Indonesian Journal of Data and Science

Volume 5 Issue 1
ISSN 2715-9936
https://doi.org/10.56705/ijodas.v5i1.121

Research Article

Assessing the Predictive Power of Logistic Regression on Liver Disease Prevalence in the Indian Context

Izmy Alwiah [1,*], Umar Zaky [2], Aris Wahyu Murdiyanto [3]

[1] UIN Alauddin Makassar, Makassar, Indonesia, izmy.alwiah@gmail.com
[2] Universitas Teknologi Yogyakarta, Yogyakrta, Indonesia, umarzaky@uty.ac.id
[3] Universitas Jenderal Achmad Yani Yogyakarta, Yogyakarta, Indonesia, ariswahyu@unjaya.ac.id

Correspondence should be addressed to Izmy Alwiah; izmy.alwiah@gmail.com

Received 10 January 2024; Accepted 15 February 2024; Published 31 March 2024

Abstract:

This study assesses the predictive power of Logistic Regression in forecasting liver disease prevalence within the Indian demographic, specifically leveraging a dataset of 584 patient records from the NorthEast of Andhra Pradesh. Utilizing biochemical markers as predictive variables, the research employed a 5-fold cross-validation approach to validate the model's efficacy, focusing on performance metrics such as accuracy, precision, recall, and F1-Score. The findings reveal moderate to high accuracy and precision levels, indicating Logistic Regression's potential as a viable predictive tool in medical diagnostics, particularly in settings constrained by resources. However, the observed variability across different folds highlights the model's sensitivity to data partitioning, suggesting further refinement is needed to enhance reliability and generalizability. This study contributes to the existing body of knowledge by demonstrating Logistic Regression's applicability in predicting liver disease in a specific Indian context, offering insights into its practical implementation in healthcare screenings for early disease detection. Moreover, it underscores the need for more robust models or additional features to capture liver disease complexity more accurately. Future research directions include exploring complex models, expanding dataset diversity, and investigating individual predictor impacts on prediction accuracy. This research opens avenues for integrating predictive models into healthcare protocols, aiming to improve liver disease detection and patient outcomes.

Keywords: Logistic Regression, Liver Disease, Predictive Modelling, Machine Learning, Medical Diagnostics.
Dataset link: https://www.kaggle.com/datasets/fatemehmehrparvar/liver-disorders

1. Introduction

Liver disease, particularly liver cirrhosis, is emerging as a significant public health concern across the globe, with its prevalence noticeably increasing due to factors such as heightened alcohol consumption rates, chronic hepatitis infections, and the rise of obesity-related liver disease. These factors contribute to the complexity and urgency of addressing liver disease, especially in countries like India, where demographic and geographic diversities present unique challenges in healthcare management. The early detection of liver pathology plays a critical role in determining patient outcomes, significantly impacting survival rates and quality of life for affected individuals. However, despite the advancements in medical technology and diagnostics, there exists a notable disparity in the early diagnosis rates among different sub-populations, particularly among female patients, who are often diagnosed at later stages compared to their male counterparts. This diagnostic disparity raises concerns over the effectiveness of current predictive models and their applicability across diverse populations.

The primary problem this research aims to solve is the urgent need for improved predictive models that can accurately identify individuals at risk of liver disease early in the disease's progression, particularly in the Indian context. Such models are crucial for closing the diagnostic gap and ensuring that all sub-populations have equal access to early detection and subsequent treatment options. By focusing on the predictive power of Logistic Regression [1],

[2] analysis in interpreting biochemical markers, this study seeks to address the limitations of current diagnostic approaches and offer a more inclusive, accurate, and practical tool for early liver disease detection.

The objectives of this research are twofold: firstly, to assess the effectiveness of Logistic Regression [3], [4] as a predictive tool for liver disease based on biochemical markers and secondly, to explore the potential disparities in prediction accuracy between different sub-populations, with a particular focus on gender-based differences. Through this, the study aims to contribute to the development of more reliable and equitable diagnostic models that can be utilized in diverse demographic settings.

