INTERNATIONAL RESEARCH JOURNAL OF
HUMANITIES AND INTERDISCIPLINARY STUDIES
(Peer-reviewed, Refereed, Indexed & Open Access Journal)
DOI : 03.2021-11278686 ISSN : 2582-8568 IMPACT FACTOR : 8.428 (SJIF 2026)

A Data-Driven Exploratory Analysis and Machine Learning Approach for Clinical Risk Stratification

Amol Avinash Shinde [1], Dr. D.V. Sahasrabuddhe[2] , Akanksha Nandkumar Jamdade[3], Durgesh Babaso Mane[4]

[1] [2] [3] [4] Assistant Professor,
[1] [3] [4] Krantiagrani Dr. G. D. Bapu Lad Mahavidyalaya, Kundal (Maharashtra, India)
[2] BVDU, Institute of Management and Rural Development Administration,
Sangli (Maharashtra, India)
***E-mail:** write2amol439@gmail.com[1], dhanashreevs@gmail.com[2],*
akankshajamadade2000@gmail.com[3], durgeshmane128@gmail.com[4]

DOI No. 03.2021-11278686 **DOI Link ::** https://doi-ds.org/doilink/01.2026-46214395/IRJHIS2601031

Abstract:

The clinical risk assessment processes involved in providing the highest level of preventive care for the patient have been established as a key/profound feature of modern healthcare systems, and with the continuing increase in availability of patient health data, the use of machine learning to predict clinical risk for patients has gained significant popularity among the medical community. This study provides a data-driven exploratory analysis of the use of machine-learning (ML) algorithms in order to develop a comparative evaluation of various algorithms for clinical risk assessment through exploring the data in EDA (Exploratory Data Analysis) prior to ML modeling via using a dataset of 2200 records created in CSVs (Comma Separated Values). EDA is performed on the dataset to investigate the variability of the values within the dataset, the characteristics of the dataset and the frequency distributions of the target classes prior to the modeling of the data. Multiple versions of ML algorithms were applied to the same 2200 record dataset and the evaluation was based on standard evaluation metrics. The results of this study indicate that by conducting an EDA on the dataset, the prediction accuracy and interpretability of the resulting ML models for clinical risk assessment were improved significantly.

Keywords: *Exploratory Data Analysis, Machine Learning, Clinical Risk Stratification, Classification, Predictive Analytics, Healthcare Data*

1. Introduction:

The fast increase in digitalization of healthcare has resulted in significant amounts of patient data created by different sources (e.g., Electronic Health Records, wearables, and clinical monitoring systems). Successful use of this patient data is crucial to recognizing health risks early and supporting informed clinical decision-making (Topol, 2019) [1]. Traditional statistics often struggle with the ability to model the complex and non-linear nature of most clinical datasets.

Machine learning provides tools for discovering and identifying the hidden patterns in healthcare data and developing comprehensive predictive models (Jordan and Mitchell, 2015) [2]. Unfortunately, machine learning's effectiveness is strongly dependent upon the format and quality of the data utilized. Exploratory Data Analysis (EDA) helps us understand the characteristics of our data, and it allows for the detection of anomalies and the discovery of relationships between variables before the implementation of machine learning algorithms (Tukey, 1977) [3].

This research will utilize an EDA to help determine how to develop, implement and compare supervised machine learning models for clinical risk stratification of individual patients based upon their clinical profile, using a dataset that contains 2200 patient records.

2. Dataset Description:

The data used in this study consists of 2,200 examples stored in CSV format, with each example representing an individual. Each individual has a number of attributes: age, heart rate, cholesterol level, blood sugar level and their clinical risk category - which indicates the group they belong to (i.e. Low, Moderate, Severe).

During data preprocessing, we checked for missing values and inconsistencies in the data and performed normalisation of the numerical features where necessary. In addition, we created a binary variable for the clinical risk categorisation, as these labels are categorical and therefore must be converted to numerical format to allow for machine learning algorithms to process them (Han, Kamber, and Pei, 2012)[4].

3. Exploratory Data Analysis (EDA):

We performed Exploratory Data Analysis (EDA) on this dataset to gain a complete understanding of its structure and quality. The process allowed us to review the data from all aspects (features, distribution, variability, relationships) as well as determine whether there was a balance across classes(Wickham et al., 2014) [5]. Our visual and statistical methods provided the necessary information to guide the way we processed our data, selected our features, and chose machine learning methods that would provide the best results.

3.1 Age Variability Analysis:

The purpose of age variability analysis is to provide a visual representation of the spread, median and range of ages among the patient population represented in this dataset, using the Box Plot as a tool for analysis. We can see in Figure-1 from the Box Plot that by understanding the age variability in relation to other factors, we are able to identify those patients who may have an extreme age (either very young or very old), which could impact their clinical risk (the impact of this will be discussed in detail later). Age variability is important because it is a critical factor that affects how a disease progresses as well as a patient's clinical risk.

