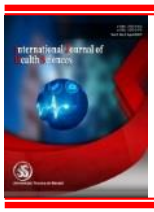

International Journal of Health Sciences

Available online at www.sciencescholar.us
Vol. 8 No. 3, December 2024, pages: 289-297
E-ISSN: 2550-696X
https://doi.org/10.53730/ijhs.v8n3.15189

Predicting Hospital Readmissions in Diabetes Patients: A Comparative Study of Machine Learning Models

Alekhya Gandra [a]

Manuscript submitted: 09 May 2024, Manuscript revised: 18 August 2024, Accepted for publication: 24 September 2024

Corresponding Author [a]

Keywords

diabetes;
healthcare analytics;
hospital readmission;
machine learning;
predictive modelling;

Abstract

Objective**:** Diabetes is a chronic condition affecting millions worldwide, requiring continuous medical care to manage and prevent complications *(Powers & D'Alessio, 2016)*. Hospital readmission rates among diabetes patients are high, contributing to significant healthcare costs and resource strain *(Zeng & Liu, 2015).* This study aims to predict hospital readmissions for diabetes patients using machine learning (ML) techniques to identify high-risk patients and reduce unnecessary readmissions. Methodology**:** Six machine learning algorithms—Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, and CATBoost—were used to classify patients based on their likelihood of being readmitted within 30 days. The Diabetes 130-US hospitals dataset from the UCI Machine Learning Repository *(Strack et al., 2014*) was employed, leveraging demographic, clinical, and discharge-related variables to build predictive models. Performance metrics such as accuracy, precision, recall, and AUC-ROC were used to evaluate the models. Results**:** The CATBoost classifier performed the best, achieving an AUC score of 0.70 and an accuracy of 64.2%. Key predictive features included the number of inpatient visits, medications, and the duration of hospital stays. The results emphasize the value of machine learning in predicting hospital readmissions, offering actionable insights for healthcare providers. Conclusion**:** This study demonstrates the effectiveness of machine learning, particularly CATBoost, in identifying high-risk diabetes patients for targeted interventions. These models can improve patient outcomes and optimize healthcare resources by reducing unnecessary readmissions. Future research should explore integrating real-time health data from wearables and the influence of social determinants on readmission rates to enhance predictive accuracy.

[a] Atlanta, Georgia, United States

Contents

Abstract 289
1 Introduction 290
2 Materials and Methods 290
3 Results and Discussions 291
3.1 Results 291
3.2 Discussions 294
4 Conclusion 295
Acknowledgments 295
References 296
Biography of Authors 297

1 Introduction

Diabetes is a widespread chronic condition that requires ongoing medical attention to manage blood sugar levels and prevent serious complications. According to Powers & D'Alessio (2016), diabetes affects millions globally, making it one of the most challenging public health issues. Hospital readmissions among diabetes patients are particularly frequent, adding to healthcare costs and straining resources (Zeng & Liu, 2015). These readmissions, which often occur within 30 days of discharge, are typically triggered by factors such as poor medication adherence, inadequate post-discharge care, and the progression of comorbidities like heart disease and hypertension (Strack et al., 2014).

Traditional methods for predicting hospital readmissions have relied on statistical models such as logistic regression or rule-based approaches, which use a limited set of variables (McHugh et al., 2013). While these methods provide a basic level of predictive power, they often fall short in accuracy and scalability. Traditional models struggle to capture the complex, non-linear relationships between variables critical in predicting hospital readmissions (van Walraven et al., 2010; Halfon et al., 2002). Machine learning (ML) offers a more refined approach capable of addressing these limitations by leveraging larger datasets and more sophisticated algorithms (Zheng et al., 2017).

By employing ML models, healthcare providers can analyze vast amounts of patient data to uncover hidden patterns and correlations that traditional models may miss (Miotto et al., 2016). These models can also account for a broader range of features, including time-based factors such as the frequency of hospital visits and patient-specific characteristics like medication adherence, leading to more accurate predictions. Additionally, ML models can be continuously updated as more data becomes available, making them suitable for real-time decision-making in healthcare settings (Deo, 2015). Early prediction of hospital readmissions allows healthcare providers to implement targeted interventions, such as medication adjustments or enhanced patient education, reducing the likelihood of readmission and improving overall patient outcomes (Kansagara et al., 2011; Artetxe et al., 2018).