The research is guided by the hypothesis that Logistic Regression can serve as a robust predictive model for liver disease in the Indian population, offering significant improvements in early detection rates across all sub-populations, including those traditionally marginalized in healthcare settings. Additionally, it hypothesizes that gender-based disparities in prediction effectiveness can be identified and addressed through model adjustments, leading to more equitable health outcomes.

The scope of this research is limited to the analysis of patient records from the NorthEast of Andhra Pradesh, India, focusing on the applicability of Logistic Regression in predicting liver disease prevalence based on specific biochemical markers. While the findings may offer valuable insights into the predictive power of Logistic Regression in this context, the generalizability of the results to other regions or populations may be limited. Furthermore, the study acknowledges potential limitations in data completeness and the representation of all demographic groups within the dataset.

This research contributes to the existing body of knowledge by providing a detailed analysis of the effectiveness of Logistic Regression in predicting liver disease in the Indian context, highlighting potential areas for improvement in early diagnosis rates. Additionally, by identifying and addressing gender-based disparities in prediction accuracy [5]–[7], this study offers a pathway toward more equitable healthcare outcomes, aligning with broader public health goals of reducing the burden of liver disease through improved diagnostic tools and strategies.

2. Method:

This study employs a quantitative research design, utilizing a predictive modeling approach to assess the efficacy of Logistic Regression [8]–[10] in forecasting liver disease prevalence among the Indian population. The research design is structured to compare the predictive accuracy of the model across different sub-populations, with a special focus on identifying potential gender-based disparities. This comparative analysis is underpinned by the application of statistical methods to evaluate model performance metrics such as accuracy, precision, recall, and F-measure [11]–[13]. Our research is designed in five well-structured main stages, and their aspects are illustrated in **Figure 1**.

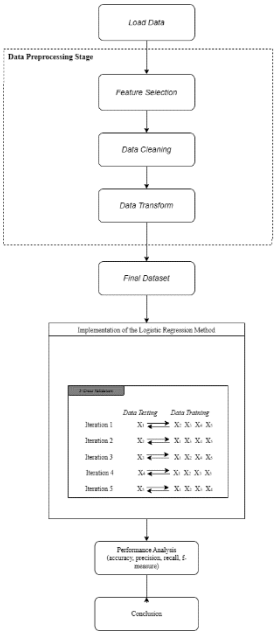

Figure 1. General Research Design Stages

Data Collection Process

The dataset comprises 584 patient records collected from the NorthEast of Andhra Pradesh, India, including both individuals diagnosed with liver disease (416 patients) and a control group without liver disease (167 patients). The selection criteria for the dataset were based on the availability of comprehensive biochemical marker information essential for the predictive model, including age, gender, total bilirubin, direct bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT, and Alkphos. Splitting dataset can show on **Figure 2**.

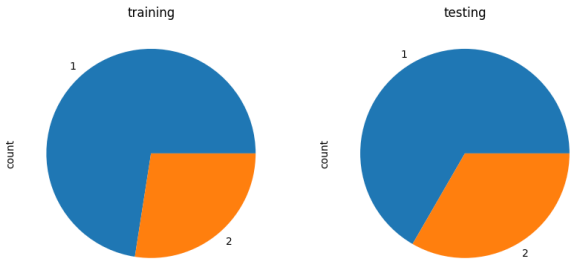

Figure 2. Splitting Dataset 10 % testing, 90% training

Tools and Technology Used

The study leverages Python for data preprocessing, analysis, and modeling, specifically employing libraries such as Pandas for data manipulation, Scikit-learn for implementing Logistic Regression and cross-validation techniques, and Matplotlib for data visualization. The choice of Python and its libraries is due to their wide acceptance in the scientific community for data science and machine learning tasks, offering robust support for statistical analysis and model development.

Data Collection Process

Data were retrospectively collected from medical records of patients who underwent liver function tests at healthcare facilities in the NorthEast of Andhra Pradesh, ensuring a comprehensive dataset reflective of the population's health status concerning liver disease.

Data Analysis Methods

Classification Algorithm: Logistic Regression

The analysis process began with data pre-processing, including cleaning (handling missing values via median imputation) and transforming data (encoding categorical variables and normalizing numerical variables). The Logistic Regression model was then applied [8], [9], [14].