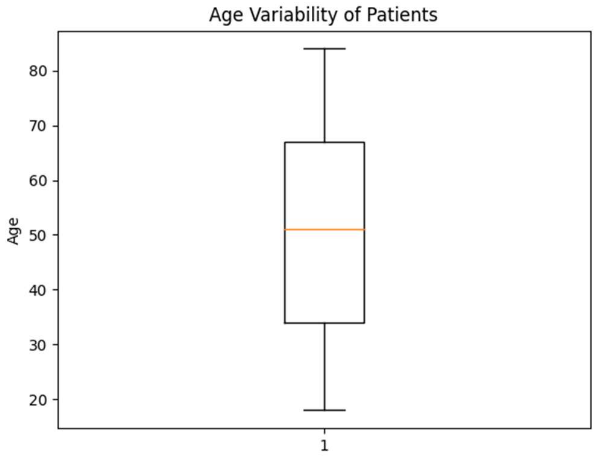

Figure 1: Age Variability of Patients

3.2 Cholesterol Level Distribution:

Cholesterol levels in patients can be described and analysed using a histogram in Figure-2. The histogram will show both the average cholesterol level (ARL - an average of all cholesterol levels in the dataset) and the different frequency distributions of normal cholesterol levels (total number of patients in the ARL/total number of patients in the dataset). The values shown on the histogram will also allow for the identification of those patients who may have an extreme cholesterol level and could therefore represent an increased clinical risk (an increase in cardiac death and cardiovascular disease).

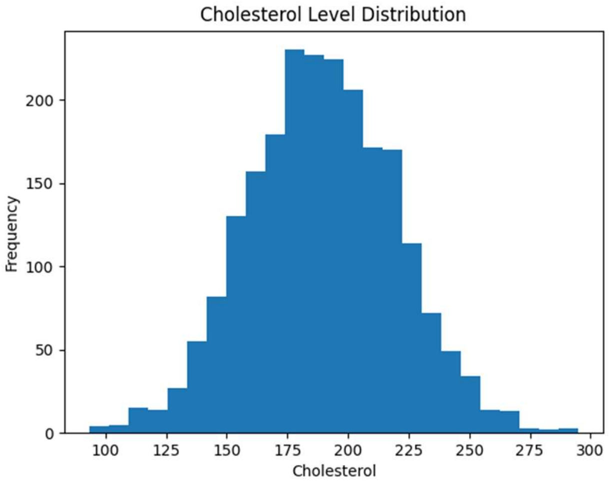

Figure 2: Cholesterol Level Distribution

3.3 Clinical Risk Level Proportions:

The pie chart demonstrates the distribution of patients into 3 main categories of clinical risk (low, moderate and severe) that shown in Figure-3. Determining these proportions will assist in determining the relative balance of classes in the dataset when training an unbiased machine learning model. If one class of risk is over-represented in the dataset, it will impact the bias of the predictive modelling toward that specific class of risk. Identifying the proportions of the risk categories will provide support to make informed decisions regarding the use of resampling techniques and the evaluation metric application to verify that fair and accurate clinical risk prediction have been achieved.

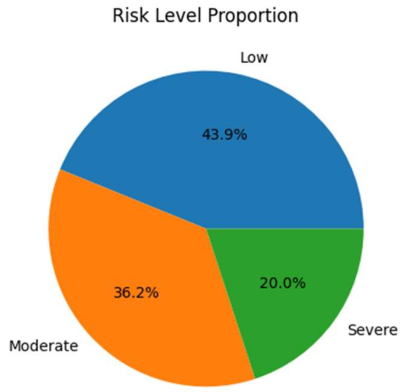

Figure 3: Risk Level Proportion

3.4 Age and Blood Sugar Relationship:

The scatter plot shows in Figure-4 a relationship between age and blood sugar levels; visually it is easy to see trends, groupings and relationships between age and blood sugar levels. This clinical relationship is noteworthy as blood sugar levels are likely to rise with increasing age due to metabolic changes. Observing the patterns in the scatter plot will help verify if age can be used as a good predictor for blood sugar changes, therefore supporting the relevancy of features and increasing the effectiveness of clinical risk predictive models.

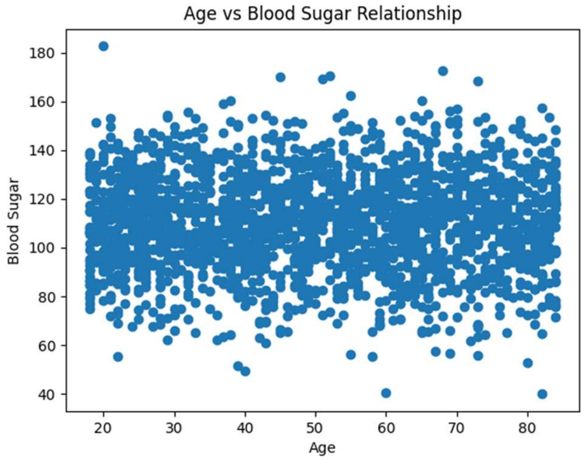

Figure 4: Age vs Blood Sugar Relationship

4. Developing Machine Learning Model:

Based on the findings from Exploratory Data Analysis (EDA), we were able to use several different approaches for building models. The approaches we used were Logistic Regression, Decision

Trees (C4.5), Random Forests, Support Vector Machines and K-Nearest Neighbour algorithms (KNN). The data was randomly split into 80% training and 20% testing set and we also applied feature scaling to numerical attributes for better model convergence and fair distance and margin between samples (Hastie et al., 2017)[9].

