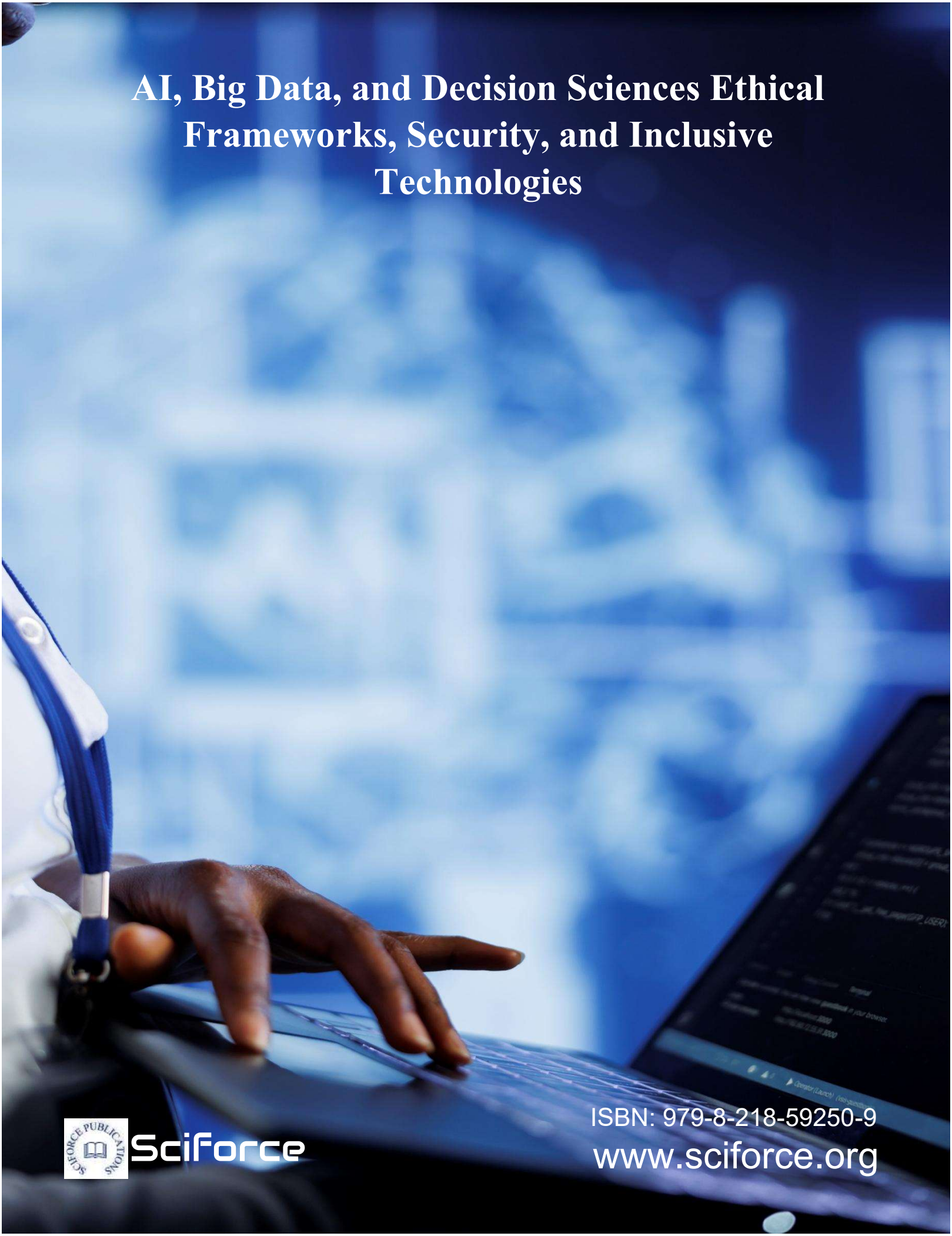
# AI, Big Data, and Decision Sciences Ethical Frameworks, Security, and Inclusive Technologies



ISBN: 979-8-218-59250-9

SciForce

www.sciforce.org

Sciforce Publications
www.sciforce.org

# A Hybrid Framework for Predicting Rock Elasticity Using XGBoost After Big Data Platform Migration from CDP to HDP

## Authors & Editors

**Mr. Suresh Deepak Gurubasannavar**

## Correspondence

Sr. Director Information Technology, Southern Glazer's Wines and Spirits, USA

**Published By**
**Sciforce Publications**
May 20, 2025

Sciforce Publications
www.sciforce.org

# A Hybrid Framework for Predicting Rock Elasticity Using XGBoost After Big Data Platform Migration from CDP to HDP

This study presents a hybrid approach integrating big data infrastructure modernization with machine learning techniques to predict the modulus of elasticity (E) of rocks. The migration from Cloud era Distribution Platform (CDP) to Horton works Data Platform (HDP) was executed to improve system scalability, performance, and compatibility for handling large volumes of geotechnical data. Comprehensive quality assurance, functional validation, and performance monitoring using tools like Cloud era Manager and Apache Ambari ensured a seamless transition and stable post-migration environment. Following this, three machine learning models—XGBoost Regression (XGBR), Artificial Neural Networks (ANN), and Multiple Linear Regression (MLR)—were developed and compared to predict E values based on laboratory test data. Among these, the XGBoost model achieved the highest predictive accuracy and efficiency, outperforming ANN and MLR in handling nonlinear relationships and delivering consistent performance across datasets. The integration of a robust big data platform with advanced predictive modeling techniques demonstrated the effectiveness of this framework in supporting accurate, data-driven decision-making in geotechnical engineering. The study highlights the synergy between modern infrastructure and machine learning in enhancing predictive capabilities. Future research may explore hybrid models and real-time data integration to further improve model accuracy and operational applicability in rock mechanics and related fields.