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

Where p is the probability of having liver disease β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the predictors $X_1, X_2, \ldots, X_n$.

For model validation, a 5-fold cross-validation was implemented:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{M}_i \quad (2)$$

Where $CV_{(n)}$ is the cross-validation score, n is the number of folds, and M_i is the performance metric for fold i.

Performance Comparison Analysis

The performance of the classifier is evaluated using 5-fold cross-validation [15]–[19], where the dataset is divided into five subsets, and the model is trained and tested five times, each time using a different subset as the test set and the remaining as the training set. The performance metrics are calculated as follows [20]–[22]:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$Pericision = \frac{TP}{(TP + FP)} \quad (3)$$

$$Recall = \frac{TP}{(TP + FN)}$$

$$F - measure = \frac{2(presisi \times recall)}{(presisi + recall)}$$

Where TP, TN, FP and FN represent the numbers of true positives, true negatives, false positives, and false negatives, respectively.

3. Results and Discussion

Results

The application of Logistic Regression for predicting liver disease in the Indian context, specifically using a dataset from the NorthEast of Andhra Pradesh, yielded insightful results through the implementation of a 5-fold cross-validation technique. The performance metrics across the five folds revealed an average accuracy ranging from 69.23% to 74.14%, precision from 59.62% to 75.62%, recall mirroring the accuracy rates (due to the binary nature of the outcome), and F1-Scores between 61.29% and 69.73%. These metrics underscore the model's capacity to predict liver disease with a moderate to high degree of reliability, given the variability in performance across different data splits. The detailed results are presented in **Table 1** and visualized in **Figure 3** for a clearer understanding and comparison of the metrics across different iterations.

Table 1. Performance Metrics Across 5-Fold Cross-Validation for the Logistic Regression

K-n	Metrics			
	Accuracy	***Precision***	***Recall***	***F-Measure***
K-1	69%	60%	69%	61%
K-2	73%	71%	73%	70%
K-3	69%	68%	69%	68%
K-4	69%	66%	69%	67%
K-5	74%	76%	74%	66%
$\sum$ ***Avg***	71%	68%	71%	66%

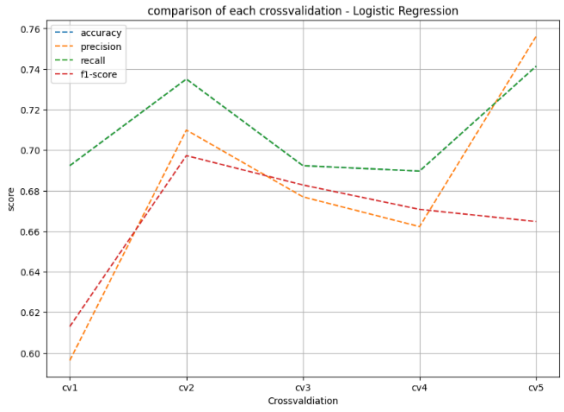

Figure 3. Visualisation Performance Metrics Across 5-Fold Cross-Validation for the Logistic Regression.

The data processing and analysis, employing logistic regression, highlighted the model's sensitivity and specificity in identifying liver disease cases. However, the variation in performance metrics across the folds indicates the model's sensitivity to data partitioning, suggesting a potential for overfitting or underfitting in specific data scenarios.

Discussion

The findings from this study contribute to the existing literature by demonstrating the feasibility of using logistic regression in a resource-constrained setting to predict liver disease prevalence with a reasonable degree of accuracy. The precision rates, indicating the model's ability to identify true liver disease cases among those predicted as positive, suggest that logistic regression can effectively reduce false positives, which is crucial in medical diagnostics to avoid unnecessary anxiety or treatment for patients.

Comparing these results with previous research indicates that while logistic regression holds promise, there is a considerable scope for improvement, especially in enhancing recall and F1-Score metrics. Previous studies have often highlighted the challenge of balancing sensitivity and specificity in disease prediction models, with a higher emphasis on ensuring that true disease cases are not missed (high recall). Our findings align with this perspective, underscoring the need for models that maintain high precision without compromising recall.