This study explores and compares several machines learning models, including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM, and CATBoost, to determine their effectiveness in predicting hospital readmissions among diabetes patients (Björk, 2001; Conget, 2002). By doing so, this study seeks to provide healthcare providers with more accurate tools to identify high-risk patients and reduce the frequency of preventable hospitalizations.

2 Materials and Methods

This study's publicly available *Diabetes 130-US hospitals dataset* was obtained from the *UCI Machine Learning Repository* (Strack et al., 2014). This dataset contains over 100,000 admissions for diabetes patients from 130 hospitals, spanning ten years. Each patient record includes attributes such as demographics, clinical data, medication history, and previous admissions, providing a rich source of information for machine learning models (Kononenko, 2001; Mair et al., 2000).

Key Features:

- *Demographics*: Age, gender, race.
- *Clinical Variables*: Number of lab tests, medications, duration of hospital stays, and number of diagnoses.
- *Comorbidities*: Conditions like hypertension and heart disease.
- *Discharge-Related Variables*: Discharge disposition, admission source, and length of stay.
- *Previous Admissions*: Admission history within the last year.

The target variable is readmission, defined as a binary outcome: 1 if the patient was readmitted within 30 days and 0 if they were not. Given medical data's complexity and noisy nature, preprocessing steps were critical. Missing values were imputed using mean imputation for numerical features and mode imputation for categorical features (Zhang et al., 2020). Categorical variables were encoded using one-hot encoding, and the data was standardized to improve model performance (Miotto et al., 2016).

The dataset was split into 70% training and 30% testing sets to evaluate model performance. Six machine learning models were employed: **Logistic Regression (LOJ)**, **Random Forest Classifier (RF)**, **Gradient Boosting Classifier (GBM)**, **XGBoost Classifier (XGB)**, **LightGBM Classifier (LGBM)**, and **CATBoost Classifier (CATB)**. Each model was implemented using the Scikit-learn library, and their performance was evaluated using accuracy, recall, precision, F1-score, and the AUC-ROC metric (Breiman, 2001; Friedman, 2001).

To optimize model performance, Hyperparameter tuning was conducted for Random Forest, LightGBM, and CATBoost. The hyperparameters for Random Forest included the maximum depth of trees, number of estimators, and minimum samples for a split (Breiman, 2001). For LightGBM, the learning rate, number of estimators, and subsample size were optimized (Friedman, 2001). CATBoost hyperparameters were tuned to adjust the learning rate, iterations, and depth (Prokhorenkova et al., 2018).

3 Results and Discussions

3.1 Results

The results of the model evaluations are summarized in Table 1. **CATBoost** emerged as the top-performing model with an AUC score of 0.70 and an accuracy of 64.2%, followed closely by **LightGBM** and **XGBoost**, which also achieved comparable results. While still valuable, the *Logistic Regression* and *Random Forest* models performed relatively lower than AUC scores of 0.65 and 0.64, respectively. Key predictive features identified by the models included the number of inpatient visits, the number of medications prescribed, and the duration of hospital stays (Ogundokun et al., 2020; Houthooft et al., 2015). These features played a critical role in predicting readmissions, highlighting the importance of both clinical and administrative data in building robust predictive models for hospital readmission (Strack et al., 2014).

Table 1
Results of each model for training and validation sets for each of the six models

#	Classifier	Data_set	AUC	Accuracy	Recall	Precision	Score	Specificity
1	LOJ	train	0.674693	0.623063	0.546765	0.644591	0.591661	0.699193
2	LOJ	val	0.643858	0.600353	0.535354	0.617215	0.573377	0.665782
3	RF	train	0.667576	0.620419	0.589412	0.62782	0.60801	0.651357
4	RF	Val	0.640279	0.598149	0.576636	0.604234	0.590112	0.619805
5	GBM	train	0.700067	0.646419	0.615294	0.655594	0.634805	0.677476
6	GBM	val	0.653895	0.61269	0.596399	0.618116	0.607063	0.629089
7	XGB	train	0.700067	0.646419	0.615294	0.655594	0.634805	0.677476
8	XGB	val	0.653895	0.61269	0.596399	0.618116	0.607063	0.629089
9	LGBM	train	0.697602	0.646126	0.613971	0.655622	0.634113	0.67821