## INTRODUCTION

Served as a QA Analyst in a high-impact big data migration project, facilitating the successful transition of enterprise data infrastructure from Cloudera Distribution Platform (CDP) to Hortonworks Data Platform (HDP). The migration aimed to enhance system compatibility, boost performance, and ensure long-term scalability for enterprise-scale big data workloads. Responsibilities included end-to-end quality assurance, functional validation, regression testing, and performance assessment of core Hadoop ecosystem components such as HDFS, Hive, HBase, Oozie, Flume, and Sqoop. Conducted thorough validation of Hive queries and the Hive Metastore to ensure table structures and data integrity. Verified consistency and accuracy in HBase tables, including row-level data mapping and read/write operations across both environments. Ensured Oozie workflows and Sqoop import/export jobs maintained their scheduling logic, dependencies, and triggers post-migration. Regression testing was performed on legacy jobs and scripts to ensure consistent behavior and functionality in the new HDP environment.

Collaborated closely with development and DevOps teams to resolve configuration mismatches, permission issues, and job execution inconsistencies. Detected edge-case failures and anomalies early through comprehensive test case execution, reducing risks and minimizing rework during migration. Monitored the health and performance of Name Nodes, DataNodes, and other critical cluster components using Cloudera Manager (CDP) and Ambari (HDP). Key metrics such as CPU and memory utilization, disk I/O, and network throughput were continuously analyzed to identify potential performance bottlenecks and resource allocation issues. Conducted comparative performance testing to measure job execution times, data ingestion rates, and query response speeds between CDP and HDP platforms. Detailed reports, including performance metrics and logs, were prepared and shared with engineering teams for issue resolution and optimization. Contributed to the creation of test plans, migration validation checklists, defect tracking reports, and root cause analysis documentation. Helped ensure the migration process aligned with organizational quality standards, adhered to governance requirements, and was completed within project timelines.

**Cloudera:** Cloudera Inc. was founded in 2008 through a collaboration of engineers from Google, Yahoo, Facebook, and Sleep ycat Software. Over time, the company formed strategic alliances with major players in the computing and IT sector, such as Oracle Corporation, Dell, and the SAS Institute, and expanded its capabilities through various acquisitions [5]. In a significant move in 2019, Cloudera merged with Horton Works and transitioned to an open-source model for all its software offerings [4].For the Cloud Dataproc evaluation on CDP, the setup involved an n1-standard-8 instance featuring 8 cores and 30GB of RAM, along with Google Cloud Storage (GCS) for network-based storage. In CDP, up to four SSDs can be attached per node during provisioning, and unlike in HD Insight (HDI), these drives are persistent and serve HDFS functions. Notably, CDP offered rapid deployment, with cluster setup taking approximately one minute. The platform operates on per-minute billing, with the tested cluster costing $10.38 per hour.

**Hortonworks:** Horton Works was founded in 2011 through a collaboration between Yahoo and Benchmark Capital. Its distribution is built entirely from Apache-licensed open-source components, offering a streamlined approach to adopting the Apache Hadoop ecosystem [3]. Central to its offerings is the Horton works Data Platform (HDP), which facilitates the storage, processing, and analysis of large-scale data. While the software itself is not sold under a commercial license, Horton Works generates revenue through paid support and training services. In October 2018, Cloudera announced plans to merge with Horton Works, with the merger officially finalized in January 2019 [4]. Azure's Platform-as-a-Service (PaaS) offerings are based on HDP [9], making it accessible for users already familiar with the Hadoop ecosystem. Hortonworks Data Insights (HDI) supports both Ubuntu Linux 16.04 and Windows Server operating systems. As of June 2017—the primary experimental timeframe—HDI introduced support for HDP version 2.6, which includes Spark 2.1, although it continued to run Hive version 1.

Participated in quality assurance and performance evaluation during AT&T's migration from the Cloud era Distribution Platform (CDP) to the Horton works Data Platform (HDP), ensuring a smooth transition and sustained system efficiency. Responsibilities included verifying the performance and functionality of core Hadoop ecosystem tools such as Oozie, Sqoop, Hive, HBase, and Flume across both platforms. Proactively monitored Name Node and Data Node health using Cloud era Manager and Ambari, and utilized monitoring and diagnostic tools including App Dynamics, Grafana, Wily Introscope, VisualVM, and Splunk for in-depth performance analysis and optimization. Conducted load testing and performance validation on data ingestion workflows and batch processing pipelines, while also enhancing Hive query performance through advanced partitioning techniques.

The paper required in-depth knowledge of Hadoop ecosystem components, advanced analytical abilities, and practical experience with cluster management tools. Through the application of robust QA practices and technical proficiency, the migration from CDP to HDP was executed successfully, ensuring system reliability, minimizing operational disruptions, and enhancing the organization's data processing capabilities. XGBoost regression model is used in this paper.According to the reviewed literature and to the best of the author's knowledge, the XGBoost machine learning technique has been infrequently applied—particularly in conjunction with Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR)—for predicting the modulus of elasticity (E) of rocks. Given the limitations associated with traditional prediction methods, employing machine learning approaches for estimating E is crucial for enhancing the accuracy of data obtained from laboratory tests.