The practical implications of this research are significant, particularly for healthcare professionals and policymakers in India. By integrating such predictive models into healthcare screenings, especially in areas with limited access to advanced medical diagnostics, early detection of liver disease can be improved, thereby facilitating timely intervention and management. However, this study is not without its limitations. The variability in model performance across different folds highlights the influence of data partitioning on prediction outcomes, suggesting the need for a more robust model or additional features that can capture the complexity of liver disease more accurately. Moreover, the dataset, being from a specific region of India, may not fully represent the broader Indian population, potentially limiting the generalizability of the findings.

For future research, it is recommended to explore more complex models or ensemble methods that may offer better performance in terms of both precision and recall. Additionally, incorporating more diverse and larger datasets could help in developing a model with higher generalizability. Investigating the impact of individual predictors on the model's performance could also provide insights into the biochemical markers most indicative of liver disease, offering a more targeted approach to early diagnostics.

4. Conclusion

This study embarked on evaluating the predictive power of Logistic Regression in identifying liver disease within the Indian context, particularly utilizing a dataset from the NorthEast of Andhra Pradesh. The outcomes, illuminated by a 5-fold cross-validation, showcased a moderate to high accuracy and precision in disease prediction, yet underscored variability across different data splits. These findings affirm the initial hypothesis that Logistic Regression can serve as a robust tool for predicting liver disease, albeit with room for enhancement, particularly in recall and F1-Score metrics. The discussion further placed these results within the broader research landscape, emphasizing the model's potential utility in resource-constrained settings while noting the balance required between sensitivity and specificity—a critical aspect in the diagnostic process. This study contributes to the burgeoning field of predictive health analytics by demonstrating the applicability of logistic regression in a specific geographical and disease context, providing a foundation for future exploratory and predictive endeavours in similar settings.

Looking ahead, the variability in model performance across the dataset folds suggests an imperative for more sophisticated modelling techniques or the incorporation of additional predictive features to capture the multifaceted nature of liver disease more accurately. Further research leveraging larger and more diverse datasets could enhance the model's generalizability across different populations and healthcare settings. Additionally, exploring the impact of individual biochemical markers on the predictive accuracy could unearth more targeted diagnostic strategies, potentially revolutionizing early detection and treatment paradigms for liver disease. The practical implications of this research are profound, offering a blueprint for integrating predictive models into healthcare screening protocols, thereby bolstering early detection efforts and optimizing patient outcomes in the face of liver disease. This study, while a step in the right direction, paves the way for future investigations to refine and expand upon the predictive capacities of logistic regression and similar analytical tools in the domain of public health and disease prevention.

References:

[1] F. Huang, "Logistic Regression Fitting of Rainfall-Induced Landslide Occurrence Probability and Continuous Landslide Hazard Prediction Modelling," *Diqiu Kexue - Zhongguo Dizhi Daxue Xuebao/Earth Sci. - J. China Univ. Geosci.*, vol. 47, no. 12, pp. 4609–4628, 2022, doi: 10.3799/dqkx.2021.164.

[2] L. S. Van Velzen, "Classification of suicidal thoughts and behaviour in children: results from penalised logistic regression analyses in the Adolescent Brain Cognitive Development study," *Br. J. Psychiatry*, vol. 220, no. 4, pp. 210–218, 2022, doi: 10.1192/bjp.2022.7.

[3] T. M. Jawa, "Logistic regression analysis for studying the impact of home quarantine on psychological health during COVID-19 in Saudi Arabia," *Alexandria Eng. J.*, vol. 61, no. 10, pp. 7995–8005, 2022, doi: 10.1016/j.aej.2022.01.047.

[4] Y. Zhang, "Multi-label feature selection based on logistic regression and manifold learning," *Appl. Intell.*, vol. 52, no. 8, pp. 9256–9273, 2022, doi: 10.1007/s10489-021-03008-8.

[5] B. Cao, "Performance analysis and comparison of PoW, PoS and DAG based blockchains," *Digit. Commun. Networks*, vol. 6, no. 4, pp. 480–485, 2020, doi: 10.1016/j.dcan.2019.12.001.

[6] S. Rahman, "Performance analysis of boosting classifiers in recognizing activities of daily living," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, 2020, doi: 10.3390/ijerph17031082.

