2010.
17. D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," *Int. J. Bio-Science Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.
18. B. Farran, A. M. Channanath, K. Behbehani, and T. A. Thanaraj, "Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study," *BMJ Open*, vol. 3, no. 5, p. e002457, Jan. 2013.
19. S. An, W. Liu, and S. Venkatesh, "Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression," *Pattern Recognit.*, vol. 40, pp. 2154–2162, 2007.
20. M. Ji, Q. He, J. Han, and S. Spangler, "Mining strong relevance between heterogeneous entities from unstructured biomedical data," *Data Min. Knowl. Discov.*, vol. 29, no. 4, pp. 976–998, 2015.
21. C. R. Hakkenberg, K. Zhu, R. K. Peet, and C. Song, "Mapping multi-scale vascular plant richness in a forest landscape with integrated LiDAR and hyperspectral remote-sensing," *Ecology*, 2018.
22. S. Basu, "Semi-supervised Clustering: Learning with Limited User Feedback," *PhD Texas Austin*, 2003.
23. Z. Izakian, M. Saadi Mesgari, and A. Abraham, "Automated clustering of trajectory data using a particle swarm optimization," *Comput. Environ. Urban Syst.*, 2016.
24. C. Y. Hung, W. C. Chen, P. T. Lai, C. H. Lin, and C. C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 3110–3113, 2017.
25. R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *J. R. Stat. Soc. Ser. B*, 1996.
26. R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani, "A significance test for the lasso," *Annals of Statistics*. 2014.
27. S. M. Alzahani, A. Althopity, A. Alghamdi, B. Alshehri, and S. Aljuaid, "An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction," *Lect. Notes Inf. Theory*, vol. 2, no. 4, pp. 310–315, 2015.
28. A. Maier, C. Syben, T. Lasser, and C. Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift fur Medizinische Physik*. 2019.
29. A. Suresh, R. Kumar, and R. Varatharajan, "Health care data analysis using evolutionary algorithm," *J. Supercomput.*, pp. 1–10, 2018.
30. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression: Third Edition*. 2013.
31. S. Clémençon, P. Bertail, E. Chautru, and G. Papa, "Optimal survey schemes for stochastic gradient descent with applications to M - estimation," *ESAIM - Probab. Stat.*, 2019.
32. S. Clemencon, P. Bertail, and E. Chautru, "Scaling up M-estimation via sampling designs: The Horvitz-Thompson stochastic gradient descent," in *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, 2015.
33. S. Seo and P. D. Gary M. Marsh, "A review and comparison of methods for detecting outliersin univariate data sets," *Dep. Biostat. Grad. Sch. Public Heal.*, 2006.
34. C. R. Bilder and J. M. Tebbs, "An Introduction to Categorical Data Analysis," *J. Am. Stat. Assoc.*, 2008.
35. C. Health, U. States, C. Rate, T. Method, L. Regression, and F. L. Regression, "Analysis of Case-control Studies Logistic Regression," *Res. Methods - II*, 2013.

36. R. Batuwita and V. Palade, "Class Imbalance Learning Methods for Support Vector Machines," in *Imbalanced Learning*, 2013.
37. D. Ramyachitra and P. Manikandan, "Imbalanced Dataset Classification and Solutions: a Review," *Int. J. Comput. Bus. Res. ISSN (Online*, 2014.
38. D. Chu and X. Zhang, "Sparse uncorrelated linear discriminant analysis," in *30th International Conference on Machine Learning, ICML 2013*, 2013.
39. W. Cai, G. Guan, R. Pan, X. Zhu, and H. Wang, "Network linear discriminant analysis," *Comput. Stat. Data Anal.*, 2018.
40. S. R. McCurdy, "Ridge regression and provable deterministic ridge leverage score sampling," in *Advances in Neural Information Processing Systems*, 2018.
41. R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief. Bioinform.*, 2017.
42. A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. 2018.

AUTHORS PROFILE

Amandeep Kaur has completed her Bachelor and Masters in Computer Science & Engg. She is a research scholar at IKGPTU KAPURTHALA, PUNJAB. She is working as Assistant Professor in CSE Department at BHSBIET Lehragaga, PUNJAB. She has teaching experience of thirteen years. Her area of interests are artificial intelligence, cloud computing and data mining. She has presented and published papers in National & International conferences and journals.

Dr.Anuj Kumar Gupta, is working as professor and Head at Chandigarh Group of Colleges. He has completed his PhD in Computer Science & Engineering from IKG Punjab Technical University. He has a vast teaching and research experience of above 18 years. His area of research is Wireless Networks & Security. He has guided 20+ M.Tech. thesis and 6 PhD thesis. He has published over 80 research papers in various National & International Journals and Conferences.