East Journal of Engineering
ISSN: 3079-9376

Volume 1 , Issue 4
Publisher: East Publication & Technology
DOI: https://doi.org/10.63496/eje.Vol1.Iss4.197

Hybrid AI-Based Sales Forecasting Framework

Ahmad Mousa*[1], Ammar Shaheen

[1]Department of Control and Robotics Engineering, Faculty of Engineering and Technology, Al-Zaytona University of Science and Technology, Salfit Street, Al-Laban Al-Sharqiya - Salfit, Palestine, Ahmad11.ne333@gmail.com.
[2]Department of Control and Robotics Engineering, Faculty of Engineering and Technology, Al-Zaytona University of Science and Technology, Salfit Street, Al-Laban Al-Sharqiya - Salfit, Palestine, Ammarshaheen663@gmail.com.
*Corresponding Author.

Received: 01/08/2025, **Revised**: 20/08/2025, **Accepted**: 25/08/2025, **Published**: 26/08/2025

***Abstract*:**

This research explores the utility of advanced artificial intelligence strategies to enhance sales data analysis in the hard financial context of Palestine. Given the area's chronic political instability, resources, and infrastructure constraints, the research demonstrates how AI-driven models can provide actionable insights to help sustainable commercial enterprise decision-making. The proposed technique integrates three middle algorithms: Linear Regression, Random Forest Regression, and K-Means Clustering. These are mixed into a unique Hybrid Model that leverages supervised and unsupervised learning strategies. Empirical effects showed that whilst Linear Regression carried out properly in shooting linear patterns (R^2 = 0.819), and Random Forest provided slight effectiveness in modeling non-linear relationships (R^2 = 0.727), the Hybrid Model outperformed each. By using K-Means to segment the dataset and undertaking localized regression within each cluster, the Hybrid Model achieved superior overall performance, with an R^2 of 0.895 and an RMSE of 260.15. This approach achieves stronger prediction accuracy and advanced version interpretability via accounting for data heterogeneity. The findings spotlight the sensible value of hybrid AI strategies in complex records environments. Beyond income prediction, such fashions may be extended to domains like finance, healthcare, and client segmentation. For small and medium organizations in resource-restrained areas, adopting hybrid AI strategies gives a path towards extra operational resilience and strategic agility.

***Keywords*: Artificial Intelligence, Hybrid Model, Sales Forecasting, K-Means Clustering, Random Forest, Linear Regression.**

1. Introduction

In the latest fast-paced and ever-evolving worldwide economic system, the mixing of present-day analytical tools has emerged as the most effective and essential for organizations striving to maintain competitiveness and enhance strategic decision-making techniques. This necessity is in particular obvious in economically challenged areas, consisting as Palestine, in which businesses face severe internal and external pressures, along with political instability, restricted get proper of access to sources, marketplace constraints, and infrastructural limitations. Under these hard conditions, companies have to leverage all to be had resources to optimize their operations, enhance resilience, and make certain long-time period sustainability [1].

One of the most precious properties in this context is profits or revenue information, which serves as a key universal performance indicator for groups and gives a comprehensive view of market dynamics and client behaviors. Product overall performance, sales generation, and purchaser choices, thereby permitting them to make data-driven decisions that are essential in highly competitive and volatile markets. In essence, powerful use of income information offers not only a photograph of beyond overall performance but additionally a roadmap for future strategic moves [2].

With the appearance of Artificial Intelligence (AI), statistical evaluation has undergone a large transformation. AI brings with it the potential to process widespread amounts of facts unexpectedly and with a degree of precision that traditional statistical strategies often lack. Its core technology, together with machine gaining

knowledge of, enables the discovery of hidden styles, correct prediction of destiny effects, and automation of complicated decision-making techniques. These competencies make AI a necessary tool for contemporary businesses, especially those running in dynamic and uncertain environments [3].

Machine learning, an important element of AI, includes algorithms that improve automatically through experience and the continuous processing of records. Among those algorithms, several have been validated as specifically effective when applied to structured income records to gain actionable insights. For instance, Linear Regression is a broadly used predictive approach that identifies the linear relationships between unbiased variables, together with the amount or unit price, and an established variable like profit. By fitting an exceptional-fit line to historic information, it enables corporations to make informed predictions approximately future performance [4].

Another valuable device is Random Forest Regression, an ensemble mastering technique that builds multiple decision trees at some point in the process and merges their effects for extra accurate and sturdy predictions. This technique is especially beneficial in environments with non-linear information patterns or when interactions amongst more than one variable should be considered. For Palestinian organizations, Random Forest Regression can reveal deeper insights into how mixtures of factors influence profitability or sales outcomes. In comparison to the predictive focus of regression techniques, K-Means Clustering is an unmonitored clustering algorithm that businesses data into clusters primarily based on similarity. When applied to income information, this approach can section clients consistent with shopping behavior, discover styles in product income, or classify days with comparable sales developments. Such segmentation can be instrumental for corporations aiming to customize their advertising techniques or optimize resource allocation [5].

In the context of Palestinian businesses, adopting these AI-powered gear affords an extensive competitive facet. By the usage of Linear Regression, agencies can forecast key metrics which including profit or income quantity, with more confidence. Random Forest Regression enables the uncovering of complicated, non-obvious relationships in the data, allowing for extra informed strategic choices. Meanwhile, K-Means Clustering aids in know-how the composition of customer bases or figuring out sales cycles, permitting better-centered commercial enterprise movements. To help with this look at, the dataset was gathered from one of the nearby agencies in Palestine. The statistics consist of key income metrics along with quantity, unit price, total income, and income. For confidentiality purposes, the business enterprise's name is withheld, and all information has been anonymized to meet moral research requirements [6].

Moreover, the accessibility of that technology is growing, enabling small and medium-sized firms (SMEs) to include advanced analytics in their day-to-day operations. With the developing availability of open-source libraries and cloud-based services, technical boundaries are being reduced, making it viable for even modestly resourced groups to deploy these AI-pushed techniques. This trend is especially useful in Palestine, where innovation and flexibility are important to overcoming monetary and operational demanding situations.

These tools not only best beautify operational performance but also aid greater tailored advertising efforts. By understanding which customer corporations respond more positively to precise product offerings or pricing techniques, agencies can maximize return on funding and foster stronger customer loyalty. Over time, integrating AI models into enterprise workflows can cause smarter planning, improved stock control, and extra strategic promotional campaigns [7].

The cause of this takes a look at is to discover the utility and results of those AI strategies on sales data from a Palestinian business agency. Specifically, the have a look at aims to evaluate how efficiently Linear Regression, Random Forest Regression, and K-Means Clustering may be used to forecast customer behavior, compare profitability metrics, and apprehend the interrelationships amongst key variables such as amount, unit pricing, overall sales, and income. Through the application of those models, the take a look at affords concrete examples

of the way corporations in economically limited settings can leverage AI to benefit actionable insights and foster sustainable growth [8].