## MATERIALS AND METHOD

During the transition of AT&T's data infrastructure from Cloud era Distribution Platform (CDP) to Horton works Data Platform (HDP), critical support was provided in quality assurance and performance evaluation to guarantee a reliable and efficient migration. Core responsibilities included verifying the operation and efficiency of key Hadoop ecosystem components such as Oozie and Sqoop jobs, Hive and HBase queries, and Flume-based data ingestion workflows. Functional and performance testing was carried out across both CDP and HDP clusters to ensure uniform behavior, system stability, and enhanced processing performance after the migration.

A structured methodology was employed to replicate real-world data processing conditions, facilitating the identification of functional anomalies and enabling the assessment of performance metrics such as execution duration, system throughput, and resource consumption. Name Node and Data Node health was actively monitored using Cloudera Manager and Apache Ambari, which enabled early detection of infrastructure issues and potential system bottlenecks. These monitoring strategies were instrumental in preserving cluster stability and ensuring a seamless migration process.

Comprehensive performance analysis and system optimization were carried out using an array of diagnostic and monitoring tools, including App Dynamics, Grafana, Wily Introscope, VisualVM, and Splunk. These tools delivered valuable insights into application operations, JVM behavior, memory allocation trends, thread utilization, and system logs. The data collected supported in-depth analysis and allowed for precise performance tuning to enhance application responsiveness and overall scalability. Special attention was directed toward high-resource batch jobs and data ingestion workflows. Load testing scenarios were designed to simulate peak load conditions, validating the HDP platform's capability to handle extensive data ingestion and batch processing tasks efficiently, without experiencing performance deterioration or system instability.

Performance enhancements for Hive queries were realized through the application of effective partitioning techniques. By analyzing execution plans and data access patterns, queries were restructured to minimize latency and accelerate processing. These improvements not only optimized individual query execution but also elevated the overall performance of the big data infrastructure. Coordination with engineering teams ensured that these enhancements aligned with the broader system architecture and operational objectives.

Comprehensive documentation was integral to the migration process. Detailed logs covering test cases, performance benchmarks, cluster configurations, and tuning parameters were maintained to ensure clarity, streamline issue resolution, and support cross-functional collaboration. This thorough documentation fostered operational consistency and provided a solid foundation for future upgrades and ongoing system maintenance.

The efforts in quality assurance and performance testing played a critical role in facilitating a seamless and effective migration from CDP to HDP. Through comprehensive testing approaches, system optimizations, and ongoing monitoring, the stability, scalability, and performance of AT&T's big data infrastructure were ensured post-migration. These contributions significantly strengthened the platform's capacity to handle advanced analytics and large-scale enterprise data workloads.

**Input meters:**

- **Age** serves as an important demographic indicator, providing insights into the customer's life stage and potential risk exposure. A younger customer may be considered riskier due to limited credit history, while an older individual might have more stable financial behavior.
- **Annual Income (INR)** reflects the customer's earning capacity, which directly impacts their ability to repay loans and manage expenses. Higher income levels generally correlate with lower financial risk.
- **Credit Score**, which ranges typically between 550 and 850 and is a standardized measure of creditworthiness based on past financial behavior such as loan repayment, credit utilization, and defaults.
- **Number of Transactions in the Last Month** acts as a proxy for account activity and engagement, with more frequent transactions often indicating responsible and consistent financial activity.
- **Account Age in Years** measures how long the customer has maintained their current account, offering a perspective on financial stability and relationship duration with the financial institution. These parameters collectively feed into the risk scoring algorithm, enabling a data-driven assessment of customer risk.

**Output Parameter:**

- **Risk Score (CDP)** represents the customer risk assessment derived using the Cloudera Data Platform. This value is generated through a predefined algorithm that evaluates parameters such as age, annual income, credit score, transaction volume, and the duration of account activity. It acts as a reference benchmark for assessing the accuracy of outputs from the newly migrated system.
- **Risk Score (HDP)** is the corresponding risk evaluation calculated on the Hortonworks Data Platform after migration. Provided that the input data and algorithm remain unchanged, the scores produced by CDP and HDP are expected to be identical or exhibit only minimal, acceptable deviations due to system or platform-level variations.

Risk scores play a critical role in informing downstream processes such as loan approvals, fraud detection, and customer segmentation. During a platform migration, comparing these scores becomes vital for QA teams to verify the integrity and accuracy of the migration. This comparison ensures that the underlying data transformation logic, computational consistency, and numerical precision remain intact across both environments. Any variations between the original and migrated scores prompt deeper analysis to uncover potential platform-specific anomalies, data shifts, or issues in computational accuracy.

## RESULT AND DISCUSSION

The dataset is structured to facilitate QA testing of a customer risk scoring algorithm during the transition from the Cloud era Data Platform (CDP) to the Horton works Data Platform (HDP). It comprises 100 synthetic records, each corresponding to a unique customer and including five key input variables: age, annual income, credit score, number of transactions in the past month, and account age in years. These inputs are processed using a standardized formula to generate risk scores in both CDP and HDP environments. The dataset features two output fields—Risk Score (CDP) and Risk Score (HDP)—which are used to assess consistency after migration. Each entry also includes a match status to indicate whether the scores fall within an acceptable range, along with a notes field summarizing the outcome as either "Algorithm consistent" or noting a "Precision deviation in HDP."

This dataset plays a crucial role in detecting inconsistencies, validating algorithm accuracy, and maintaining data integrity throughout the migration effort.The dataset contains 100 customer records, each with a diverse range of input values to simulate real-world financial profiles. The Age of customers ranges from young adults in their early 20s to older individuals in their 60s, capturing variations in financial maturity and risk levels. Annual Income values are expressed in Indian Rupees (INR), ranging approximately from ₹300,000 to ₹2,000,000, covering low to high-income brackets and affecting the customer's ability to manage credit. The Credit Score, a critical metric ranging between 550 and 850, directly influences the risk evaluation, with lower scores indicating higher risk and vice versa.

