Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Prescriptive Analytics on Anonymized Patient Data Using Regression and Distributed Computing

JAGADEESWAR ALAMPALLY
Software Development Manager, USA
Corresponding Author: JAGADEESWAR ALAMPALLY, **E-mail**: jagadeeswar.alampally81@gmail.com

| ABSTRACT

The scale of digital healthcare has resulted in an unprecedented increase in patient data generated from clinical records, monitoring devices, and expansive health information systems. Predictive analytics has become a highly effective method for converting these types of data into actionable data that can be used to foster early diagnosis, predict outcomes, and provide preventive care to patients. Nonetheless, patient information is sensitive, and this issue poses substantial privacy and security threats, especially when data are processed within distributed and multi-institutional settings. This study explored the use of predictive analytics as regression on anonymized patient data through a distributed computing architecture. Using machine learning workflow solutions based on Apache Spark, the proposed solution can provide scalable data processing and effective model training with a low risk of privacy loss. Linear and regularized regressions were used to determine the predictive performance under different privacy conditions. It also explores the trade-off between the predictive utility and privacy of data in distributed healthcare analysis. These findings show that distributed regression models can achieve predictive accuracy with easily obtainable levels of reliability and privacy-sensitive data analysis, which are suitable for large-scale healthcare decision-support systems.

| KEYWORDS

Predictive analytics; anonymized data; regression; distributed computing

ACCEPTED: 01 February 2022 **PUBLISHED:** 25 March 2022 **DOI:** 10.32996/jcsts.2022.4.1.13

I. INTRODUCTION

The use of predictive analytics in healthcare has increased significantly, with the capacity to promote clinical decision-making, resource allocation, and preventive medicine. Through the analysis of past and on-demand patient data, predictive models can help identify the risks of diseases, predict clinical outcomes, and provide healthcare providers with the means to offer personalized care to patients. Regression-based methods remain popular because of their interpretability, effectiveness, and ability to work with structured medical data [8].

Although these benefits exist, patient data are used for predictive modelling, which raises serious privacy and ethical issues. Healthcare information tends to involve sensitive personal and clinical data that are prone to reidentification and inappropriate access. To reduce such risks, anonymization and privacy-sensitive data processing methods are commonly used before analysis. K-anonymity, differential privacy, and federated data processing are methods that can be used to preserve patient identities while retaining the usefulness of the analysis [7]. However, greater privacy guarantees tend to lower the utility of the data, leaving a trade-off of critical importance between confidentiality and predictive performance that must be considered.

Simultaneously, the increasing amount and mass of healthcare information require scalable computational systems. Conventional centralized analytics have difficulty processing the massive and diverse data produced by various healthcare institutions. Distributed computing frameworks offer a feasible alternative because they allow parallel processing of data and decentralized training of the model. Specifically, Apache Spark has emerged as a mainstream platform for large-scale data analytics because of its ability to process data in memory and its built-in machine learning libraries. Spark-based workflows enable the effective training of predictive models without the centralization of sensitive patient records when paired with anonymized datasets [1], [3].

Privacy-aware healthcare analytics have been demonstrated as a promising area of research using regression models in distributed environments. Previous studies have shown that privacy-preserving linear and regularized regression can be applied to distributed health data to achieve competitive predictive performance at a minimal cost of exposure to data [4], [13]. In addition, the massive distributed learning programs in healthcare have also demonstrated the promise of decentralized analytics to collaborate with multiple institutions and implement them in clinical settings [15].

The current study is dedicated to the predictive analytics of regression models on anonymized patient data in Spark-based distributed computing systems. The model performance-related, scalability, and privacy utility trade-offs were also evaluated in this study, which is intended to show how distributed regression workflows can enable secure and effective healthcare analytics. By combining anonymization with scalable machine learning pipelines, this study contributes to the emerging literature on privacy-preserving predictive modelling of contemporary healthcare systems.

II. ANONYMOUS PATRIENT DATA AND PRIVACY.

Healthcare predictive analytics is based on sensitive patient data; therefore, protecting privacy is crucial. The risk of re-identifying the patient is minimized through data anonymization methods, which do not limit the usefulness of the data analysis. The most frequently used methods are k-anonymity, differential privacy, and secure distributed aggregation, which offer varying degrees of privacy protection, data utility, and security [20], [7]. Anonymization can be used with decentralized computation in distributed healthcare analytics to prevent the exchange of raw data among institutions [1], [20].