#	Classifier	Data_set	AUC	Accuracy	Recall	Precision	Score	Specificity
10	LGBM	Val	0.653716	0.609826	0.59069	0.615842	0.603004	0.629089
11	CATS	train	0.718581	0.65942	0.617794	0.673345	0.644375	0.700954
12	CATB	val	0.657078	0.61291	0.58498	0.621269	0.602579	0.641026

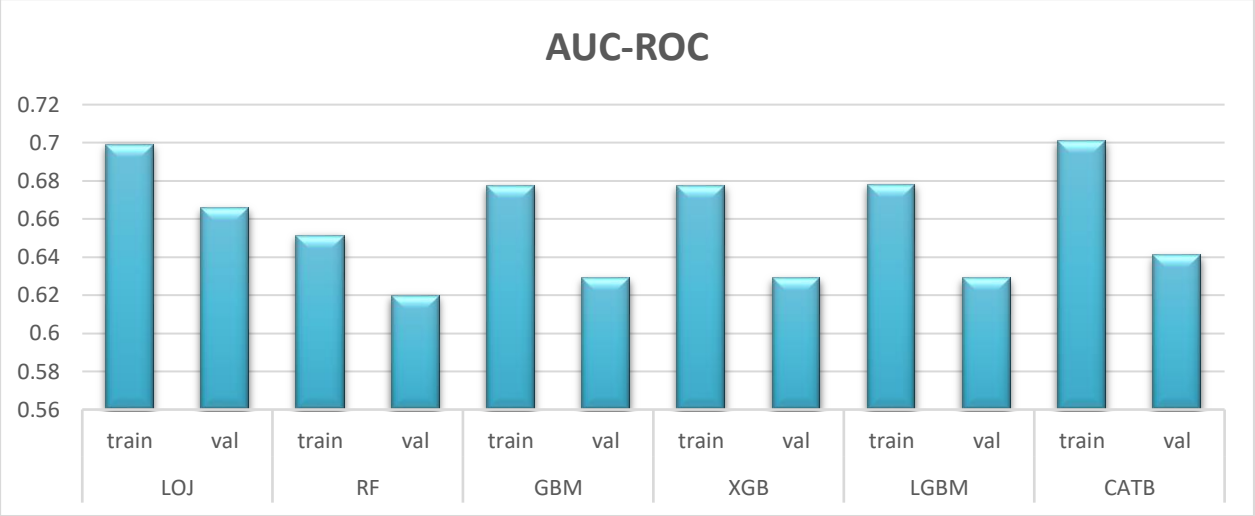

Figure 1. AUC-ROC for Training and Validation sets for each of the six models

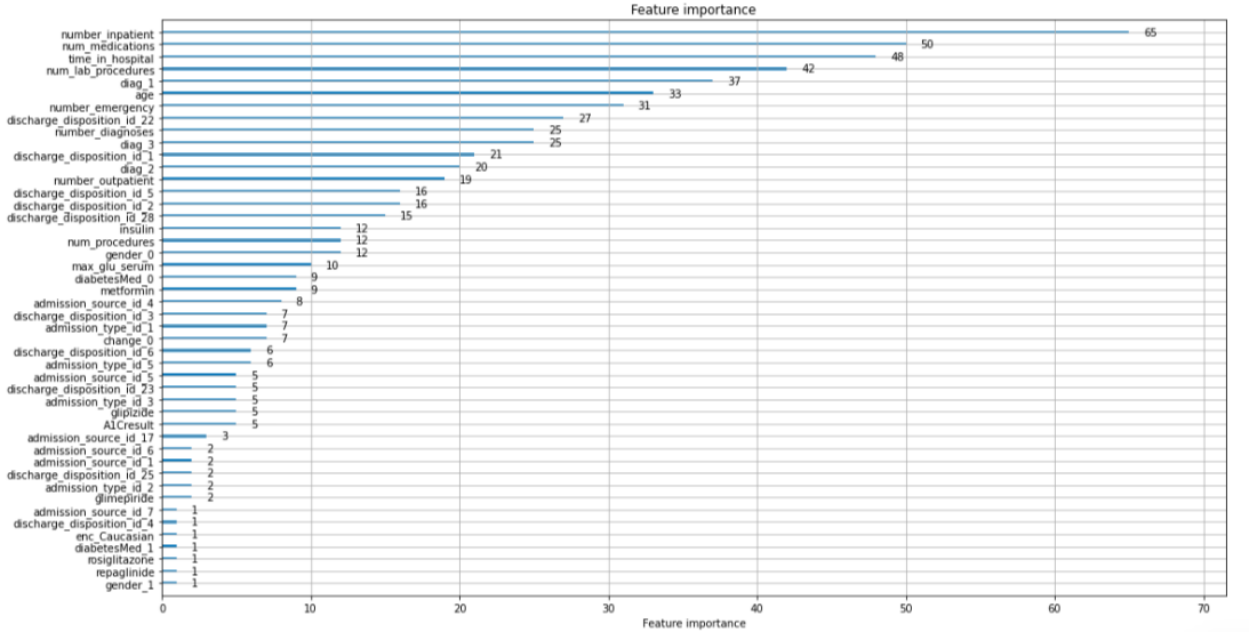

Figure 2. Feature importance based on Light-GBM Classifier (LGBM)