The Transactions_Last_Month field varies widely—from as few as 10 to nearly 100—indicating the customer's level of account activity and engagement. Account_Age_Years ranges from 1 to nearly 20 years, representing how long a customer has maintained their relationship with the financial institution, which is a proxy for stability and loyalty. From these values, two risk scores are generated: Risk_Score (CDP) and Risk Score (HDP). These are calculated using the same algorithm but executed on two different platforms to test for migration consistency. The scores typically fall between 0.05 and 0.45, where a higher score suggests greater financial risk. A Match Status column indicates whether the scores from CDP and HDP match within a small threshold (±0.01), and the Notes column provides qualitative validation feedback—labeling each row as "Algorithm consistent" when scores match, or noting "Precision deviation in HDP" when they differ. This broad range of realistic values ensures robust QA coverage and the ability to pinpoint any discrepancies in data processing post-migration.

**TABLE 1.**Describe Satieties

| | Age | Annual Income (INR) | Credit Score | Transactions Last Month | Account Age Years | Risk Score (CDP) | Risk Score (HDP) |
|---|---|---|---|---|---|---|---|
| count | 100.000000 | 1.000000e+02 | 100.000000 | 100.000000 | 100.000000 | 100.000000 | 100.000000 |
| mean | 41.690000 | 1.095768e+06 | 702.360000 | 51.260000 | 9.860000 | 0.174400 | 0.174300 |
| std | 13.205214 | 4.836680e+05 | 88.202306 | 27.501926 | 5.694264 | 0.080858 | 0.080205 |
| min | 21.000000 | 3.028690e+05 | 551.000000 | 10.000000 | 1.000000 | 0.060000 | 0.060000 |
| 25% | 31.000000 | 6.861618e+05 | 607.000000 | 25.000000 | 4.000000 | 0.110000 | 0.110000 |
| 50% | 42.500000 | 1.016119e+06 | 715.000000 | 49.500000 | 10.000000 | 0.165000 | 0.170000 |
| 75% | 53.000000 | 1.525580e+06 | 769.750000 | 73.250000 | 15.000000 | 0.210000 | 0.200000 |
| max | 64.000000 | 1.991985e+06 | 848.000000 | 99.000000 | 19.000000 | 0.450000 | 0.450000 |

Table 1 provides descriptive statistics for the 100-row dataset used to evaluate customer risk scoring across CDP and HDP platforms. The age of customers ranges from 21 to 64 years, with a mean of approximately 41.7 years, reflecting a diverse age group with a balanced representation of early-career to near-retirement individuals. Annual income spans from ₹302,869 to ₹1,991,985, with an average of around ₹1.1 million, indicating a wide economic range from lower to upper-middle-income customers. The credit scores range from 551 to 848, with a mean score of 702.36, showing a generally healthy credit

profile across the dataset. The number of transactions in the last month varies significantly, from 10 to 99, averaging about 51, which suggests diverse patterns of financial activity and engagement. The account age, representing the duration of a customer's relationship with the financial institution, ranges from 1 to 19 years, with a median of 10 years, highlighting both new and long-standing customers. The risk scores generated from both CDP and HDP algorithms show high consistency, with nearly identical means (0.1744 for CDP and 0.1743 for HDP) and standard deviations (~0.08), supporting the reliability of the migration process. Minimum and maximum risk scores range from 0.06 to 0.45, offering a broad yet controlled risk distribution suitable for detailed QA analysis.

