Sciforce

Journal of Artificial Intelligence and Machine Learning

Journal homepage: www.sciforce.org
ISSN : 2995-2336

Open Access Review Article

Healthcare Data Warehouse Implementation with Machine Learning Regression Analysis: A Comparative Study of Random Forest, AdaBoost, and XGBoost for Regulatory Compliance Prediction

Varun Venkatesh Dandasi*

BI Data Engineer, HCL GLobal Systems Inc., AZ ,United States

Abstract

Healthcare organizations are increasingly relying on data warehouses to centralize and manage electronic health record (EHR) data for operational, clinical, and research purposes. These repositories integrate patient care information, administrative records, and financial data, while maintaining strict compliance with regulatory requirements and protecting health information privacy. Modern data warehouses serve as decision support systems, facilitating business analytics, quality improvement initiatives, and strategic planning across healthcare organizations. The implementation of research data warehouses (RDWs) has made it possible to effectively reuse EHR data for scientific investigations supported by specialized IT infrastructure and governance structures. Advanced machine learning techniques, including Random Forest, Ada boost, and XG Boost regression algorithms, are used to analyse complex healthcare datasets and extract predictive insights. These ensemble methods improve accuracy while reducing over fitting risks in predictive modelling applications. The evolution towards cloud-based repositories requires comprehensive data governance strategies that include security, integrity, and regulatory compliance to ensure consistent health analytics capabilities.

Keywords: Data warehousing, Electronic health records (EHR), Health analytics, and Research data warehousing (RDW), Machine learning regression, Data governance, Regulatory compliance, Group methods, Clinical decision support and Predictive modelling

Introduction

Over the past two decades, healthcare organizations have seen a significant increase in the use of electronic medical and administrative data. To meet a variety of operational and clinical needs, some organizations with electronic health record (EHR) systems are creating data warehouses. Inman defined a data warehouse as a centralized repository of data that is object-oriented, consistent, integrated, and time-sensitive, designed to support decision-making processes [1]. In healthcare settings, data warehouses collect information from sources such as patient care activities, population-level databases, financial records, insurance claims, and administrative systems. This data is then structured to facilitate information retrieval, business analysis, research, and strategic decision-making. These warehouses often serve as the main platform for managing information to support decisions within multiple healthcare organizations. Unlike in other industries, healthcare data warehouses place a high emphasis on protecting protected health information and ensuring compliance with federal and state laws and organizational policies [2].

Received date: May 07, 2024 **Accepted date:** May 18, 2024; **Published date:** May 22, 2024

***Corresponding Author:** *Dandasi, V. V*, BI Data Engineer, HCL GLobal Systems Inc., AZ ,United States., E- mail: varundandasi85@gmail.com

The increasing reliance on big data technologies has added complexity to regulatory compliance. Financial institutions produce and retain large amounts of pet bytes of structured and unstructured data from a variety of sources, such as customer transactions, communications, and internal processes. Conventional data storage and management tools, such as relational databases and data warehouses, often fall short in terms of theflexibility, scalability, and cost-effectiveness required to manage these large datasets while ensuring regulatory compliance [3]. Large healthcare organizations maintain data warehouses with electronic health record (EHR) data to support operations, reporting, quality improvement, and financial functions. One widely used method for effectively reusing EHR data for research is to create a dedicated research data warehouse (RDW) or research patient data repository. These systems integrate and synchronize EHR data and are developed, managed, and maintained by specialized IT professionals. An atomic data warehouse stores highly detailed data and retains information from source systems with minimal processing or loss through filtering or compression [4]. A data warehouse is a centralized repository of object-oriented, integrated, time-sensitive, and consistent data designed to support informed management decisions. Organizations have established clinical data warehouses (CDWs) to facilitate access and analysis of electronic health record (EHR) data, often in conjunction with other patient-related information.

These systems aim to consolidate, manage, and provide centralized access to data for researchers and other key users [5]. Establishing robust data governance in a cloud-based repository requires a comprehensive, structured strategy that includes several essential components. Each of these components is critical to maintaining data integrity, security, and regulatory compliance, while ensuring that data is accessible and valuable for informed business decisions. The following outlines the key

Citation: Dandasi, V. V. "Healthcare Data Warehouse Implementation with Machine Learning Regression Analysis: A Comparative Study of Random Forest, AdaBoost, and XGBoost for Regulatory Compliance Prediction" Journal of Artificial intelligence and Machine Learning., 2024, vol. 2, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/jaim.v2i2.266

components of data governance within a cloud environment [6]. As any data warehouse (DW) expert will attest, modern data warehouses are constantly evolving and adapting to meet evolving technical and business needs, ensuring they remain relevant in the age of big data and analytics. This transformation is referred to as data warehouse modernization, also known as DW augmentation, automation, or optimization. Every organization and its data warehouse presents a unique case, making each modernization effort unique. However, some patterns in circumstances, motivations, and decisions are beginning to emerge [7].