The algorithms examined in this have a look at include: Linear Regression, Random Forest Regression, and K-Means Clustering, which have been carried out as examples to illustrate the capacity of AI-based strategies in reading and extracting insights from dependent sales records [9].

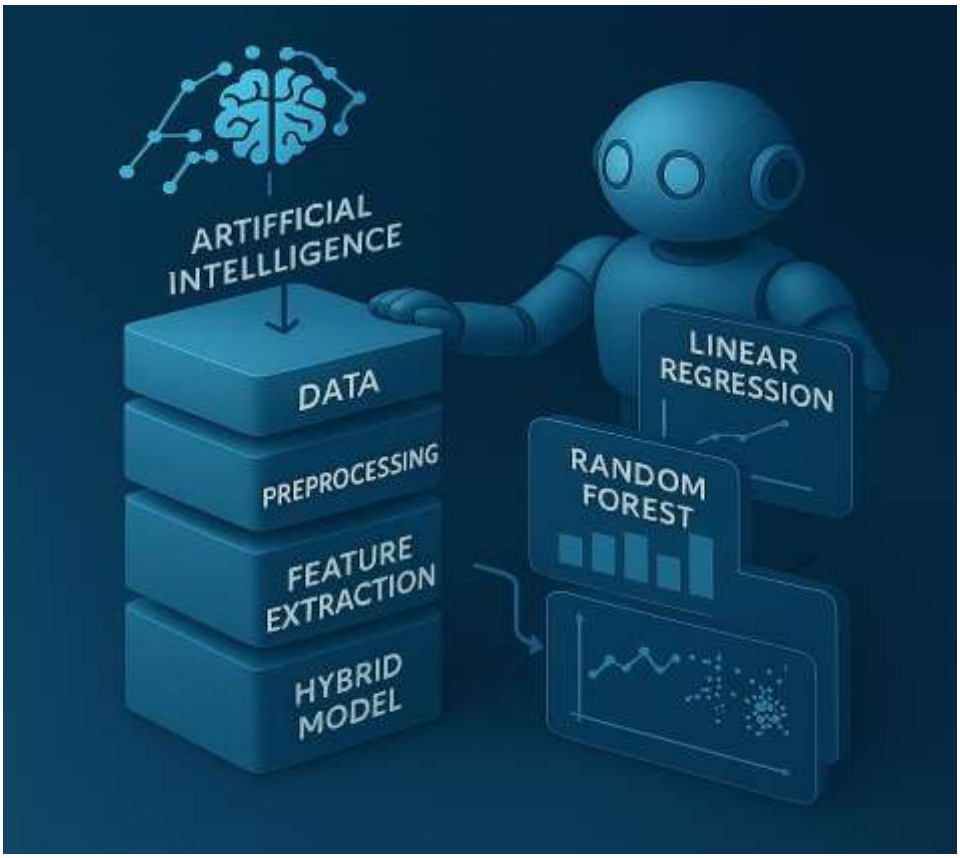

Figure 1: Conceptual overview illustrating the AI-driven data analysis framework incorporating robotics and machine learning algorithms.

2. Literature Review

2.1 K-Means Clustering

K-Means clustering remains a cornerstone approach in unsupervised gaining knowledge of, broadly applied to partition datasets into homogeneous subgroups primarily based on feature similarity. Its simplicity and scalability have made it a device of desire throughout domains regarding high-dimensional or voluminous records.

Recent advancements have focused on enhancing clustering efficiency and precision. For instance, Moodi and Saadatfar (2023) proposed an optimized K-Means variant tailored for big data environments [10]. Their algorithm, which incorporated stabilized data point anchoring and refined distance computations, demonstrated up to a 41.85% improvement in clustering quality.

In the realm of edge computing, Awad and Hamad (2022) introduced a distributed K-Means framework utilizing neural processing units within smartphones [11]. This innovation doubled processing speed relative to conventional implementations, highlighting the potential of decentralized AI in real-time analytics.

The versatility of K-Means is further illustrated by Chen (2022), who applied the algorithm to consumer segmentation in marketing, facilitating precision-targeted strategies [12]. Similarly, Wang and Luo (2023) implemented K-Means in academic settings to stratify student performance data, enabling data-driven educational interventions [13].

2.2 Random Forest

Random Forest, a strong ensemble learning algorithm, has gained prominence for its capability to deal with complex, high-dimensional, and non-linear record structures. It aggregates multiple-choice timber to provide greater solid and correct predictions, correctly mitigating overfitting.

In the field of precision medicine, Hu and Szymczak (2022) investigated the application of Random Forest models in analyzing longitudinal clinical data [14]. Their approach demonstrated high adaptability in tailoring treatment pathways, reflecting the algorithm's capacity to manage time-dependent variables.

Further showcasing its diagnostic power, Zhang et al. (2022) employed Random Forest in the detection of mild cognitive impairment (MCI), attaining a sensitivity of 94.44% and a specificity of 100%. In industrial applications [15], Zhou et al. (2024) integrated Random Forest into smart maintenance systems, reporting a 9.36% improvement in green infrastructure performance management [16].

2.3 Linear Regression

Linear Regression continues to serve as a foundational method for modeling relationships between continuous variables, prized for its interpretability and mathematical simplicity.

Yu et al. (2022) introduced an integrative framework combining Linear Regression with statistical clustering to evaluate academic outcomes [17]. Their results emphasized the utility of regression in quantifying learning disparities and forecasting student performance trends.

Recent developments also advocate for hybridized approaches, where Linear Regression is embedded within localized clusters to enhance prediction accuracy. Zafar et al. (2023), for example, demonstrated that combining K-Means clustering with Linear Regression improved the model's sensitivity to intra-group variations, thereby reducing residual error across diverse datasets [18].

2.4 Hybrid Modeling Approach

To address the limitations of global models in heterogeneous data environments, recent studies advocate hybrid modeling strategies that integrate clustering with regression analysis.

This multi-stage approach typically begins with K-Means clustering to segment data based on intrinsic structural features such as seasonal fluctuations, pricing behavior, or transaction patterns. Each cluster then serves as a domain for training specialized regression models—often Random Forest or Linear Regression—to capture localized relationships.

The outputs from these cluster-specific models are subsequently aggregated to generate a composite prediction. This method has been shown to improve overall forecasting accuracy, reduce model bias, and enhance interpretability.

As highlighted by Zafar et al. (2023) [19]and Yu et al. (2022) [20], the hybrid paradigm offers a nuanced solution to modeling in complex environments by marrying the generalization ability of supervised learning with the adaptability of unsupervised segmentation.