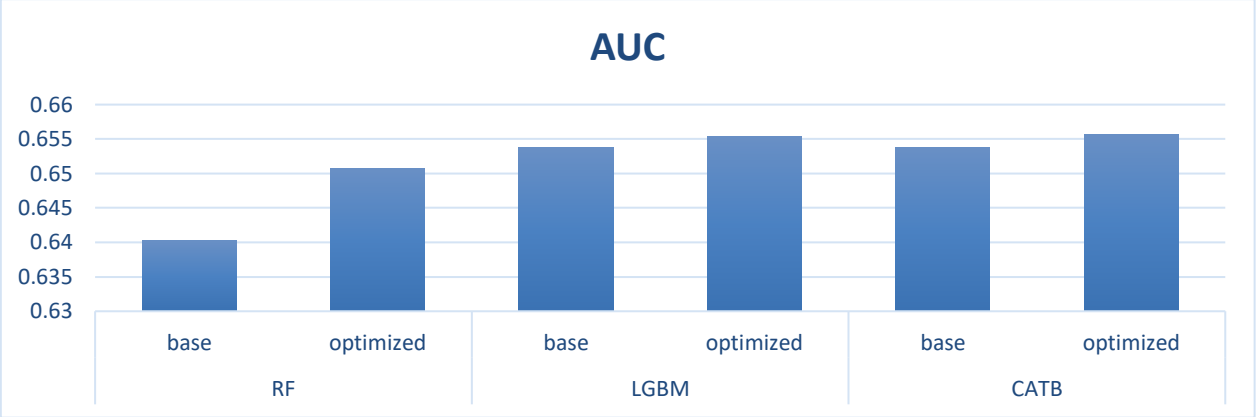

Figure 3: Hyperparameter Tuning Results for RF, LGBM, and CATB models

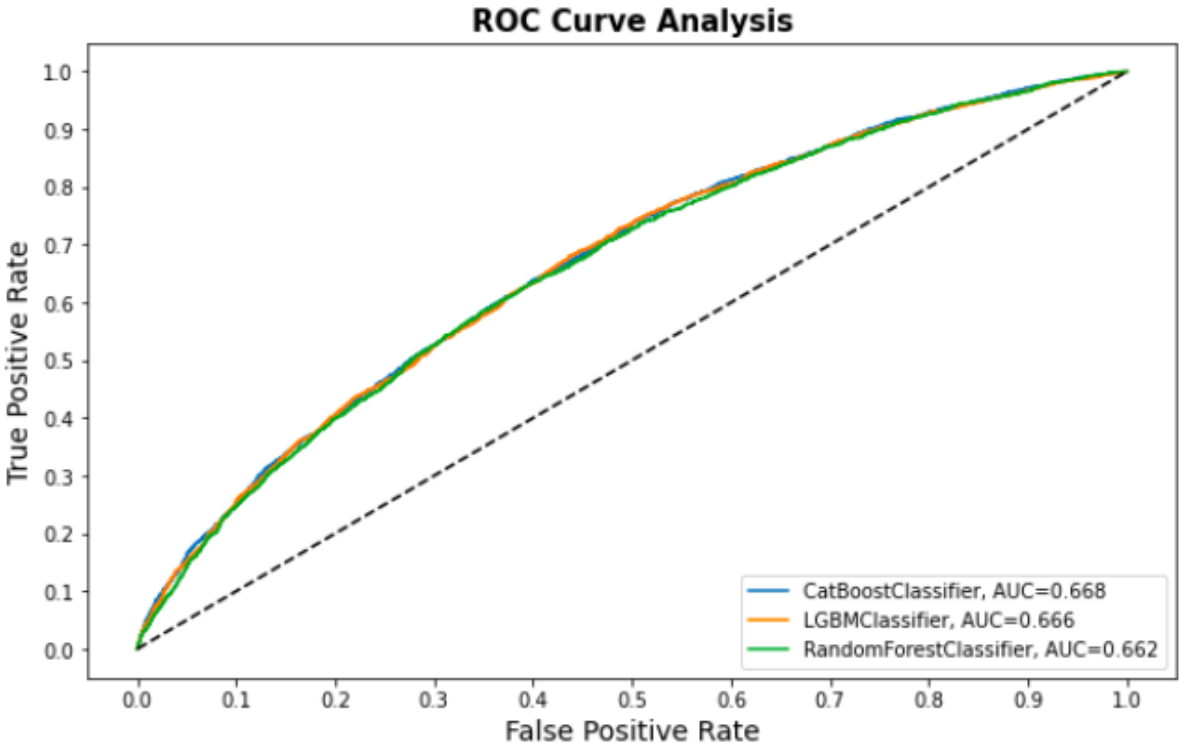

Figure 4. ROC Curve Analysis for RF, LGBM, and CATB models

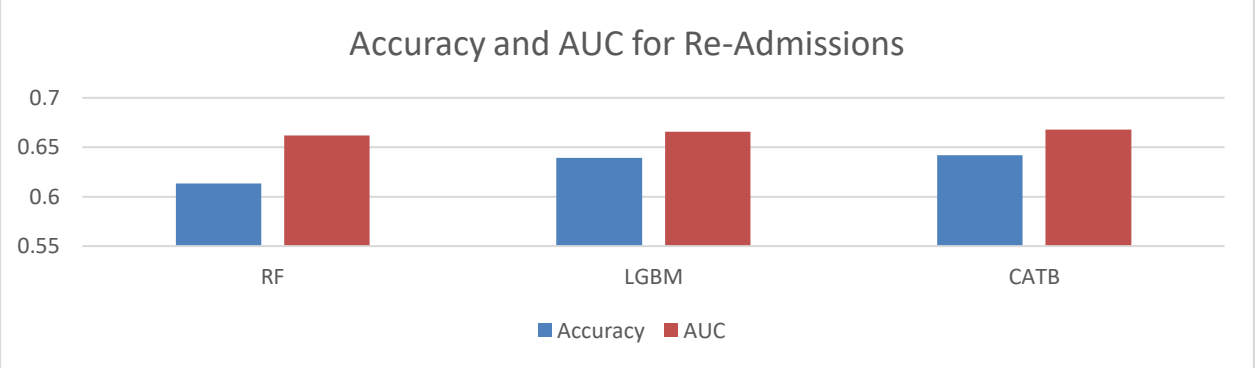

Figure 5. Accuracy and AUC for Readmissions for RF, LGBM, and CATB models

3.2 Discussion

The results from this study demonstrate that **CATBoost** outperformed other machine learning models in predicting hospital readmissions among diabetes patients. One of CATBoost's key strengths is its ability to handle categorical variables directly without needing one-hot encoding (Prokhorenkova et al., 2018). This is especially useful in healthcare datasets, where many features such as race, gender, and discharge disposition are categorical. By leveraging these categorical features more effectively, CATBoost uncovered relationships other models may have missed, contributing to its superior performance.

Another factor that contributed to **CATBoost's** success was its robustness to imbalanced data. In healthcare, the positive class (readmitted patients) is often much smaller than the negative class (non-readmitted patients), creating a class imbalance issue. CATBoost's regularization techniques helped mitigate this imbalance, allowing the model to better distinguish between the two classes (Fernández et al., 2018).

While **LightGBM** and **XGBoost** also performed well, they rely more heavily on feature engineering for categorical variables, which may have limited their performance compared to CATBoost. Though effective in capturing non-linear relationships, Random Forest was prone to overfitting, especially with an imbalanced dataset (Breiman, 2001). **Logistic Regression**, as expected, underperformed compared to more sophisticated models due to its linear nature, which limits its ability to capture complex relationships between features (Hosmer et al., 2013).