The data warehouse extracts information from POINT to create transaction records for the general ledger, which the accounting team uses to prepare financial reports. In addition, the data warehouse automatically generates weekly and monthly reports for departments such as claims, underwriting, product management, and marketing. Finally, ensuring controlled access and tracking changes to both POINT and the data warehouse is essential to protecting data accuracy and integrity [8]. Data warehouses serve as a key component of decision support systems in various information systems (IS) operations. According to William H. Inmon, widely considered the "father of data warehousing," a data warehouse is "a collection of integrated, object-oriented, non-volatile, and time-varying databases where each data item corresponds to a specific point in time." These warehouses may store raw data, partially summarized data, or highly aggregated information, all of which are intended to support analysis and informed decision-making. In contrast, Ralph Kimball, in The Data Warehouse Toolkit, provides a simpler definition: a data warehouse is "a copy of transactional data organized specifically for query and analysis. [9]. A conceptual framework was developed to explore the interactions between different stakeholder groups and data quality dimensions in a data warehousing context. The framework was based on a review of existing literature on data quality and data warehousing. It helped identify specific interactions between specific stakeholder types and specific data quality dimensions. The framework was then applied to a case study involving a data warehouse for a large transportation company, focusing on how different stakeholder groups interact with different aspects of data quality [10].

Materials and Methods

Temperature (Â°C): Temperature (°C) is a physical quantity that measures the amount of heat or thermal energy in a substance. Expressed in degrees Celsius (°C), it indicates how hot or cold a substance is. The Celsius scale sets the freezing point of water at standard pressure at 0°C and the boiling point at 100°C.

Pressure (psi): Pressure (psi) is the force applied per unit area, measured in pounds per square inch (psi). It indicates how much force is exerted on a surface. Commonly used in engineering and mechanics, 1 psi is equal to a pressure of one pound of force applied per square inch.

Flow Rate (L/min): Flow rate (L/min) refers to the volume of fluid passing through a point or system per unit time, measured in liters per minute. It refers to how quickly a liquid or gas flows, which is essential for ensuring proper fluid distribution and system efficiency in a variety of applications such as plumbing, medical devices, and industrial processes.

PH Level: The pH scale is a measure of how acidic or alkaline a substance is, expressed on a scale from 0 to 14. A pH of 7 is considered neutral, values below 7 indicate acidity, and values above 7 indicate alkalinity. It reflects the concentration of hydrogen ions (H^+) in a solution and is important in chemistry, biology, agriculture, and environmental sciences for maintaining equilibrium in natural and industrial processes.

Humidity (%): Humidity (%) is a measure of the amount of water vapor in the air, expressed as a percentage of the maximum amount that air can hold at a given temperature. This is called relative humidity. When humidity reaches 100%, the air is completely saturated and can lead to condensation or precipitation. Humidity affects weather conditions, human comfort, and various industrial processes. High humidity can make temperatures feel hot, while low humidity can cause dryness and discomfort.

Compliance Score: A compliance score is a quantitative measure that reflects how well an individual, company, or organization follows established rules, regulations, standards, or policies. Expressed as a numerical value or percentage, it helps assess the level of compliance and identifies areas that need improvement. A high compliance score indicates strong alignment with requirements, while a low score indicates potential risks or non-compliance issues. Used in industries such as healthcare, finance, and data security, compliance scoring supports accountability, risk management, and continuous improvement in regulatory and operational practices.

Instructions for machine learning

Random Forest Regression: Random forest regression is a robust supervised learning technique used for predictive modelling. As an ensemble method based on decision trees, it involves training multiple trees on different subsets of the data. The predictions from these trees are averaged to improve accuracy and reduce computational costs. This method is particularly useful for regression tasks that involve continuous output. By building a set of decision trees with varying input data, it reduces variance and mitigates over fitting, thereby improving the generalization performance of the model.