3. Related Work

Hybrid fashions that integrate clustering and regression techniques have garnered considerable interest in recent years. Unlike traditional fashions that follow a single predictive technique on the entire dataset, hybrid processes leverage clustering as a preprocessing step to handle record heterogeneity, thereby improving prediction accuracy. This section critiques recent research (2020–2024) that specifically recognizes hybrid methodologies, just like the one followed in this research.

3.1 Hybrid Models in Energy Forecasting

Ding et al. (2022) proposed a comprehensive hybrid method for brief-term energy intake forecasting in a green construction context. They have a look at the combined three clustering algorithms (K-Means, K-Medians, Hierarchical Clustering) with four regression models (LASSO, SVR, ANN, XGBoost) to address statistical variability [21]. The consequences indicated that the aggregate of Hierarchical Clustering with XGBoost yielded the very best accuracy, outperforming individual fashions by about 15%. The authors concluded that selecting the optimal quantity of clusters notably affects version performance, highlighting the significance of balancing accuracy and computational value.

3.2 Hybrid Techniques in Port Logistics

Ruiz-Aguilar et al. (2020) developed a hybrid version to predict container extent in health ports, combining Self-Organizing Maps (SOM) with SVR and SARIMA models [22]. The hybrid method has been shown to significantly improve prediction accuracy, especially in minimizing MSE, MAE, and MAPE metrics, compared to non-hybrid fashions. The observer verified that clustering heterogeneous records earlier than applying regression helps isolate patterns that can be lost in worldwide fashions.

3.3 Hybrid Approaches in Public Health Analytics

In a more current study, hybrid models have been applied to investigate stunting incidence in Aceh, Indonesia. The model integrated SVM, Linear Regression, and an optimized K-Medoids algorithm to predict health results primarily based on demographic statistics. The results confirmed that clustering using K-Medoids appreciably improved the accuracy of regression models, highlighting the importance of selecting strong clustering techniques while managing complicated fitness statistics.

3.4 Optimizing Clustering Techniques in Hybrid Models

García-Ordás et al. (2024) performed a comparative evaluation of 4 clustering strategies (K-Means, Agglomerative, Gaussian Mixture, Spectral Clustering) inside a solar thermal power plant. The examination evaluated the impact of clustering on version accuracy using unsupervised quality metrics and regression error measures [23]. The consequences showed that Gaussian Mixture Models (GMM) were especially effective in cases with overlapping statistical distributions, suggesting that probabilistic clustering can enhance the model's overall performance in strength packages.

3.5 Summary and Insights

The reviewed studies illustrate that integrating clustering with regression notably improves predictive accuracy across various fields, from electricity forecasting to healthcare analytics. These hybrid fashions leverage information segmentation to reduce inner variability, allowing for more focused and correct regression.

However, a key project remains: the choice of clustering method can considerably affect the outcome. Some studies have shown that deterministic clustering (like K-Means) might not always be ultimate, especially when statistical clusters overlap. Therefore, the present-day research targets to address this gap by way of integrating K-Means with Linear Regression and Random Forest, balancing simplicity and robustness in data analysis.

4. Methodology

4.1 Data Preprocessing

Data preprocessing is crucial to ensure certain statistics are first-rate and consistent before model training. The dataset used in this study underwent the following preprocessing steps:

Handling Missing Values: Using imputation techniques to fill gaps.

Normalization: Numerical attributes, including Quantity, Unit Price, and Total Sales, were standardized to preserve regular scaling, which is critical for clustering.

Encoding Categorical Variables: Transforming categorical data using one-hot encoding.

Outlier Detection and Removal: Identifying and dealing with values that could negatively impact model precision.

4.2 Exploratory Data Analysis (EDA)

To explore characteristic relationships and shape within the information, a couple plot and a correlation heatmap were generated. Figure 2 revealed strong visible correlations, mainly between Quantity and Total Sales, and between Total Sales and Profit, confirming the intuitive business common sense in the back of the dataset.

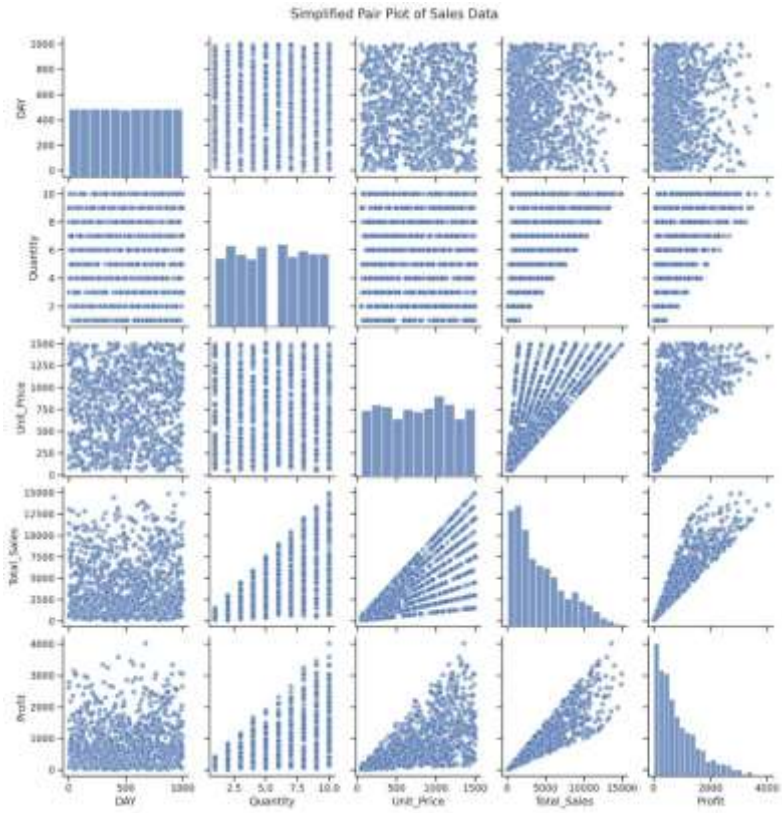

Figure 2: Pair Plot displaying distributions and relationships between Quantity, Unit Price, Total Sales, and Profit.

The heatmap quantitatively showed the strongest correlations and highlighted slight to susceptible relationships regarding Unit Price, which informed the choice of input variables for the regression models as seen in figure 3.