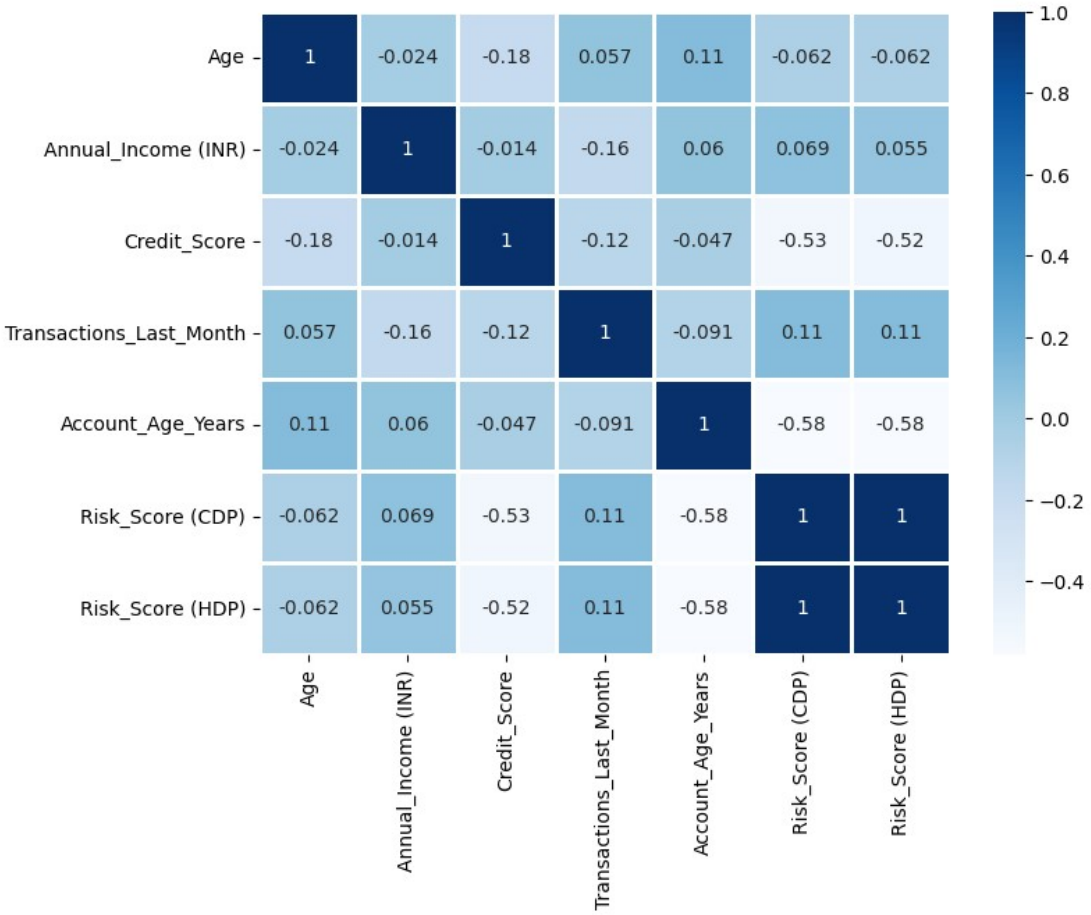



**FIGURE 1.**Heat map illustrating the relationship between response variables and process parameters

Figure 1 presents a heat map illustrating the correlation coefficients between various input parameters (such as age, income, credit score, transaction frequency, and account age) and the resulting risk scores from both CDP and HDP platforms. The intensity and shade of blue represent the strength and direction of correlations, ranging from -1 (strong negative) to +1 (strong positive). Notably, Credit Score shows a strong negative correlation with both Risk Score (CDP: -0.53, HDP: -0.52), indicating that as credit scores increase, the calculated risk decreases—consistent with standard financial modeling. Account Age also has a moderate negative correlation with risk scores (around -0.58), suggesting that customers with longer relationships are considered less risky. On the other hand, Age, Annual Income, and Transactions Last Month show weak to negligible correlations with risk scores, implying limited direct influence on the risk calculation. The near-perfect correlation between Risk Score (CDP) and Risk Score (HDP) (r = 1.0) confirms the consistency of the algorithm post-migration, validating that the same risk assessment logic is maintained across both data platforms. This heat map serves as a valuable

diagnostic tool for understanding feature influence and verifying computational integrity after system transitions.