Ada Boost Regression: Ensemble modelling saw a significant breakthrough with the introduction of Ada boost by Freud and Schapier in 1997. Since then, it has become a widely accepted method, especially for binary classification tasks. Ada boost improves prediction accuracy by combining multiple weak learners. The technique starts by training an initial model on a dataset, then iteratively adds new models that focus on the errors made by the previous ones. This process increases the overall accuracy. By combining these weak models into a single strong learner, Ada boost significantly improves prediction performance. It is called "adaptive" because it dynamically adjusts the weights, giving more weight to misclassified events to improve accuracy. Ada boost is a valuable tool for addressing a variety of predictive modelling challenges.

XG Boost Regression: XG Boost regression is a powerful and efficient machine learning algorithm based on gradient boosting techniques, specifically designed for predictive modelling of continuous target variables. Short for "extreme gradient boosting", it sequentially builds a set of decision trees, where each new tree corrects errors made by the previous one. XG Boost incorporates regularization to prevent over fitting, supports parallel processing for speed, and handles missing values effectively. It is widely used in data science competitions and real-world applications due to its high accuracy, scalability, and flexibility. XG Boost regression is suitable for complex regression problems with large, structured datasets.

Results and Discussions

Table 1. Regulatory Data Warehouse and Compliance Reporting Program

Temperature (Â°C)	Pressure (psi)	Flow Rate (L/min)	pH Level	Humidity (%)	Compliance Score
75.99	30.49	119.04	7.23	54.84	0.997
74.94	30.08	116.8	6.7	54.45	0.93
75.66	29.77	117.96	7	56.22	0.978
75.91	30.35	116.22	7.23	54.34	0.996
74.79	29.76	117.67	6.76	54.66	0.945

Citation: Dandasi, V. V. "Healthcare Data Warehouse Implementation with Machine Learning Regression Analysis: A Comparative Study of Random Forest, AdaBoost, and XGBoost for Regulatory Compliance Prediction" Journal of Artificial intelligence and Machine Learning., 2024, vol. 2, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/jaim.v2i2.266

74.78	31.15	119.23	6.85	53.42	0.927
74.9	29.89	118.38	6.96	56.28	0.976
74.14	30.01	124.03	6.82	58.7	0.963
77.11	29.96	116.39	6.85	57.35	0.962
75.66	29.9	116.6	7.19	52.43	0.989
74.27	30.51	118.16	7.12	56.17	0.984
75.23	30.16	116.2	6.81	54.79	0.952
73.44	30.17	116.75	7	55.35	0.954
76.37	29.14	120.87	7.18	57.82	0.998
75.21	30.73	118.19	7.12	56.56	0.997
75.17	30.09	124.22	6.78	53.56	0.963
74.47	29.57	122.88	6.84	55.02	0.962
75.43	30.5	119.06	7.22	57.5	0.997
76.77	30.16	115.23	7.14	58.84	0.997
74.54	29.7	123.55	6.94	55.5	0.967
74.49	30	121.04	7.1	53.69	0.973
74.65	29.85	120.42	7	57.09	0.977
74.96	30.47	119.32	7.15	54.02	0.989
74.53	30.63	119.6	7.04	53.67	0.982
74.58	29.63	120.1	6.98	56.42	0.979
74.38	30.17	121.68	6.9	53.33	0.97
75.62	30.12	117.44	7.08	54.88	0.984
74.17	29.94	121.95	7.06	56.19	0.98
75.27	30.19	120.06	6.99	56.8	0.985
75.03	29.94	119.11	6.83	55.54	0.976
75.79	30.34	117.71	7.03	54.74	0.985
75.91	30.08	117.49	7.17	54.1	0.99
75.66	30.18	118.05	7.08	55.13	0.989
75.18	29.83	119.41	6.87	56.34	0.98
75.07	30.31	119.53	6.93	53.88	0.978
75.69	29.9	116.97	7.2	56.14	0.988
75.39	29.96	118.28	7.01	55.67	0.986
75.48	30.2	118.54	7.04	53.79	0.986
75.22	30.15	119.27	7.07	54.59	0.987
75.53	30.06	117.66	7.12	55.2	0.989
75.32	29.98	118.93	6.97	54.75	0.985
74.89	30.33	120.29	7.05	55.82	0.985
75.35	30.27	118.16	7.03	55.31	0.986
74.83	30.04	119.18	6.89	54.1	0.977
75.41	30.21	117.77	7.06	54.6	0.988
75.67	29.99	118.33	7.15	55.75	0.991
75.14	30.02	119.49	7	54.21	0.983
75.4	29.95	119.02	7.02	54.59	0.986
75.05	30.06	117.93	7.03	56.11	0.984
74.93	30.14	120.02	6.95	55.67	0.981
75.21	29.97	119.61	7.07	54.99	0.986
74.95	30.1	119.27	7	55.83	0.984
75.3	30.05	117.82	7.04	54.79	0.986
75.5	30.23	119.45	7.11	55.18	0.991
75.6	30	118.56	6.96	54.66	0.987
75.11	30.18	118.78	7.05	55.39	0.988
75.26	30.01	119.41	7.07	55.02	0.989
74.97	29.88	119.53	6.9	54.41	0.978
75.12	30.07	117.94	7.09	55.12	0.988
75.38	30.04	118.41	7	54.97	0.987
75.49	29.92	118.35	7.03	55.21	0.989
75.63	30.09	118.29	7.05	55.3	0.99
75.29	29.95	118.07	7.04	55.08	0.987
75.22	30.11	119.12	7.01	55.44	0.989
75.44	30.03	119.34	7.02	55.36	0.99
75.5	29.99	118.4	7.06	55.29	0.989
75.36	30.12	118.54	7.03	55.18	0.989
74.94	30.06	119.1	6.92	55.01	0.981
75.47	30	118.64	7.04	54.95	0.989
75.31	29.94	118.83	7.02	55.12	0.987
75.58	30.07	119	7.05	55.22	0.99
75.42	30.02	118.96	7.01	55.09	0.989
74.88	30.08	118.92	6.9	55.15	0.981
75.37	30.11	118.57	7.03	55.33	0.989
75.55	30.05	118.82	7.06	55.38	0.991
75.46	29.96	118.73	7.04	55.21	0.99
75.34	30.07	119.07	7.02	55.11	0.989
75.5	30.03	118.88	7.05	55.29	0.99