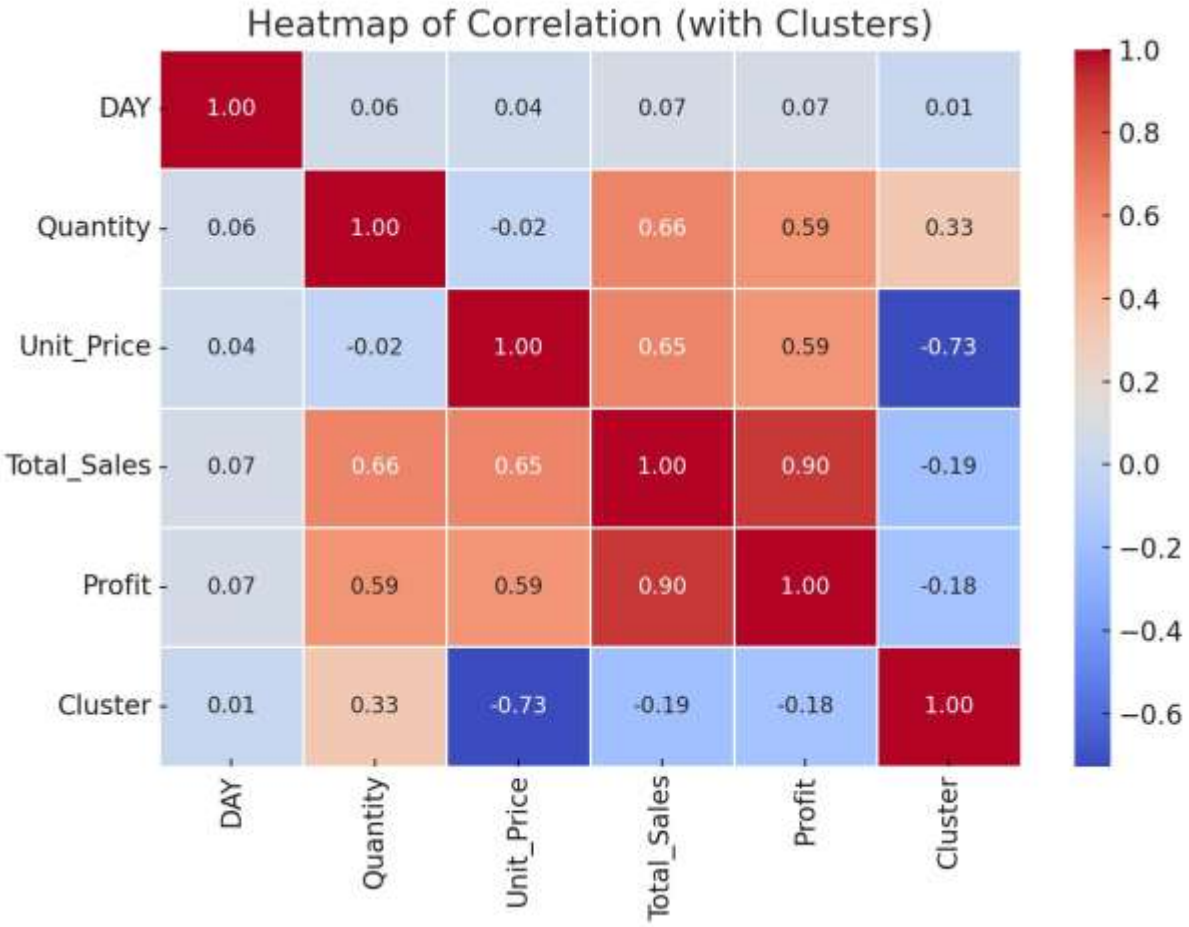

Figure 3: Correlation Heatmap illustrating Pearson correlations amongst numerical variables.

4.3 Clustering Approach: K-Means

To address the heterogeneity inside the records, K-Means Clustering turned into implemented. This method clusters data points into clusters based on similarity, allowing for more refined model training within each group.

Mathematical Formulation of K-Means:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left| x_i^{(j)} - \mu_j \right|^2 \qquad (1)$$

Where, k represents the number of clusters, n represents the number of data points $x_i^{(j)}$ represents the data point belonging to cluster j, and μ_j represents centroid of cluster j.

Where the input features, illustrating the separation of information points into distinct clusters based on similarity see figure 4.