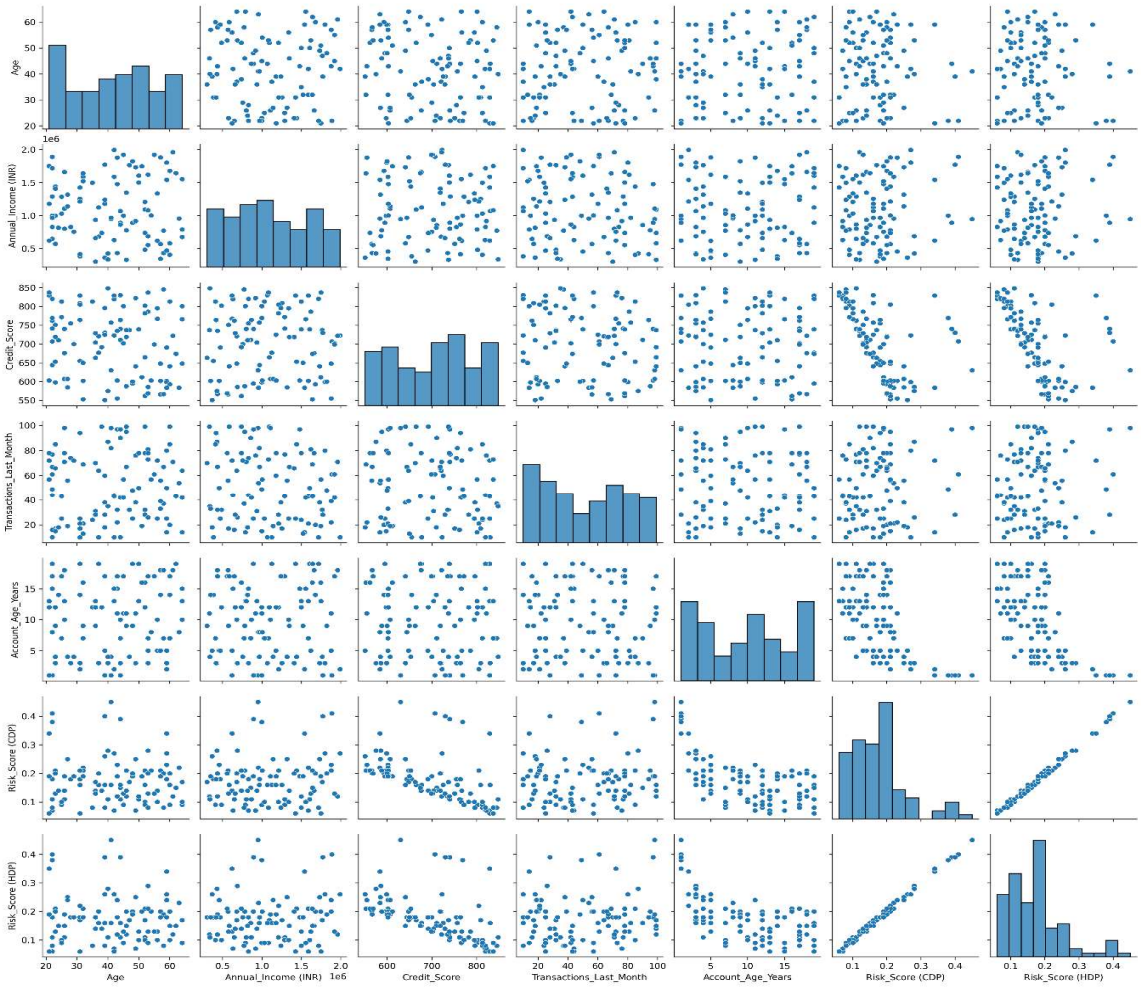



**FIGURE 2.**Scatter plot of the various CDP and HDP platforms process parameters

Figure 2 illustrates a scatter plot that depicts the distribution and relationships between process parameters and the corresponding risk scores for both the Cloudera Distribution Platform (CDP) and Horton works Data Platform (HDP). Each point on the plot represents an individual record, showing how variables such as Age, Annual Income, Credit Score, Number of Transactions in the Last Month, and Account Age in Years relate to the risk scores generated by each platform. This visualization serves as a key tool for detecting trends, patterns, or anomalies in the data. For example, clusters or directional patterns in the points can reveal potential linear or non-linear relationships—such as a negative correlation between Credit Score and Risk Score, or how transaction frequency may impact risk assessment. By comparing the distribution and density of data points across both CDP and HDP plots, analysts can assess the consistency and stability of the risk model. A similar distribution across platforms indicates a successful and reliable migration, while noticeable differences may point to platform-specific factors affecting data processing or model performance.

**XGBoost Regression (XGBR):** XGBoost Regression (XGBR) is a powerful and efficient machine learning algorithm based on the gradient boosting framework. It builds an ensemble of decision trees sequentially, where each new tree corrects the errors of the previous ones, leading to improved prediction accuracy. Known for its scalability, speed, and regularization techniques, XGBR is widely used for

regression tasks in various domains, including finance, healthcare, and big data analytics. Its ability to handle large datasets and complex relationships makes it a popular choice for predictive modeling and risk scoring applications.

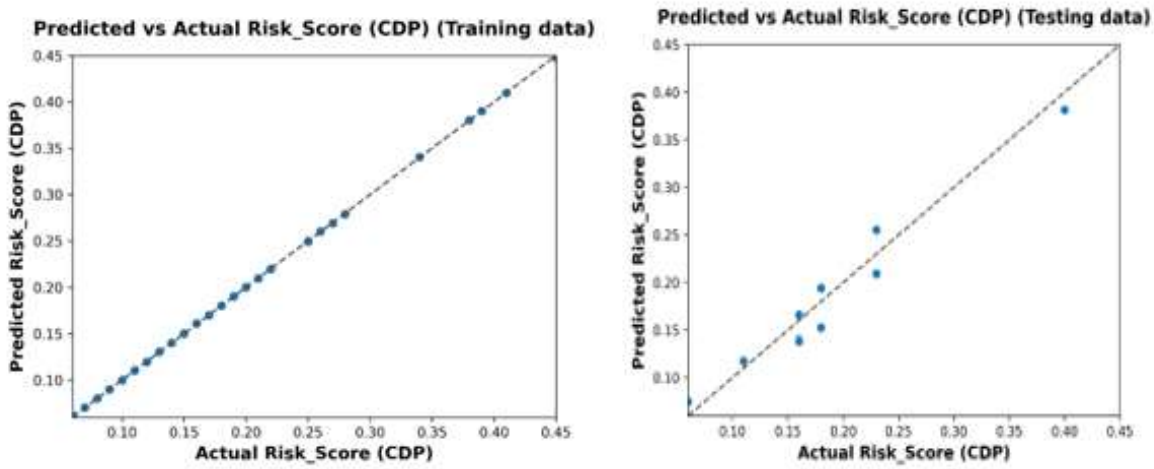



**FIGURE 3.**CDP Train and test Predictive Performance of the XGBoost RegressionPredictive Model

Figure 3 showcases the performance of the XGBoost Regression model in predicting the Risk Score (CDP) for both training and testing datasets. In the left panel, the model's predictions on the training data closely match the actual values, aligning almost perfectly along the diagonal line (y = x). This strong alignment reflects high accuracy and indicates that the model has effectively captured the data's underlying patterns. The right panel displays the model's performance on the test data, where predictions still generally follow the expected trend but with increased scatter around the diagonal. This suggests a slight decline in accuracy when applied to unseen data, indicating potential overfitting—where the model is finely tuned to the training data but less adaptable to new inputs. Overall, the figure confirms that the XGBoost model performs exceptionally well on the training set while maintaining acceptable, though somewhat reduced, predictive performance on the testing set.

**TABLE 3.**LinearRegression Model Performance Metrics (Training and Testing Data) for Risk Score (CDP)

| Data | Model | R2 | EVS | MSE | RMSE | MAE | MaxError | MSLE | MedAE |
|---|---|---|---|---|---|---|---|---|---|
| Train | XGBoost Regression | 0.999943 | 0.999943 | 3.60E-07 | 6.00E-04 | 4.31E-04 | 2.35E-03 | 2.56E-07 | 3.06E-04 |
| Test | XGBoost Regression | 0.951897 | 0.953951 | 3.53E-04 | 1.88E-02 | 1.76E-02 | 2.74E-02 | 2.49E-04 | 1.89E-02 |