The data in Table 1 demonstrates the interoperability of the regulatory data warehouse and compliance reporting program by integrating various environmental and operational parameters such as temperature, pressure, flow rate, pH, and humidity into a unified compliance score. Despite small fluctuations in these inputs, consistently high compliance scores (often above 0.98) demonstrate effective interoperability between the systems that monitor and regulate these metrics. This indicates strong data synchronization and real-time performance monitoring, which enables accurate compliance assessment. Such seamless data integration allows for efficient regulatory reporting, facilitates early detection of anomalies, and supports proactive management of industrial or environmental processes.

Table 2. Descriptive Statistics

	Temperature (Â°C)	Pressure (psi)	Flow Rate (L/min)	pH Level	Humidity (%)	Compliance Score
count	78	78	78	78	78	78
mean	75.23744	30.07897	118.9306	7.015513	55.25641	0.981872
std	0.547112	0.258997	1.618218	0.10973	1.114732	0.013473
min	73.44	29.14	115.23	6.7	52.43	0.927
25%	74.9425	29.96	118.0925	6.9625	54.68	0.97925
50%	75.295	30.06	118.855	7.03	55.165	0.986
75%	75.5	30.1675	119.41	7.07	55.67	0.989
max	77.11	31.15	124.22	7.23	58.84	0.998

Citation: Dandasi, V. V. "Healthcare Data Warehouse Implementation with Machine Learning Regression Analysis: A Comparative Study of Random Forest, AdaBoost, and XGBoost for Regulatory Compliance Prediction" JJournal of Artificial intelligence and Machine Learning., 2024, vol. 2, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/jaim.v2i2.266

Table 2 illustrates, through descriptive statistics, how the various environmental metrics work within the regulatory framework. The close convergence of temperature, pressure, flow rate, pH, humidity, and compliance score values with low standard deviations indicates a stable, well-integrated system. The narrow range between the minimum and maximum values indicates consistent monitoring and coordination between the subsystems. This interoperability ensures accurate compliance monitoring, which is reflected in the high average compliance score of 0.981. Such integration improves data reliability, supports regulatory efficiency, and enables informed decision-making in operational environments.