[7] A. A. Ewees, "Performance analysis of Chaotic Multi-Verse Harris Hawks Optimization: A case study on solving engineering problems," *Eng. Appl. Artif. Intell.*, vol. 88, 2020, doi: 10.1016/j.engappai.2019.103370.

[8] Z. Zhao, "Logistic Regression Analysis of Risk Factors and Improvement of Clinical Treatment of Traumatic Arthritis after Total Hip Arthroplasty (THA) in the Treatment of Acetabular Fractures," *Comput. Math. Methods Med.*, vol. 2022, 2022, doi: 10.1155/2022/7891007.

[9] V. N. Vasu, "Prediction of Defective Products Using Logistic Regression Algorithm against Linear Regression Algorithm for Better Accuracy," *2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2022*. pp. 161–166, 2022, doi: 10.1109/3ICT56508.2022.9990653.

[10] A. Bailly, "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models," *Comput. Methods Programs Biomed.*, vol. 213, 2022, doi: 10.1016/j.cmpb.2021.106504.

[11] G. Giri, I. A. Musdar, H. Angriani, and ..., "Enhancing Disease Management in Mango Cultivation: A Machine Learning Approach to Classifying Leaf Diseases," *Indones. J. ...*, 2023, doi:10.56705/ijodas.v4i3.111.

[12] N. Rismayanti and A. P. Utami, "Improving Multi-Class Classification on 5-Celebrity-Faces Dataset using Ensemble Classification Methods," Indones. J. Data …, 2023, doi: 10.56705/ijodas.v4i2.78.

[13] S. Hidayat, H. M. T. Ramadhan, and ..., "Comparison of K-Nearest Neighbor and Decision Tree Methods using Principal Component Analysis Technique in Heart Disease Classification," Indones. J. …, 2023, doi: 10.56705/ijodas.v4i2.70.

[14] B. H. Reddy, "Classification of Fire and Smoke Images using Decision Tree Algorithm in Comparison with Logistic Regression to Measure Accuracy, Precision, Recall, F-score," *14th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics, MACS 2022*. 2022, doi: 10.1109/MACS56771.2022.10022449.

[15] S. Ortiz-Toquero, "Classification of Keratoconus Based on Anterior Corneal High-order Aberrations: A Cross-validation Study," *Optom. Vis. Sci.*, vol. 97, no. 3, pp. 169–177, 2020, doi: 10.1097/OPX.0000000000001489.

[16] T. A. Reist, "Cross validation of aerodynamic shape optimization methodologies for aircraft wing-body optimization," *AIAA J.*, vol. 58, no. 6, pp. 2581–2595, 2020, doi: 10.2514/1.J059091.

[17] Z. Xiong, "Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation," *Comput. Mater. Sci.*, vol. 171, 2020, doi: 10.1016/j.commatsci.2019.109203.

[18] O. Karal, "Performance comparison of different kernel functions in SVM for different k value in k-fold cross-validation," *Proc. - 2020 Innov. Intell. Syst. Appl. Conf. ASYU 2020*, 2020, doi: 10.1109/ASYU50717.2020.9259880.

[19] M. Rafało, "Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis," *ICT Express*, vol. 8, no. 2, pp. 183–188, 2022, doi: 10.1016/j.icte.2021.05.001.

[20] S. Rahmah, H. Azis, D. Widyawati, and A. U. Tenripada, "Prediksi potensi donatur menggunakan model Logistic Regression," *Indones. J. Data Sci.*, vol. 4, no. 1, pp. 31–37, 2023, doi:10.56705/ijodas.v4i1.64.

[21] H. Azis, D. Widyawati, and ..., "Prediksi potensi donatur menggunakan model Logistic Regression," *Indones. J. ...*, 2023, doi:10.56705/ijodas.v4i1.64.

[22] H. Azis, F. T. Admojo, and E. Susanti, "Analisis Perbandingan Performa Metode Klasifikasi pada Dataset Multiclass Citra Busur Panah," *Techno.Com*, vol. 19, no. 3, 2020, doi: 10.33633/tc.v19i3.3646.