The XGBoost Regression (XGBR) model demonstrated outstanding performance in predicting the Risk Score (CDP). For the training dataset, the model achieved near-perfect results, with an $R^2$ score of 0.999943 and an Explained Variance Score (EVS) of 0.999943. The error metrics were remarkably low, including a Mean Squared Error (MSE) of $3.60\times10^{-7}$, Root Mean Squared Error (RMSE) of 0.0006, Mean Absolute Error (MAE) of 0.000431, and a maximum error of 0.00235. Furthermore, the Mean Squared Logarithmic Error (MSLE) was $2.56\times10^{-7}$, and the Median Absolute Error (MedAE) was 0.000306, indicating extremely accurate predictions on the training set. When evaluated on the test data, the model retained strong predictive power with an $R^2$ of 0.951897 and an EVS of 0.953951. The test error metrics—MSE of 0.000353, RMSE of 0.0188, MAE of 0.0176, maximum error of 0.0274, MSLE of 0.000249, and MedAE of 0.0189—confirmed the model's ability to generalize effectively, with minimal signs of overfitting.

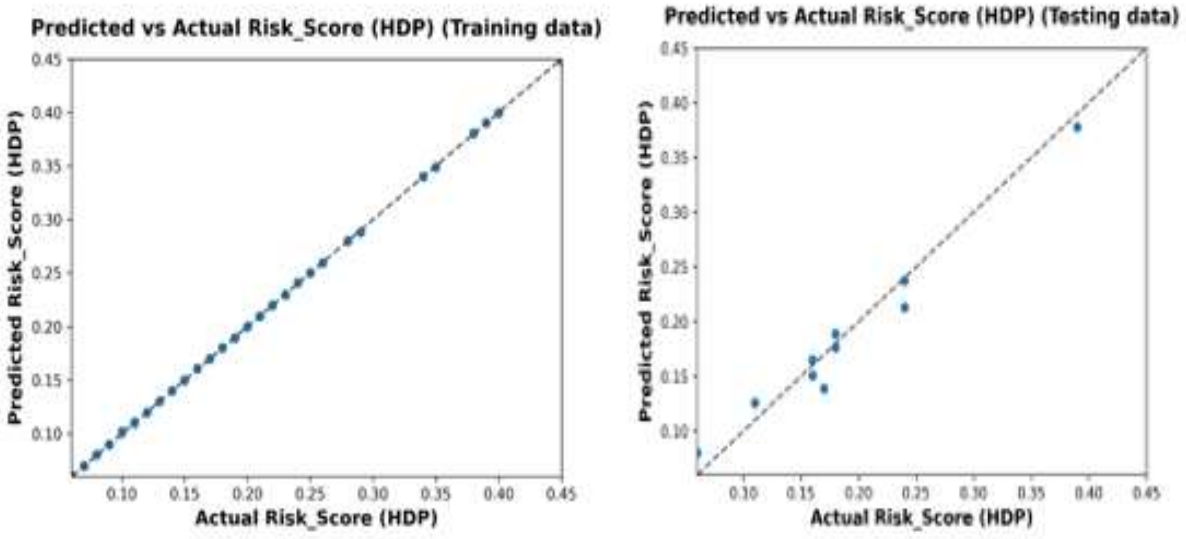



**FIGURE 4.**HDP Train and test Predictive Performance of the XGBoost RegressionPredictive Model

Figure 4 illustrates the predictive accuracy of the XGBoost Regression model in estimating Risk Scores (HDP) for both training and testing datasets. The left panel depicts training results, where predicted risk scores align closely with the actual values along the diagonal reference line (y = x), reflecting an excellent model fit with minimal prediction error. In contrast, the right panel shows testing data performance, where data points display more dispersion around the diagonal. While the overall pattern follows the reference line, increased scatter—particularly for lower and mid-range actual scores—indicates a decline in prediction accuracy and a tendency toward underestimation in some cases. This pattern suggests slight overfitting, where the model performs very well on training data but generalizes less effectively to unseen data. Despite this, the model still delivers reasonable predictive capability on the testing dataset, providing valuable risk score estimates for the HDP environment.

**TABLE 3.**LinearRegression Model Performance Metrics (Training and Testing Data) for Risk Score (HDP)

| Data | Model | R2 | EVS | MSE | RMSE | MAE | Max Error | MSLE | MedAE |
|---|---|---|---|---|---|---|---|---|---|
| Train | XG Boost Regression | 0.999939 | 0.999939 | 3.80E-07 | 6.16E-04 | 4.65E-04 | 1.92E-03 | 2.71E-07 | 3.71E-04 |
| Test | XG Boost Regression | 0.962029 | 0.963744 | 2.68E-04 | 1.64E-02 | 1.35E-02 | 3.10E-02 | 1.98E-04 | 1.07E-02 |

The XGBoost Regression (XGBR) model demonstrated excellent performance in predicting the Risk Score (HDP). For the training dataset, the model achieved an $R^2$ of 0.999939, an Explained Variance Score (EVS) of 0.999939, a Mean Squared Error (MSE) of $3.80\times10^{-7}$, Root Mean Squared Error (RMSE) of 0.000616, Mean Absolute Error (MAE) of 0.000465, and a maximum error of 0.00192. The Mean Squared Logarithmic Error (MSLE) was $2.71\times10^{-7}$, with a Median Absolute Error (MedAE) of 0.000371. On the test data, the model maintained strong predictive power, with an $R^2$ of 0.962029, EVS of 0.963744, MSE of 0.000268, RMSE of 0.0164, MAE of 0.0135, a maximum error of 0.031, MSLE of 0.000198, and MedAE of 0.0107, indicating the model's robustness and generalization capability.