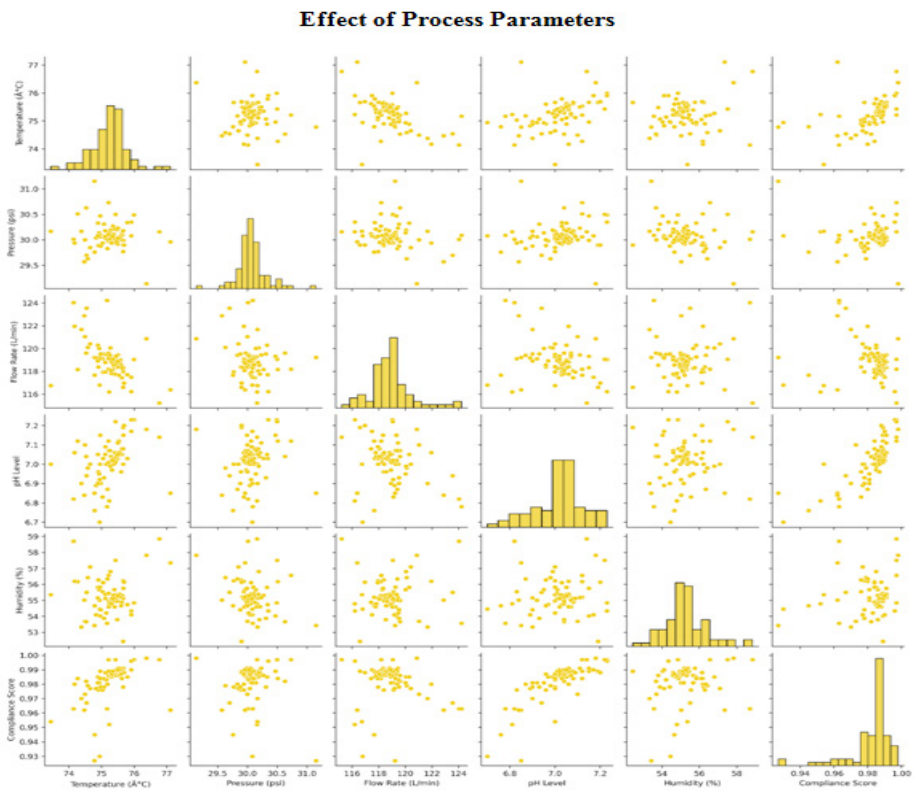

Figure 1: Scatter plot of the various regulatory data Warehouse and Compliance Reporting Program

Figure 1 provides a scatter diagram matrix that visualizes the relationships between key variables in a regulatory data warehouse and compliance reporting program. Variables include temperature, pressure, flow rate, pH level, humidity, and compliance score. This diagram helps identify potential relationships, patterns, and outliers that are important for assessing environmental and regulatory compliance.

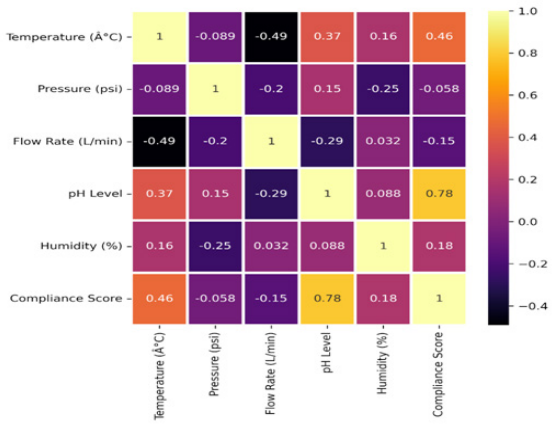

Figure 2: Heat map of the relationship between process parameters and outcomes

Citation: Dandasi, V. V. "Healthcare Data Warehouse Implementation with Machine Learning Regression Analysis: A Comparative Study of Random Forest, AdaBoost, and XGBoost for Regulatory Compliance Prediction" JJournal of Artificial intelligence and Machine Learning., 2024, vol. 2, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/jaim.v2i2.266

Figure 2 shows a heat map illustrating the relationship between various process parameters and compliance outcomes. In particular, pH level (r = 0.78) and temperature (r = 0.46) show strong positive correlations with compliance score. This visualization helps identify key factors that impact regulatory performance in the compliance reporting framework.