Clinical implications

The findings from this study have significant clinical implications. By identifying high-risk diabetes patients before readmission, healthcare providers can implement targeted interventions to prevent unnecessary hospitalizations (Lestari et al., 2022). For instance, patients with frequent inpatient visits or complex medication regimens can be flagged for closer follow-up, medication reviews, and discharge planning (Hansen et al., 2011). Proactive interventions such as remote patient monitoring or telemedicine consultations can also be used to manage patients more effectively in outpatient settings (Basu et al., 2023). Moreover, predictive models such as CATBoost can help healthcare administrators optimize resource allocation by focusing on high-risk patients, thus improving care efficiency while reducing costs (Bates et al., 2014).

Resource allocation

Predictive models like CATBoost can also help healthcare administrators optimize resource allocation. By identifying high-risk patients before they are readmitted, hospitals can allocate specialized resources—such as case managers or patient navigators—to ensure they receive the care and attention they need, potentially improving care efficiency and reducing overall costs (Hansen et al., 2011). Given the growing cost burden of hospital readmissions, particularly for chronic conditions like diabetes, the ability to preemptively target

patients for interventions represents a significant opportunity to enhance the overall effectiveness of healthcare systems (Caron et al., 2013; Khalifa & Zabani, 2016).

4 Conclusion

This study demonstrates the potential of machine learning models in predicting hospital readmissions among diabetes patients. Of the six models compared, CATBoost outperformed the others, achieving the highest accuracy (64.2%) and AUC score (0.70). Key predictive features such as the number of inpatient visits, medications prescribed, and duration of hospital stay were identified as significant contributors to readmission risk. The results suggest that machine learning models, particularly those capable of handling complex and imbalanced healthcare datasets, can provide critical insights into patient care. The implications of these findings extend beyond predictive accuracy. By applying these models in clinical practice, healthcare providers can identify high-risk patients early, enabling them to intervene with targeted care plans that reduce the likelihood of hospital readmissions. Additionally, integrating predictive modeling into hospital workflows could allow for better resource allocation, ensuring that healthcare systems focus on patients most likely to benefit from increased attention and intervention.

Future research

Future research should explore integrating *social determinants of health (SDOH)*, such as income, education, access to healthcare, and social support systems, into predictive models. Incorporating these factors would provide a more comprehensive view of patient health risks and offer deeper insights beyond medical data alone (Artiga & Hinton, 2018). In addition, real-time data from wearable devices and *Internet of Things (IoT)* technologies could significantly improve the predictive power of these models. Continuous monitoring of vital signs and other health metrics through wearables would provide dynamic data on critical indicators such as blood glucose levels, physical activity, and heart rate, which are highly relevant for managing chronic diseases like diabetes (Piwek et al., 2016). Lastly, although complex models like CATBoost exhibit high accuracy, their need for interpretability may limit clinical adoption. Future work should focus on improving model transparency by using explainability techniques such as *SHAP (Shapley et al.)* or *LIME*, enabling healthcare providers to understand better and trust the predictions made by these models (Lundberg & Lee, 2017).

Acknowledgments
I am grateful to two anonymous reviewers for their valuable comments on the earlier version of this paper.

References

Artetxe, A., Beristain, A., & Grana, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer methods and programs in biomedicine*, *164*, 49-64. https://doi.org/10.1016/j.cmpb.2018.06.006

Artiga, S., & Hinton, E. (2018). Beyond health care: the role of social determinants in promoting health and health equity. *Kaiser Family Foundation*, *10*.

Basu, S., Berkowitz, S. A., Davis, C., Drake, C., Phillips, R. L., & Landon, B. E. (2023). Estimated costs of intervening in health-related social needs detected in primary care. *JAMA Internal Medicine*, *183*(8), 762-774.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health affairs*, *33*(7), 1123-1131.

Björk, S. (2001). The cost of diabetes and diabetes care. *Diabetes research and clinical practice*, *54*, 13-18. https://doi.org/10.1016/S0168-8227(01)00304-7

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Caron, F., Vanthienen, J., & Baesens, B. (2013). Healthcare analytics: Examining the diagnosis–treatment cycle. *Procedia Technology*, *9*, 996-1004. https://doi.org/10.1016/j.protcy.2013.12.111

Conget, I. (2002). Diagnóstico, clasificación y patogenia de la diabetes mellitus. *Revista española de cardiología*, *55*(5), 528-535. https://doi.org/10.1016/S0300-8932(02)76646-3