5. Performance Evaluation and Comparative Analysis:

Performance metrics used to evaluate the performance of models are accuracy, precision, recall, F1-score, and the confusion matrix for each model as documented by Powers (2011)[10]. The results of comparing the models suggest that Random Forest and Support Vector Machines performed best because they are able to model non-linear relationships and therefore had improved performance compared to the other models in capturing very complicated feature interactions that shown in Table-1.

Sr. No.	Model	Accuracy (%)	Precision	Recall	F1-Score
1.	Logistic Regression	84.6	0.83	0.82	0.82
2.	Decision Tree (C4.5)	86.1	0.85	0.84	0.84
3.	Random Forest	91.8	0.91	0.92	0.92
4.	Support Vector Machine	90.9	0.90	0.91	0.91
5.	K-Nearest Neighbors	82.4	0.81	0.80	0.80

Table 1: Comparative Performance Metrics of Machine Learning Models

After evaluating the performance of these models, it became evident that the implemented machine learning models have clear differences in how each model performs. In particular, Random Forest has the highest accuracy and F1-score, which indicates high performance in prediction as well as robustness to noise because it uses the ensemble approach. In addition to the Random Forest model performing well, SVM performed very well and takes advantage of its ability to model complex decision boundaries. Logistic Regression also had stable and interpretable results; however, it demonstrated limitations in capturing nonlinear patterns. KNN had the lowest level of performance because it was too sensitive to the feature scaling and class overlap in the data. Analysis of the confusion matrix supports the conclusion that the models which use ensemble and margin-based methods produce fewer misclassifications leading to more accurate and trustworthy predictions of health risks based on the evidence from this research study.

6. Results and Discussion:

Based on the findings from this research, EDA increases the effectiveness of the models used for clinical risk stratification through the use of exploration and comparison techniques, which guide both pre-processing and feature selection processes. The findings suggest that the ensemble and kernel-based algorithms performed better than the simpler classifiers, supporting previous research (Breiman, 2001)[11] indicating that in the field of healthcare, selecting models based on data is important, as opposed to simply selecting models based on intuition.

7. Conclusion:

An exploratory and comparative study (i.e., the research) based on actual data has determined that 2,200 patient records were analysed to create a model for the prediction of clinical risk through various machine-learning methods which have the potential of being applied to assess patients' risk. The use of exploratory data analysis in conjunction with machine-learning methods increases the accuracy and interpretability of the predictions generated using these algorithms; thus, further supporting the potential of exploratory data analysis in producing more robust predictive systems for healthcare.

8. Future Work:

Subsequent studies should evaluate larger amounts of information, sophisticated deep learning methodologies, and state-of-the-art methods for understanding artificial intelligence systems to enhance patient safety within the context of decision-support healthcare systems (Doshi-Velez & Kim 2017) [12].

References:

1. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. ***Nature Medicine, 25***(1), 44–56.
2. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. ***Science, 349***(6245), 255–260.
3. Tukey, J. W. (1977). ***Exploratory data analysis***. Addison-Wesley.
4. Han, J., Kamber, M., & Pei, J. (2012). ***Data mining: Concepts and techniques*** (3rd ed.). Morgan Kaufmann.
5. Wickham, H., Cook, D., Hofmann, H., & Buja, A. (2014). Graphical inference for infovis. ***IEEE Transactions on Visualization and Computer Graphics, 20***(12), 1981–1990.
6. World Health Organization. (2021). ***Cholesterol and cardiovascular risk***. WHO Press.
7. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. ***IEEE Transactions on Knowledge and Data Engineering, 21***(9), 1263–1284.
8. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. ***New England Journal of Medicine, 380***(14), 1347–1358.
9. Hastie, T., Tibshirani, R., & Friedman, J. (2017). ***The elements of statistical learning*** (2nd ed.). Springer.
10. Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness and correlation. ***Journal of Machine Learning Technologies, 2***(1), 37–63.
11. Breiman, L. (2001). Random forests. ***Machine Learning, 45***(1), 5–32.

12. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. ***arXiv preprint arXiv:1702.08608***.
13. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. ***New England Journal of Medicine, 375***(13), 1216–1219.
14. Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record analysis. ***IEEE Journal of Biomedical and Health Informatics, 22***(5), 1589–1604.
15. Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. ***Nature Medicine, 25***(1), 24–29.