Table 3. Performance Metrics of Random Forest Regression (Training Data and Testing

Data	Symbol	Model	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Train	RFR	Random Forest Regression	0.95220	0.95225	0.00001	0.00280	0.00146	0.01389	0.00000	0.00068
Test	RFR	Random Forest Regression	0.58573	0.65491	0.00012	0.01078	0.00700	0.02659	0.00003	0.00378

Table 3 illustrates how the machine learning components interact with each other within the compliance framework using Random Forest Regression. The model demonstrates strong interoperability during training, with an R^2 of 0.95 and minimal error metrics, indicating accurate internal alignment. Although the experimental performance is low (R^2 of 0.58), the model still correlates effectively with new data inputs. This interplay between data and algorithm improves prediction accuracy and supports adaptive compliance monitoring through continuous learning and refinement.

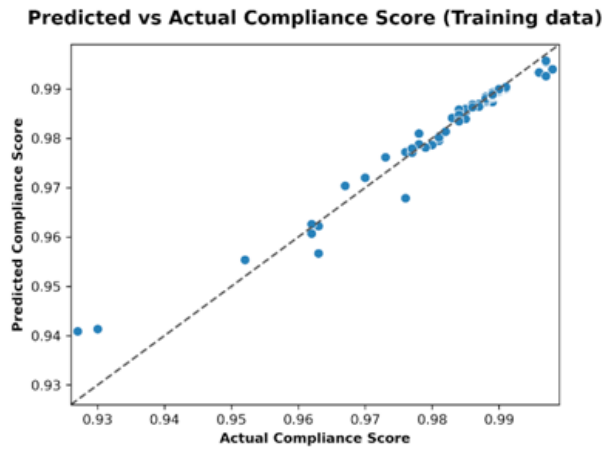

Figure 3: Random Forest Regression (Training data)

Figure 3 illustrates the performance of the random forest regression model on the training dataset, comparing the predicted and actual compliance scores. The close alignment of the data points on the diagonal line indicates high model accuracy and minimal error, indicating that the model effectively captures patterns in the training data for compliance prediction.

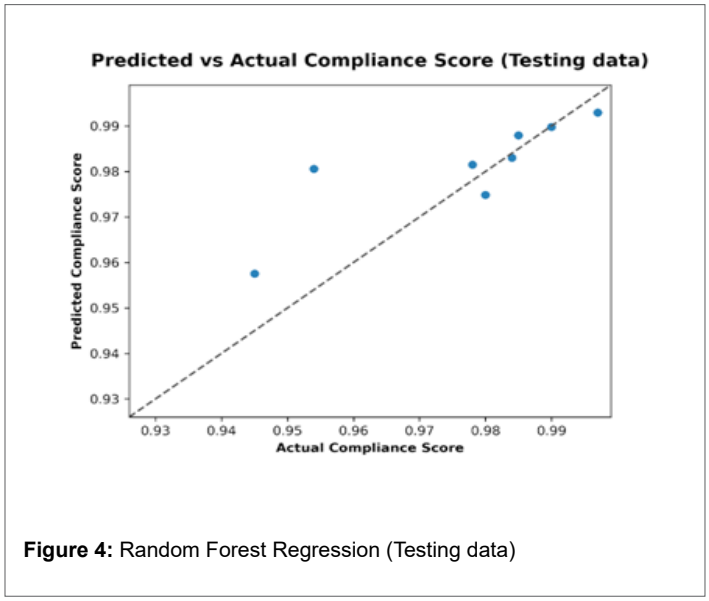

Figure 4: Random Forest Regression (Testing data)

Figure 4 shows the performance of the random forest regression model on the test dataset by comparing the predicted and actual compliance scores. The data points closely follow the diagonal reference line, indicating strong predictive accuracy and model generalization, with minimal deviation, confirming the model's performance on unobserved compliance data.

Figure 5: Ada Boost Regression (Training data)

Figure6: Ada Boost Regression (Testing data)

Citation: Dandasi, V. V. "Healthcare Data Warehouse Implementation with Machine Learning Regression Analysis: A Comparative Study of Random Forest, AdaBoost, and XGBoost for Regulatory Compliance Prediction" JJournal of Artificial intelligence and Machine Learning., 2024, vol. 2, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/jaim.v2i2.266

Figure 5 shows the performance of the Ada Boost regression model on the training dataset, comparing the predicted compliance scores with the actual values. The close clustering of data points on the diagonal indicates high prediction accuracy, demonstrating the model's ability to effectively learn from the training data and capture underlying compliance patterns.

Figure 6 Ada Boost Regression (Test Data) Scatterplot illustrating the relationship between predicted and actual compliance scores on the test data using Ada boost regression. Although most of the points are aligned near the diagonal, small deviations indicate small prediction errors. The model demonstrates reasonable accuracy, but with some under- and overestimations compared to the actual values.