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*(20), 1920-1930.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets* (Vol. 10, No. 2018). Cham: Springer.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Halfon, P., Eggli, Y., van Melle, G., Chevalier, J., Wasserfallen, J. B., & Burnand, B. (2002). Measuring potentially avoidable hospital readmissions. *Journal of clinical epidemiology*, *55*(6), 573-587. https://doi.org/10.1016/S0895-4356(01)00521-2

Hansen, L. O., Young, R. S., Hinami, K., Leung, A., & Williams, M. V. (2011). Interventions to reduce 30-day rehospitalization: a systematic review. *Annals of internal medicine*, *155*(8), 520-528.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Houthooft, R., Ruyssinck, J., van der Herten, J., Stijven, S., Couckuyt, I., Gadeyne, B., ... & De Turck, F. (2015). Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores. *Artificial intelligence in medicine*, *63*(3), 191-207. https://doi.org/10.1016/j.artmed.2014.12.009

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., & Kripalani, S. (2011). Risk prediction models for hospital readmission: a systematic review. *Jama*, *306*(15), 1688-1698.

Khalifa, M., & Zabani, I. (2016). Utilizing health analytics in improving the performance of healthcare services: A case study on a tertiary care hospital. *Journal of Infection and Public Health*, *9*(6), 757-765. https://doi.org/10.1016/j.jiph.2016.08.016

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, *23*(1), 89-109. https://doi.org/10.1016/S0933-3657(01)00077-X

Lestari, Y. D., Armi, A., Koniasari, K., Setiawan, Y., Sartika, M., Rohmah, H. N. F., Nurpratiwi, Y., & Fahrudin, A. (2022). Effectiveness of the emotional freedom techniques to reducing stress in diabetic patients. *International Journal of Health Sciences*, *6*(2), 555–562. https://doi.org/10.53730/ijhs.v6n2.6728

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765-4774).

Mair, C., Kadoda, G., Lefley, M., Phalp, K., Schofield, C., Shepperd, M., & Webster, S. (2000). An investigation of machine learning based prediction systems. *Journal of systems and software*, *53*(1), 23-29. https://doi.org/10.1016/S0164-1212(00)00005-4

McHugh, M. D., Berez, J., & Small, D. S. (2013). Hospitals with higher nurse staffing had lower odds of readmissions penalties than hospitals with lower staffing. *Health Affairs*, *32*(10), 1740-1747.

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, *6*(1), 1-10.

Ogundokun, R. O., Lukman, A. F., Kibria, G. B., Awotunde, J. B., & Aladeitan, B. B. (2020). Predictive modelling of COVID-19 confirmed cases in Nigeria. *Infectious Disease Modelling*, *5*, 543-548. https://doi.org/10.1016/j.idm.2020.08.003

Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. (2016). The rise of consumer health wearables: promises and barriers. *PLoS medicine*, *13*(2), e1001953.

Powers, A. C., & D'Alessio, D. (2016). Endocrine physiology of diabetes. *Diabetes Care, 39*(S1), S1-S102.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, *2014*(1), 781670.

van Walraven, C., et al. (2010). The utility of case-mix adjustment in readmission rate comparisons among hospitals. *BMC Health Services Research, 10*(1), 1-11.

Zhang, H., Huang, M., Yang, J., & Sun, W. (2020). A data preprocessing method for automatic modulation classification based on CNN. *IEEE Communications Letters*, *25*(4), 1206-1210.

Zhang, Z., et al. (2019). Data preprocessing in predictive modeling. *Current Medical Research and Opinion, 35*(4), 655-660.

Zheng, L., et al. (2017). Predicting hospital readmission using machine learning and data mining techniques: A systematic review. *PLoS One, 12*(4), e0174680.

Biography of Author

Alekhya Gandra

Alekhya Gandra is a skilled healthcare analytics and data engineering professional with a strong focus on leveraging machine learning and big data technologies. Her expertise includes advanced tools such as Hadoop, Spark, AWS, and Snowflake, which she uses to optimize healthcare data systems and streamline reporting processes. With a keen interest in applying AI-driven solutions to healthcare, Alekhya is dedicated to improving patient outcomes through predictive modeling and data-driven decision-making. Her research interests revolve around integrating emerging technologies, such as real-time data from wearable devices, to enhance the accuracy and efficiency of healthcare analytics.

Email: alugandra04@gmail.com