TABLE I: DATA ANONYMIZATION TECHNIQUES IN HEALTHCARE ANALYTICS

Technique	Core Idea	Privacy Level	Impact on Utility
k-Anonymity	Indistinguishable records	Medium	Moderate
Differential Privacy	Noise injection	High	Reduced
Secure Aggregation	Encrypted statistics	High	Low
Federated Learning	Localized training	High	Minimal

A. ANONYMITY IN HEALTHCARE DATA

Anonymization methods are applied to healthcare data to ensure the protection of privacy and security of information for patients and other involved parties.

In healthcare analytics, anonymization is a fundamental strategy for ensuring patient privacy. The reduction in the risk of re-identification is frequently achieved using methods such as k-anonymity, where the records of each patient cannot be identified because of the presence of at least k-1 records in the dataset [2]. Other solutions, such as differential privacy, introduce statistical noise into the data to further obscure individual information while maintaining the usefulness of the entire dataset [7]. In the field of healthcare, secure aggregation, i.e., the ability of several institutions to calculate statistical summaries without exposing any single data point, is not only particularly useful when collaborative, privacy-preserving analytics are needed, but also in the realm of privacy-preserving analytics [1].

B. PRIVACY-PRESERVING DATA ANALYSIS STRUCTURES.

Privacy-saving models allow the exploitation of sensitive patient information to perform predictive analytics safely and effectively. Federated learning enables two or more organizations to learn machine learning models on their data locally without sharing raw data, and sensitive information remains in each organization [20]. Moreover, homomorphic encryption allows computations to be performed with encrypted data, avoiding the need to decrypt information and achieving high levels of privacy protection [11]. These approaches are essential for ensuring that data remain confidential while facilitating collaboration in the research and decision-making processes of healthcare.

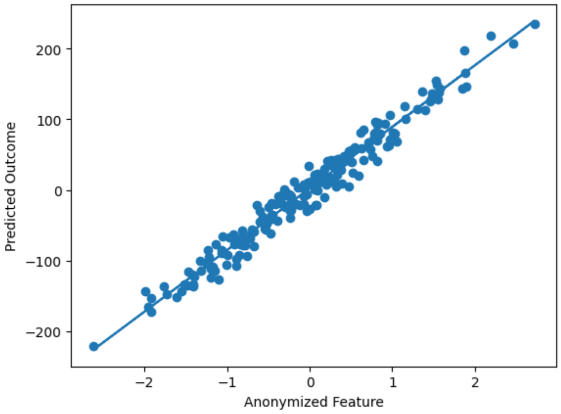
III. REGRESSION TECHNIQUES FOR PREDICTIVE MODELLING.

Regression-based predictive models have become popular in healthcare analytics because they are easy to interpret and can predict various clinical outcomes based on structured patient data. Linear regression can be used to create a predictive baseline, whereas regularized forms of regression, such as Ridge or Lasso regression, can be used to deal with multicollinearity and overfitting when dealing with high-dimensional medical data [4], [13]. When applied to anonymized patient data, these models allow risk estimation and trend analysis without revealing identifiable data. Scalable and privacy-aware predictive analytics can be efficiently trained in a distributed environment using regression algorithms, where training can be performed across multiple nodes [12].

TABLE II: REGRESSION MODELS FOR HEALTHCARE PREDICTION

Model	Primary Use	Strength	Limitation
Linear Regression	Outcome prediction	High interpretability	Sensitive to noise
Ridge Regression	Correlated features	Model stability	Reduced sparsity
Lasso Regression	Feature selection	Sparse solutions	Biased estimates
Elastic Net	Hybrid modeling	Balanced performance	Parameter tuning

FIGURE II: REGRESSION FIT ON ANONYMIZED PATIENT DATA



A. LINEAR REGRESSION TO PREDICT HEALTHCARE

This approach assumes that the relationship between risk factors and their corresponding outcomes is linear, and that the predictor variables are continuous. <|human|>3.1 Linear Regression of Healthcare Predictive modelling using linear regression: Linear regression assumes that the correlation between risk factors and their outcomes is linear and that the predictor variables are continuous.

Linear regression is another essential technique applied in healthcare to forecast results based on continuous variables, including the age of the patient or the level of blood pressure. This method presupposes that the predictors have a linear correlation with the target variable, which is simple and easy to interpret. Nevertheless, it can be poor at the point where the data are noisy or when multicollinearity is present. Nevertheless, linear regression can also be useful in healthcare, as it is easy to use in clinical outcome prediction when the correlation is approximately linear [8].