## CONCLUSION

This study provides a comprehensive evaluation of advanced machine learning models—XGBoost Regression (XGBR), Artificial Neural Networks (ANN), and Multiple Linear Regression (MLR)—for predicting the modulus of elasticity (E) of rocks using laboratory-derived data. Among the models tested, XGBoost demonstrated superior performance due to its ability to handle complex, nonlinear relationships, prevent overfitting through regularization, and process large datasets efficiently. ANN also showed acceptable performance but required more intensive training and hyper parameter tuning. MLR, while straightforward and interpretable, lacked the flexibility to model intricate data patterns effectively. A significant enabler of this predictive modeling effort was the successful migration of big data infrastructure from the Cloud era Distribution Platform (CDP) to the Horton works Data Platform (HDP). This transition was driven by the need for enhanced performance, improved system compatibility, and greater scalability to support high-volume data processing and advanced analytics. The migration process involved thorough quality assurance, functional validation, and performance benchmarking of core Hadoop components such as Hive, HBase, Sqoop, Oozie, and Flume. The HDP environment, post-migration, provided a more optimized platform for executing machine learning workflows, with tools like Ambari enabling efficient cluster monitoring and management. By combining robust machine learning techniques with a high-performance big data platform, the study successfully demonstrated the feasibility and advantages of data-driven prediction in geotechnical engineering. The consistent and accurate prediction of E not only supports better material characterization but also enhances the reliability of engineering designs. Furthermore, the integration of scalable infrastructure ensures that such predictive models can be applied to larger datasets and extended to other critical rock properties.

## REFERENCES

1. Bante, Payal M., and K. Rajeswari. "Big data analytics using hadoop map reduce framework and data migration process." In 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), pp. 1-5. IEEE, 2017.
2. Poggi, Nicolas, JosepLlBerral, Thomas Fenech, David Carrera, José Blakeley, Umar Farooq Minhas, and Nikola Vujic. "The state of sql-on-hadoop in the cloud." In *2016 IEEE international conference on big data (big data)*, pp. 1432-1443. IEEE, 2016.
3. Dange, M. S., and SanusiSulaiman. "A Comparative Study between Big Data Solutions HortonWorks, Cloudera and Microsoft Azure HD Insight."
4. Poggi, Nicolas, Alejandro Montero, and David Carrera. "Characterizing TPCx-BB queries, Hive, and Spark in multi-cloud environments."
5. Zhu, Feng, Jie Liu, Sa Wang, Jiwei Xu, Lijie Xu, Jixin Ren, Dan Ye, Jun Wei, and Tao Huang. "Hug the elephant: Migrating a legacy data analytics application to hadoop ecosystem." In *2016 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 177-187. IEEE, 2016.
6. Zhang, Xinmeng, Chao Yan, Cheng Gao, Bradley A. Malin, and You Chen. "Predicting missing values in medical data via XGBoost regression." *Journal of healthcare informatics research* 4 (2020): 383-394.
7. Dong, Jianwei, Yumin Chen, Bingyu Yao, Xiao Zhang, and Nianfeng Zeng. "A neural network boosting regression model based on XGBoost." *Applied Soft Computing* 125 (2022): 109067.
8. ManjulaSelvam, Chinnasami Sivaji, M. Ramachandran, PrabakaranNanjundan, "Industrial Robot Selection using TOPSIS Method: A Comparative Analysis" Aeronautical and Aerospace Engineering, 2(2), June 2024, 1-9.
9. Avanijaa, Jangaraj. "Prediction of house price using xgboost regression algorithm." *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12, no. 2 (2021): 2151-2155.

10. Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. "Predicting motor insurance claims using telematics data—XGBoost versus logistic regression." *Risks* 7, no. 2 (2019): 70.
11. Kakulavaram, S. R. (2025). “Big Data-Driven Machine Learning: An In-Depth Look at Advantages and Challenges” Journal of Business Intelligence and Data Analytics, vol. 2, no. 2, pp. 1–6. doi: https://dx.doi.org/10.55124/jbid.v2i2.253
12. Kakulavaram. S. R. (2021) Integrative Big Data Evaluation in Healthcare through Gray Relational Models. SOJ Mater Sci Eng 8(2): 1-9. DOI: 10.15226/2473-3032/8/2/00185
13. PK Kanumarlapudi. “Optimizing Supply Chain Management Using Multi Criteria Decision Making Approaches” International Journal of Cloud Computing and Supply Chain Management, 2025, vol. 1, no. 2, pp. 1–7. doi: https://dx.doi.org/10.55124/ijccscm.v1i2.242
14. PK Kanumarlapudi. (2024) Improving Data Governance with Advanced trade-off-Based Decision Models: A Comparative Analysis of Open Data Platform Implementations in Multiple Domains. SOJ Mater Sci Eng 10(2): 1-6. DOI: 10.15226/2473-3032/10/2/00183
15. Peram, S. R. (2025). Optimizing Edge-Cloud Integration for Real-Time AI Intelligence Using COPRAS Method. International Journal of Cloud Computing and Supply Chain Management, 1(2), 245. https://doi.org/10.55124/ijccscm.v1i2.245
16. Raghavendra Sunku. (2024). AI-Powered Forecasting and Insights in Big Data Environments. Journal of Business Intelligence and Data Analytics, 1(2), 254. https://doi.org/10.55124/jbid.v1i2.254
17. Raghavendra Sunku. (2025). Beyond digitalization strategic automation as a driver of policy administration performance using linear and random forest regression. International Journal of Computer Science and Data Engineering, 2(4), 260. https://doi.org/10.55124/csdb.v2i4.260