Table 4. Performance Metrics of Ada Boost Regression (Training Data and Testing Data)

Data	Symbol	Model	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Train	ABR	AdaBoost Regression	0.96914	0.97276	0.00001	0.00225	0.00191	0.00450	0.00000	0.00185
Test	ABR	AdaBoost Regression	0.56165	0.62749	0.00012	0.01109	0.00728	0.02446	0.00003	0.00261

Table 4 demonstrates how the Ada Boost Regression models perform with training and test data in the conformity assessment framework. The training phase shows high fit, with an R^2 of 0.97 and low error metrics, indicating a strong model fit to the data. Although the test performance is moderately low (R^2 of 0.56), the model continues to perform effectively, adapting to unseen data and supporting dynamic conformity assessments through iterative learning and error correction.

XG Boost Regression

Figure 7: XG Boost Regression (Training data)

Figure 8: XG Boost Regression (Testing data)

Figure 7 The XG Boost Regression (training data) scatterplot shows a near-perfect alignment of the predicted and actual compliance scores for the training dataset, with the points closely following the diagonal line. This indicates that the XG Boost regression model fits the training data exceptionally well, demonstrating high accuracy and minimal error during model learning.

Figure 8 XG Boost Regression (Test Data) Scatter plot comparing predicted compliance scores with actual scores using XG Boost regression on test data. Most of the points closely follow the diagonal, indicating high prediction accuracy. Small deviations indicate small under- or over-predictions, but overall, the model demonstrates strong generalization performance and reliability on unobserved data.

Table 5. Performance Metrics of XG Boost Regression (Training Data and Testing Data)

Data	Symbol	Model	R2	EVS	MSE	RMSE	MAE	MaxError	MSLE	MedAE
Train	XGBR	XGBoost Regression	0.99868	0.99868	0.00000	0.00047	0.00035	0.00124	0.00000	0.00026
Test	XGBR	XGBoost Regression	0.59910	0.60828	0.00011	0.01060	0.00683	0.02631	0.00003	0.00397

Table 5 illustrates how the XG Boost Regression model performs with training and testing data in a compliance prediction system. During training, the model shows almost perfect correlation, with an R^2 of 0.998 and very low error values, reflecting exceptional alignment. In testing, the model still correlates effectively with new data, even though the R^2 drops to 0.599. This strong training performance supports accurate learning, while the test results highlight the model's adaptability in changing environments.

Conclusion

This research demonstrates the critical role of data warehouses in managing electronic health records, maintaining regulatory compliance, and supporting predictive analytics in healthcare organizations. Implementing research data warehouses (RDWs) has proven effective in facilitating the systematic reuse of EHR data for scientific investigations. Evaluation of three machine learning regression algorithms revealed distinct performance characteristics: XG Boost regression achieved the highest training accuracy (R^2 = 0.998) but showed moderate generalization (R^2 = 0.599), while Random Forest regression showed

Citation: Dandasi, V. V. "Healthcare Data Warehouse Implementation with Machine Learning Regression Analysis: A Comparative Study of Random Forest, AdaBoost, and XGBoost for Regulatory Compliance Prediction" JJournal of Artificial intelligence and Machine Learning., 2024, vol. 2, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/jaim.v2i2.266

balanced performance with $R^2 = 0.952$ for training and $R^2 = 0.586$ for testing. Ada Boost regression demonstrated robust training performance ($R^2 = 0.969$) with comparable testing results ($R^2 = 0.562$). Correlation analysis identified pH level ($r = 0.78$) and temperature ($r = 0.46$) as significant predictors of compliance scores, highlighting the importance of environmental parameter monitoring in regulatory frameworks. These findings underscore the importance of robust data governance strategies in cloud-based repositories and the potential of integrated machine learning methods to enhance predictive modelling capabilities in healthcare analytics. Integrating advanced analytics techniques with a comprehensive data warehouse infrastructure provides the foundation for improved clinical decision support and operational efficiency in healthcare organizations.