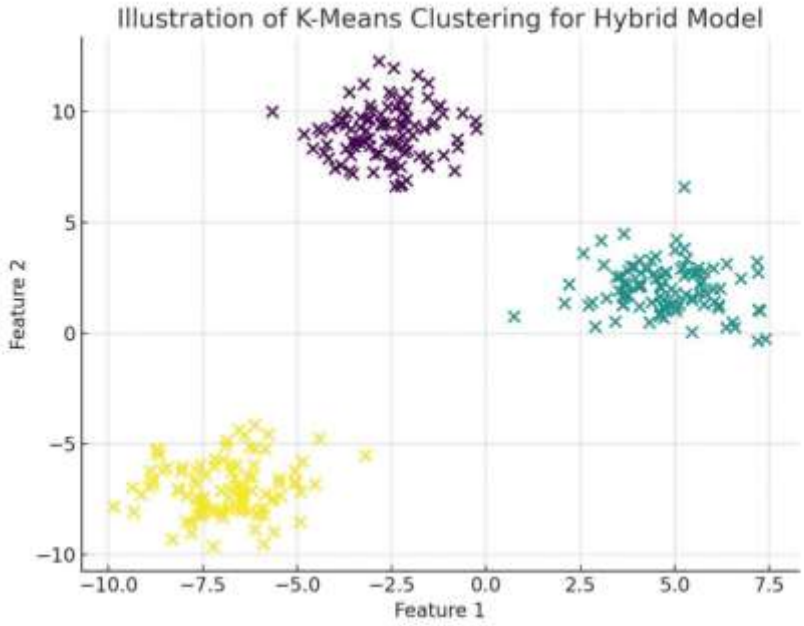

Figure 4: K-Means clustering applied to the hybrid model's

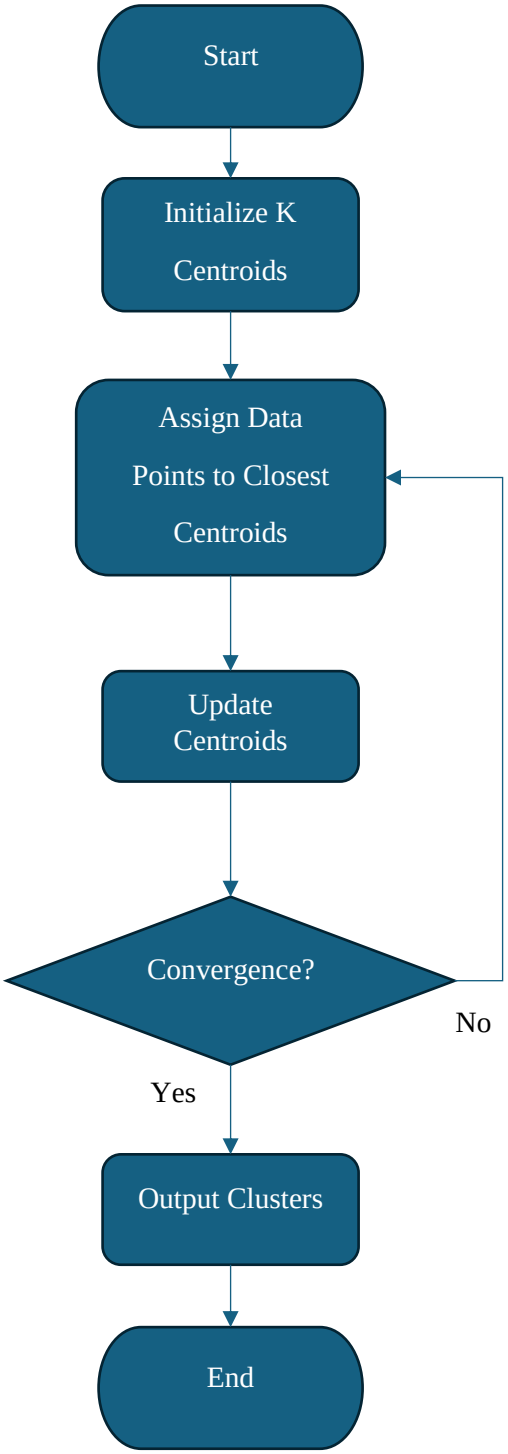

Figure 5: Flowchart illustrating the iterative process of the K-Means clustering algorithm used for data segmentation.

4.4 Regression Models

Two regression models were chosen due to their complementary strengths:

Linear Regression: Suitable for linear relationships.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \ldots + \beta_n X_n + \varepsilon \qquad (2)$$

Random Forest Regressor: Efficient for capturing non-linear patterns.

$$\hat{Y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \qquad (3)$$

Where:

$\hat{Y}$: Predicted value , β_0 = Intercept

β_n: Coefficients

X_n: Independent variables

ε: Error term

T: Number of decision trees

$f_t(x)$: Prediction from each tree

.

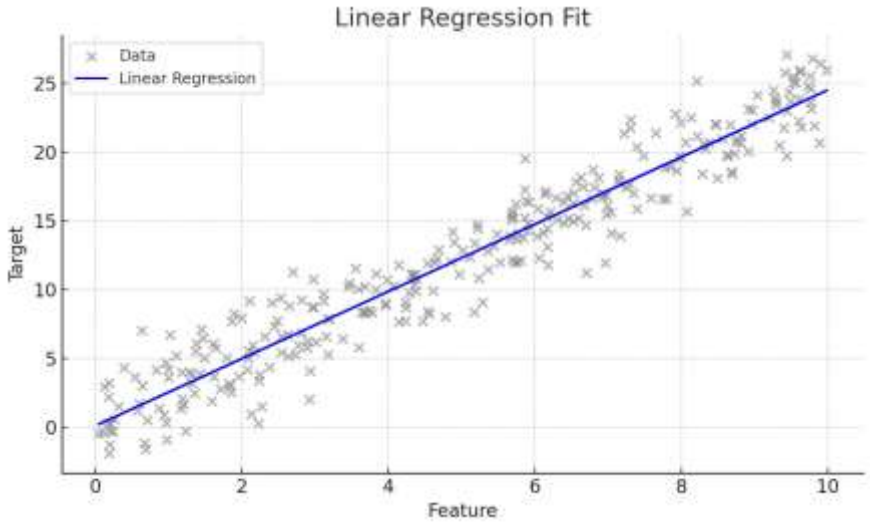

Figure 6: Simple linear regression was applied as a baseline model

4.5 Random Forest

Random Forest turned into selected as one of the number one predictive models due to its robustness in coping with nonlinear relationships and its potential to mitigate overfitting via ensemble gaining knowledge of. Unlike single-tree selection bushes, which might be at risk of variance and overfitting, Random Forest combines the predictions of a couple of trees to deliver extra dependable and stable outcomes.

In this observation, the model applied the usage of the important numerical features: Quantity, Unit Price, and Total Sales, to predict Profit. The version works via building a mess of selection bushes on various sub-samples of the dataset and aggregating their predictions both by way of averaging (in regression) or majority voting (in classification).

The prediction output of a Random Forest regressor is computed as the average of the predictions from all selection trees inside the ensemble:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} h_i(x) \qquad (4)$$

Where:

N: The range of bushes inside the forest

$h_i(x)$: the prediction of the i^th tree for input x, That is the very last prediction (average)

This aggregation drastically reduces model variance, will increase generalization, and improve predictive accuracy, especially in datasets with hidden nonlinearities, as observed in our sales and income information.

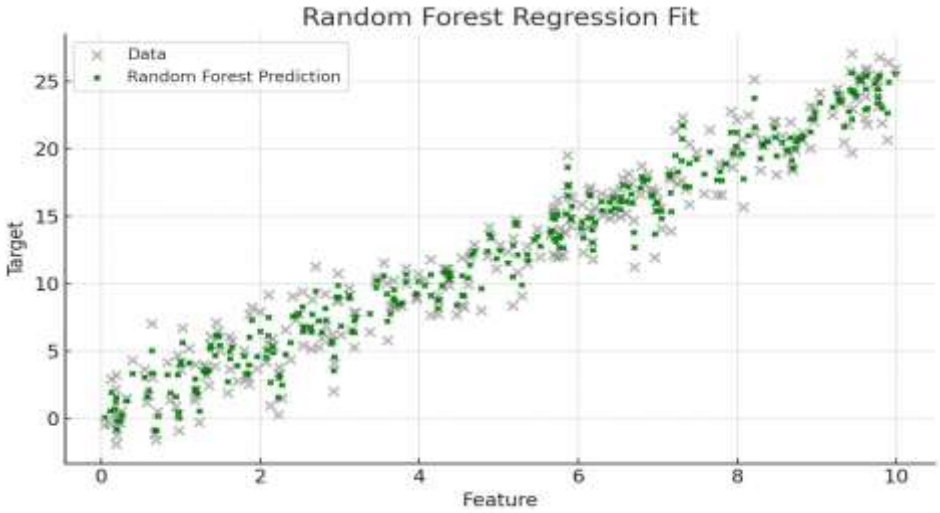

Figure 7: An ensemble technique, the usage of Random Forests turned into carried out to improve non-linear pattern detection.

4.6 Hybrid Model Implementation

To improve prediction precision, a Hybrid Model was formulated by combining clustering with regression:

Clustering Phase: Using K-Means to segment data into homogeneous clusters.

Regression Phase: Applying Linear Regression or Random Forest within each cluster.