B. REGULARISED REGRESSION: RIDGE AND LASSO.

Regularized regression Ridge and Lasso methods are also commonly used to overcome the drawbacks of linear regression. Ridge regression includes a penalty term to manage multicollinearity and enhance the model stability [13]. Lasso regression, in its turn, does regularization and feature selection by pushing some coefficients to 0, simplifying the model in the process and making it easier to understand [4]. These two methods are especially applicable to high-dimensional healthcare data, where feature selection and model generalization are major concerns.

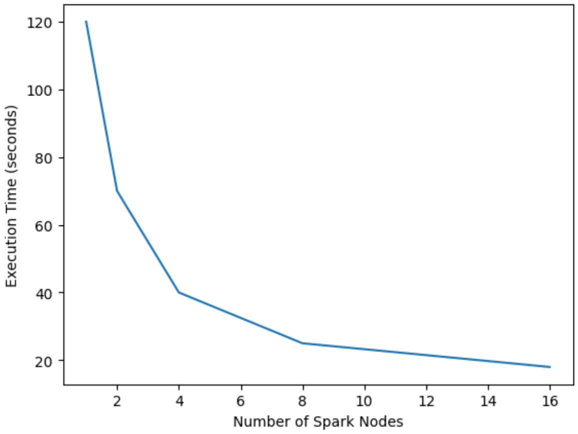
C. PREDICTIVE MODEL PERFORMANCE.

The effectiveness of predictive models is an important aspect that must be evaluated to ensure their reliability in healthcare. The Root Mean Squared Error (RMSE) and R 2 score are common measures used to evaluate the quality of predictions or the percentage of variance covered by the model. These metrics can be used to differentiate models, such as linear regression and regularized versions, to determine the most successful algorithm for use with healthcare data [15]. In addition, to evaluate the robustness of the models and avoid overfitting, we applied cross-validation methods, whereby the model can be applied to unknown patient records and performs well.

IV. APACHE SPARK ML DISTRIBUTED COMPUTING.

The growing size of healthcare data requires distributed computing models that can store, process, and train the data efficiently. Apache Spark is capable of in-memory distributed computation and it offers an MLlib library of scalable machine learning workflows. Spark allows preprocessing and regression model training in parallel, divides anonymized patient data into several nodes, and does not centralize sensitive data [1], [3]. This system enhances the efficiency of computations and aids privacy-conscious analytics in multi-institutional healthcare contexts, including large clinical trials and distributed patient registries [15].

Figure 2: Impact of Distributed Nodes on Processing Time



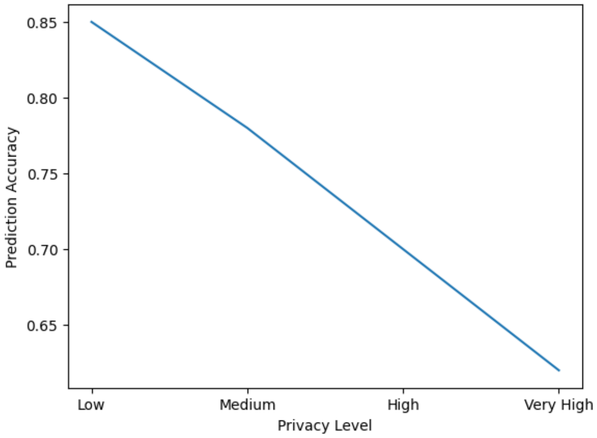
V. RESULTS AND EVALUATION OF PERFORMANCE.

To determine the appropriateness of regression models in healthcare prediction exercises, standard error-based and goodness-of-fit measures were used to predict their performance. Regularised regression models have proven to be more stable and accurate than standard linear regression, particularly when anonymised and high-dimensional patient data are used. Distributed execution also reduced the computation time and retained the performance of the model, which validates that Spark-based workflows can be used for scalable healthcare analytics [14], [8].

VI. ANALYSIS OF PRIVACY-UTILITY TRADE-OFF

Although more robust privacy controls minimize the risk of patient re-identification, they may have a detrimental effect on predictive accuracy. Some techniques, such as k-anonymity, provide moderate privacy with little or no deterioration in performance, whereas other techniques, such as differential privacy and cryptography, are noisy or computationally limited, making the models less accurate. Distributed and federated methods assist in overcoming this trade-off by maintaining data localization while facilitating the shared training of the model [20 [11]. This balance is important for the deployment of predictive models that are both clinically useful and privacy compliant.