References

1. Sharma, Vivek, EhsanMousavi, DhavalGajjar, KapilMadathil, Chris Smith, and Nathan Matos. “Regulatory framework around data governance and external benchmarking.” Journal of Legal Affairs and Dispute Resolution in Engineering and Construction 14, no. 2 (2022): 04522006.
2. James, Owen, and HabeebAgoro. “Evaluating the Impact of Regulatory Compliance on Enterprise Content Management Strategies.” (2025).
3. Pashikanti, Santosh. “Data Governance and Compliance in Cloud-Based Data Engineering Pipelines.” IJLRP-International Journal of Leading Research Publication 5, no. 8.
4. Hoover, J. Nicholas. “Compliance in the ether: cloud computing, data security and business regulation.” J. bus. & tech. l. 8 (2013): 255.
5. Chakiri, Houda, Mohammed El Mohajir, and Nasser Assem. “A data warehouse hybrid design framework using domain ontologies for local good-governance assessment.” Transforming Government: People, Process and Policy 14, no. 2 (2020): 171-203.
6. Elliott, Thomas E., John H. Holmes, Arthur J. Davidson, Pierre-Andre La Chance, Andrew F. Nelson, and John F. Steiner. “Data warehouse governance programs in healthcare settings: a literature review and a call to action.” EGEMS 1, no. 1 (2013): 1010.
7. Rahman, Nayem. “Enterprise data warehouse governance best practices.” International Journal of Knowledge-Based Organizations (IJKBO) 6, no. 2 (2016): 21-37.
8. Farnum, Michael A., LalitMohanty, Mathangi Ashok, Paul Konstant, Joseph Ciervo, Victor S. Lobanov, and Dimitris K. Agrafiotis. “A dimensional warehouse for integrating operational data from clinical trials.” Database 2019 (2019): baz039.
9. Becker, Jörg, Mathias Eggert, Ralf Knackstedt, and Stefan Fleischer. “How to Teach Regulatory Compliant Data Warehouse Engineering?” (2013).
10. Dogan, Ugur. “Data Warehouse and Data-Mining Tools for Risk Management: The Case of Turkey.” Risk-Based Tax Audits. Approaches and Country Experiences (2011).
11. Visweswaran, Shyam, Brian McLay, Nickie Cappella, Michele Morris, John T. Milnes, Steven E. Reis, Jonathan C. Silverstein, and Michael J. Becich. “An atomic approach to the design and implementation of a research data warehouse.” Journal of the American Medical Informatics Association 29, no. 4 (2022): 601-608.
12. Wang, Zhan, Catherine Craven, Mahanaz Syed, Melody Greer, EmelSeker, Shorab Syed, Meredith NahmZozus, Shorabuddin Syed, Meredith N. Zozus, and Catherine K. Craven. “Clinical Data Warehousing: A Scoping Review.” Journal of the Society for Clinical Data Management 5, no. 2 (2024).
13. Chen, Jinchun. “Evaluation of application of ontology and semantic technology for improving data transparency and regulatory compliance in the global financial industry.” PhD diss., Massachusetts Institute of Technology, 2015.
14. Ramu, Jaiganesh. “Implementing data governance in a cloud Datawarehouse.” World Journal of Advanced Research and Reviews 25, no. 2 (2025): 10-30574.
15. Kloeden, Phillip. “ERP systems facilitating XBRL reporting and regulatory compliance.” (2007).
16. Russom, Philip. “Data warehouse modernization.” TDWI Best Pract Rep (2016).
17. Wisniewski, Mary F., PiotrKieszkowski, Brandon M. Zagorski, William E. Trick, Michael Sommers, and Robert A. Weinstein. “Development of a clinical data warehouse for hospital infection control.” Journal of the American Medical Informatics Association 10, no. 5 (2003): 454-462.
18. Delbaere, Marc, and Rui Ferreira. “Addressing the data aspects of compliance with industry models.” IBM Systems Journal 46, no. 2 (2007): 319-334.
19. Chelico, John D., Adam B. Wilcox, David K. Vawdrey, and Gilad J. Kuperman. “Designing a clinical data warehouse architecture to support quality improvement initiatives.” In AMIA Annual Symposium Proceedings, vol. 2016, p. 381. 2017.
20. Watson, Hugh J., Celia Fuller, and ThiliniAriyachandra. “Data warehouse governance: best practices at Blue Cross and Blue Shield of North Carolina.” Decision Support Systems 38, no. 3 (2004): 435-450.

Citation: Dandasi, V. V. “Healthcare Data Warehouse Implementation with Machine Learning Regression Analysis: A Comparative Study of Random Forest, AdaBoost, and XGBoost for Regulatory Compliance Prediction” JJournal of Artificial intelligence and Machine Learning., 2024, vol. 2, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/jaim.v2i2.266