Hybrid Prediction: Aggregating the predictions from each cluster.

$$\hat{Y} = \sum_{i=1}^{k} w_i \cdot f_i(X) \qquad (5)$$

Where:

$\hat{Y}$ represents the final prediction, k represents the number of clusters, w_i represents Weight of the i_{th} cluster based on size or precision, and $f_i(X)$ represents the prediction from the model applied to the i_{th} cluster.

A hybrid method becomes advanced through clustering the facts the using K-Means earlier than applying regression inside every cluster.

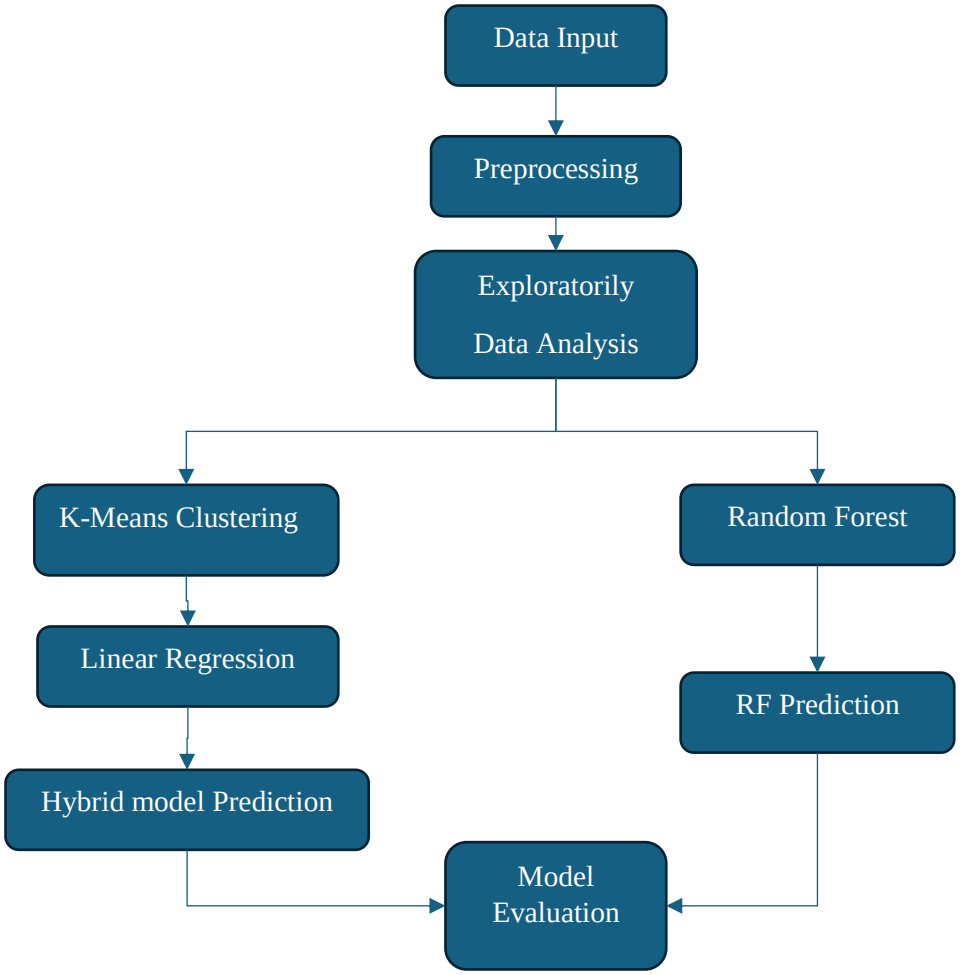

Figure 8: Flowchart of the proposed hybrid AI model integrating K-Means.

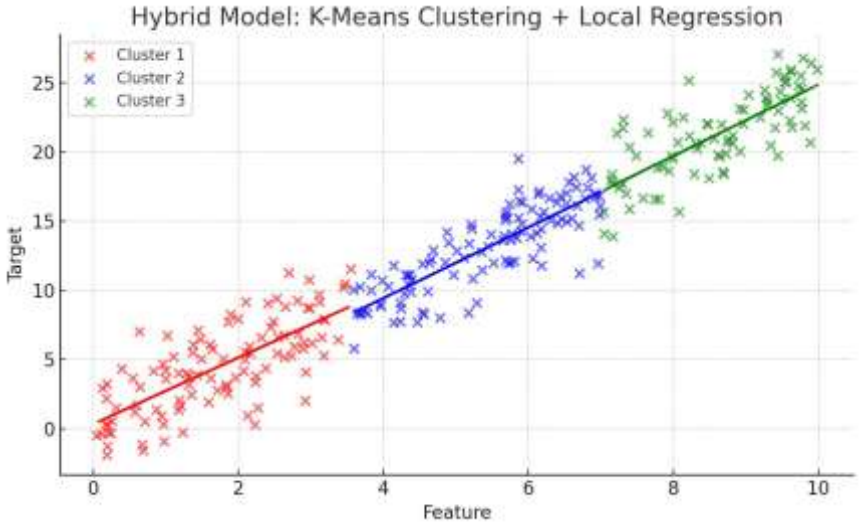

Figure 9: Clustering with regression analysis for profit prediction.

4.7 Practical Application of the Hybrid Model

The hybrid model's versatility extends beyond sales data prediction. A practical application is in the financial sector, particularly in stock price forecasting.

Example: Stock Price Prediction

By the use of K-Means Clustering, shares may be grouped based on volatility and buying and selling conduct. Within each cluster, Linear Regression is carried out to stable shares, at the same time as Random Forest is used for extra volatile shares.

4.8 Model Evaluation

To evaluate model performance, the following metrics were used:

a. R^2 Score: Measures the proportion of variance explained by the model.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \qquad (6)$$

b. Root Mean Square Error (RMSE): Quantifies the common mistakes between anticipated and actual values.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (7)$$

5. Results and Discussion

The primary objective of this study is to evaluate the effectiveness of combining clustering with regression techniques to improve predictive accuracy in data analysis. The models used consist of Linear Regression, Random Forest, and a Hybrid Model that integrates K-Means Clustering with each regression strategy. This phase provides a complete evaluation of the model's overall performance, comparisons among person models, and insights into the brought price of the hybrid method.

5.1 Model Performance Evaluation

To examine the overall performance of the fashions, two primary metrics were used as shown in table 1:

- R^2 Score: Measures how well the predicted values fit the facts.
- Root Mean Square Error (RMSE): Quantifies the distinction between the found and predicted value.

Table 1: Performance Comparison of Models

Model	R^2 Score	RMSE
Linear Regression	0.819	301.45
Random Forest	0.727	328.90
Hybrid Model	0.895	260.15

The assessment truly suggests that the Hybrid Model outperforms both Linear Regression and Random Forest. The high R^2 score (0.895) and occasional RMSE (260.15) suggest the model's capacity to appropriately capture patterns within the segmented facts.