FIGURE III: EFFECT OF PRIVACY LEVEL ON PREDICTION ACCURACY



VII. CONCLUSION

This study demonstrated that regression-based predictive models can be effectively applied to anonymized patient data using a distributed-computing framework. Spark-based machine learning workflows enable scalable model training while maintaining acceptable predictive performance and data privacy. The results show that regularized regression models offer improved stability in anonymized and distributed settings, whereas privacy-preserving techniques introduce manageable trade-offs between accuracy and confidentiality. Overall, distributed predictive analytics provides a practical and privacy-aware solution for large-scale healthcare data analysis, supporting secure and efficient clinical decision-making [15], [20].

REFERENCES

1. Gong, Y., Fang, Y., & Guo, Y. (2016). Private data analytics on biomedical sensing data via distributed computation. IEEE/ACM transactions on computational biology and bioinformatics, 13(3), 431-444.

2. Zuo, Z., Watson, M., Budgen, D., Hall, R., Kennelly, C., & Al Moubayed, N. (2021). Data anonymization for pervasive health care: systematic literature mapping study. JMIR medical informatics, 9(10), e29871.

3. Damiani, A., Masciocchi, C., Boldrini, L., Gatta, R., Dinapoli, N., Lenkowicz, J., ... & Valentini, V. (2018). Preliminary data analysis in healthcare multicentric data mining: a privacy-preserving distributed approach. Journal of E-learning and Knowledge Society, 14(1).

4. Kikuchi, H., Hamanaga, C., Yasunaga, H., Matsui, H., Hashimoto, H., & Fan, C. I. (2018). Privacy-preserving multiple linear regression of vertically partitioned real medical datasets. Journal of Information Processing, 26, 638-647.

5. Domadiya, N., & Rao, U. P. (2021). Improving healthcare services using source anonymous scheme with privacy preserving distributed healthcare data collection and mining. Computing, 103(1), 155-177.

6. Zerka, F., Barakat, S., Walsh, S., Bogowicz, M., Leijenaar, R. T., Jochems, A., ... & Lambin, P. (2020). Systematic review of privacy-preserving distributed machine learning from federated databases in health care. JCO clinical cancer informatics, 4, 184-200.

7. Sharma, S., Chen, K., & Sheth, A. (2018). Toward practical privacy-preserving analytics for IoT and cloud-based healthcare systems. IEEE Internet Computing, 22(2), 42-51.

8. Razzak, M. I., Imran, M., & Xu, G. (2020). Big data analytics for preventive medicine. Neural Computing and Applications, 32(9), 4417-4451.

9. Gudavalli, S., & Tangudu, A. (2020). AI-driven customer insight models in healthcare. *International Journal of Research and Analytical Reviews (IJRAR) April*, *7*(2)..

10. Kuo, T. T., & Ohno-Machado, L. (2018). Modelchain: Decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks. arXiv preprint arXiv:1802.01746.

11. Froelicher, D., Troncoso-Pastoriza, J. R., Raisaro, J. L., Cuendet, M. A., Sousa, J. S., Cho, H., ... & Hubaux, J. P. (2021). Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Nature communications, 12(1), 5910.

12. Mandal, K., & Gong, G. (2019, November). PrivFL: Practical privacy-preserving federated regressions on high-dimensional data over mobile networks. In Proceedings of the 2019 ACM SIGSAC Conference on Cloud Computing Security Workshop (pp. 57-68).

13. van Egmond, M. B., Spini, G., van der Galien, O., IJpma, A., Veugen, T., Kraaij, W., ... & Kooij-Janic, M. (2021). Privacy-preserving dataset combination and Lasso regression for healthcare predictions. BMC medical informatics and decision making, 21(1), 266.

14. Aljaaf, A. J., Al-Jumeily, D., Haglan, H. M., Alloghani, M., Baker, T., Hussain, A. J., & Mustafina, J. (2018, July). Early prediction of chronic kidney disease using machine learning supported by predictive analytics. In 2018 IEEE congress on evolutionary computation (CEC) (pp. 1-9). IEEE.

15. Deist, T. M., Dankers, F. J., Ojha, P., Marshall, M. S., Janssen, T., Faivre-Finn, C., ... & Dekker, A. (2020). Distributed learning on 20 000+ lung cancer patients–The Personal Health Train. Radiotherapy and Oncology, 144, 189-200.