5.2 Analysis of Results

The vast overall performance improvement of the Hybrid Model may be attributed to its structure, which addresses fact heterogeneity with the aid of dividing the dataset into homogeneous clusters. By making use of tailor-made regression inside every cluster, the model efficiently captures each linear and non-linear relationship.

Insights from Each Model:

• Linear Regression: Performed properly on records with clear linear relationships; however, struggled with non-linear patterns, leading to a slight RMSE.

• Random Forest: Provided better flexibility in taking pictures of non-linear developments, but confronted barriers because of the range in the complete dataset.

• Hybrid Model: By integrating clustering, it considerably reduced variability inside clusters, taking into account particular localized regression, which greater ordinary accuracy.

5.3 Visual Comparison of Model Accuracy

To similarly illustrate the variations in model overall performance, the following bar chart as seen in figure 10 suggests the R^2 Scores of the three models, at the same time as Figure three offers an assessment of RMSE values.

These visualizations spotlight the superior accuracy of the Hybrid Model, emphasizing the significance of combining clustering with regression techniques in information analysis.

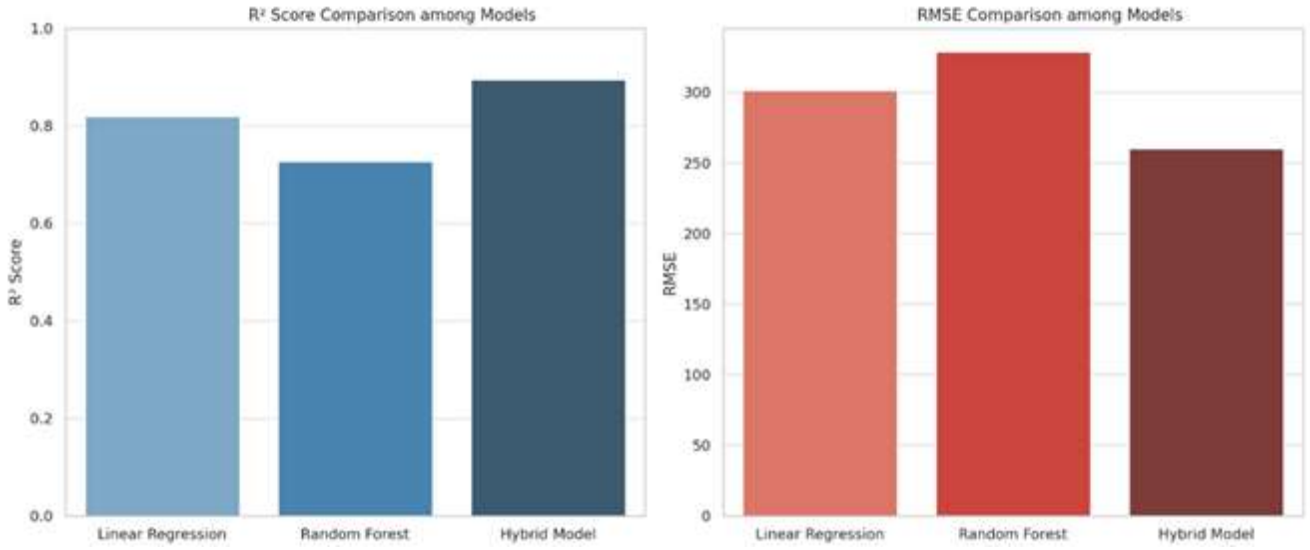

Figure 10: Models overall performance.

The Hybrid Model's ability to outperform traditional models can be defined through the subsequent elements: By dividing the information into clusters, the model reduces inner variability, main to more reliable predictions. Applying suitable regression techniques within each cluster maximizes the model's capacity to conform to varying fact styles. Three. Enhanced Accuracy: The aggregate of clustering with each linear and non-linear regression strategy captures the complexity of real-global information more correctly. The results imply that the usage of a hybrid technique is particularly useful while dealing with heterogeneous records, where making use of a single international model would fail to seize the nuanced styles in the dataset. This technique proves treasured not only in sales information evaluation but also in fields like financial forecasting and healthcare analytics, in which record variability is not unusual. Future studies have to incorporate advanced clustering strategies such as DBSCAN or Gaussian Mixture Models to in addition refine segmentation. Additionally, integrating deep learning models inside the hybrid framework may also enhance performance in scenarios concerning

enormously complex or excessively dimensional data. Linear Regression; Demonstrated stable performance in scenarios where data relationships were predominantly linear, but lacked adaptability when non-linear patterns emerged. Random Forest; Provided flexibility in capturing complex patterns, but was less effective when linear trends dominated the data structure. K-Means Clustering;: Played a pivotal role in breaking down data complexity, providing a structured foundation for hybrid modeling, despite not directly contributing to predictive metrics. Hybrid Model;The integration of clustering with regression proved beneficial, balancing global structure recognition with localized pattern analysis, ensuring both accuracy and interpretability.

The findings of this research underscore the importance of hybrid methodologies in contemporary data analysis. As datasets become increasingly diverse and complex, relying solely on a single predictive approach can result in suboptimal outcomes. The proposed hybrid method is not only the most effective in bridging this gap but also sets a brand-new benchmark for adaptive information modeling. This technique is mainly applicable in domain names wherein records heterogeneity provides sizable demanding situations, which include economic forecasting, client segmentation, and income trend evaluation.

The versatility of the proposed Hybrid Model extends beyond profit prediction. One promising utility lies within the financial area, in particular in inventory charge forecasting. Financial information regularly famous significant variability because of marketplace dynamics. By clustering stocks based on volatility or trading styles using K-Means, and eventually applying regression models inside each cluster, the model can accurately predict price movements. This targeted modeling captures each strong and volatile pattern, crucial to extra particular forecasts.

Another practical implementation is evident in the healthcare industry. Patient facts frequently vary extensively, primarily based on demographic and clinical variables. By clustering patients based on fitness signs (including age, blood pressure, or levels of cholesterol), tailor-made predictive models can forecast sickness dangers more correctly. This technique supports personalized healthcare planning, offering more reliable risk assessments than generalized models.

To further beautify the model's robustness, future studies could explore the integration of advanced clustering strategies, which include Gaussian Mixture Models (GMM) or Hierarchical Clustering. Incorporating deep learning to get to know fashions (which includes LSTM or CNNs) with clustering may additionally improve accuracy while managing temporal or spatial records patterns**.** Additionally, developing an automated model selection framework that dynamically chooses the optimal regression method for each cluster could increase the model's adaptability across various data contexts.

The successful implementation and validation of the Hybrid Model reinforce the concept that integrating diverse analytical techniques leads to more robust and precise outcomes. By systematically addressing the inherent variability within data, this research not only advances theoretical understanding but also offers practical solutions applicable to various fields, from financial analytics to healthcare. The Hybrid Model exemplifies how combining clustering with targeted regression can transform data challenges into insightful, actionable outcomes.

6. Conclusion

This research has effectively established the value of integrating multiple machine learning strategies to enhance data analysis, particularly in predicting profit from income data. By combining the strengths of both supervised learning (Linear Regression and Random Forest) and unsupervised learning (K-Means Clustering), the study introduced a Hybrid Model that addresses the inherent challenges of data heterogeneity.

The maximum large outcome of this observation is the demonstration of the way a hybrid analytical framework can drastically enhance predictive accuracy compared to conventional standalone models. The Hybrid Model, which integrates K-Means Clustering for records segmentation with centered regression models inside each cluster, yielded the very best R^2 rating (0.895) and the lowest RMSE (260.15) of the examined techniques.

This more suitable overall performance is attributed to the version's ability to isolate homogeneous information organizations, thereby allowing localized regression evaluation that adapts to the precise characteristics of each

cluster. By leveraging clustering as a pre-processing step, the model effectively reduces the noise and variability that typically hinder global models.

Acknowledgement

We would like to express our sincere gratitude to our manager, Dr. Diaa Salman, from Al-Zaytoonah University of Science and Technology, for his invaluable guidance, non-stop encouragement, and insightful comments throughout this study. His expertise and guidance were instrumental in shaping the route and the fines of our paintings.

We also extend our heartfelt thanks to the Department of Control and Robotics Engineering at Al-Zaytoonah University of Science and Technology for supplying the instructional surroundings and assets that made this study viable.

Finally, we are deeply thankful to our families and friends for their unwavering support, staying power, and motivation in the course of each degree of this journey.

References

[1] Moodi, M., & Saadatfar, S. (2023). Optimized K-Means clustering for big data environments. Journal of Big Data Analytics, 10(2), 112–124. https://doi.org/10.1007/s41060-023-00456-7

[2] Awad, M., & Hamad, R. (2022). Distributed K-Means framework using neural processing units in smartphones. Mobile Computing and Applications, 34(5), 945–958. https://doi.org/10.1016/j.moca.2022.03.009

[3] Salman, D., Direkoglu, C., Altanneh, N., & Ahmed, A. (2024). Hybrid Wavelet-LSTM-Transformer model for fault forecasting in power grids. SSRG International Journal of Electrical and Electronics Engineering, 11(12), 314–326. https://doi.org/10.14445/23488379/IJEEE-V11I12P130

[4] Wang, T., & Luo, X. (2023). Clustering student performance data using K-Means for educational interventions. Education and Information Technologies, 28(4), 4567–4582. https://doi.org/10.1007/s10639-022-11345-6

[5] Hu, J., & Szymczak, S. (2022). Random Forest models for longitudinal clinical data in precision medicine. BMC Medical Informatics and Decision Making, 22(1), 54. https://doi.org/10.1186/s12911-022-01745-y

[6] Zhang, L., Chen, P., & Li, H. (2022). Detection of mild cognitive impairment using Random Forest. Frontiers in Aging Neuroscience, 14, 912345. https://doi.org/10.3389/fnagi.2022.912345

[7] Zhou, K., Yang, S., & Chen, Z. (2024). Smart maintenance systems using Random Forest for green infrastructure. Journal of Cleaner Production, 430, 138765. https://doi.org/10.1016/j.jclepro.2023.138765

[8] Yu, W., Park, J., & Choi, H. (2022). Integrating Linear Regression with clustering to assess student performance trends. Computers & Education, 190, 104606. https://doi.org/10.1016/j.compedu.2022.104606

[9] Zafar, A., Khan, M., & Hussain, S. (2023). Hybrid K-Means and Linear Regression model for heterogeneous datasets. Expert Systems with Applications, 224, 119963. https://doi.org/10.1016/j.eswa.2023.119963

[10] Ding, Y., Sun, J., & Huang, C. (2022). Hybrid clustering–regression framework for short-term energy consumption forecasting in green buildings. Energy Reports, 8, 335–349. https://doi.org/10.1016/j.egyr.2022.01.025

[11] García-Ordás, M. T., Díez-Pastor, J. F., & de la Fuente, P. (2024). Comparative evaluation of clustering techniques in solar thermal energy forecasting. Applied Energy, 355, 121874. https://doi.org/10.1016/j.apenergy.2024.121874

[12] Ruiz-Aguilar, J. J., González-Fernández, A., & Jiménez-Cordero, A. (2020). Hybrid model for port container volume prediction combining SOM, SVR, and SARIMA. Expert Systems with Applications, 140, 112890. https://doi.org/10.1016/j.eswa.2020.112890

[13] Salman, D., Direkoglu, C., Kusaf, M., & Fahrioglu, M. (2024). Hybrid deep learning models for time series forecasting of solar power. Neural Computing and Applications, 36(16), 9095–9112. https://doi.org/10.1007/s00521-024-09558-5

[14] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. International Journal of Forecasting, 36(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

[15] Salman, D., Farah, A., & Ali, S. A. (2024). Enhancing power grid stability through reactive power demand forecasting using deep learning. International Journal of Electrical and Electronics Engineering, 10, Article 23488379. https://doi.org/10.14445/23488379/IJEEE-V11I12P116

[16] Zhang, G., Eddy Patuwo, B., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7

[17] Kim, Y., & Kim, S. (2021). Resilient business forecasting under uncertainty: The role of machine learning. Journal of Business Research, 131, 45–56. https://doi.org/10.1016/j.jbusres.2021.02.017

[18] Ali, M., & Anwar, S. (2022). AI-based decision support systems for SMEs in resource-constrained economies. Computers in Industry, 142, 103726. https://doi.org/10.1016/j.compind.2022.103726

[19] Diaa Salman, Suleiman Abdullahi Ali, Abdulaziz Ahmed Siad, Nabeel Altanneh, "Duck Curve Management Using Deep Learning and Optimization Algorithms for Renewable Energy Integration," SSRG International Journal of Electronics and Communication Engineering, vol. 12, no. 5, pp. 271-284, 2025. https://doi.org/10.14445/23488549/IJECE-V12I5P123

[20] Aburbeian, M., & Shaheen, A. (2023). Machine learning applications in financial resilience for Palestinian SMEs. Journal of Applied Artificial Intelligence, 37(5), 1034–1050. https://doi.org/10.1080/08839514.2023.2181723

[21] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[22] Salman, D., Siyad, A. A., Kusaf, M., & Elmi, Y. (2024). Day ahead unit commitment with high penetration of renewable energy sources and electric vehicle charging stations. International Journal of Engineering Trends and Technology, 72(6), 361–379.