CTICTR

CURRENT TRENDS IN INFORMATION COMMUNICATION TECHNOLOGY RESEARCH

BABCOCK UNIVERSITY

KNOWLEDGE TRUTH SERVICE

ILISHAN - REMO, NIGERIA.

CTICTR 4(1): 116 - 126 (June 2025)

ISSN: 2814-1709

Received: 08-03-2025

Accepted: 09-06-2025

https://doi.org/10.61867/pcub.v3i1a.203

MACHINE LEARNING MODELS FOR SALES PREDICTION

A. A. Udosen[a], M. E. Adesuyan[b*], O. O. Bamidele [c], D. Nteziryayo [d], and J. Mugerwa[e]

Corresponding Author Email: *adesuyan0173@pg.babcock.edu.ng*

Machine Learning Models for Sales Prediction

A. A. Udosen[a], M. E. Adesuyan[b*], O. O. Bamidele [c], D. Nteziryayo [d], and J. Mugerwa[e]

[a] *Department of Computer Science, Babcock University, Ilishan Remo, Ogun State, Nigeria.*
[b] *Department of Computer Science, Babcock University, Ilishan Remo, Ogun State, Nigeria.*
[c] *Department of Computer Science, Babcock University, Ilishan Remo, Ogun State, Nigeria.*
[d] *Department of Computer Science, Adventist University of Central Africa (AUCA), Rwanda.*
[d] *Department of Computing and Informatics, Bugema University, Kampala, Uganda*

[a] udosena@babcock.edu.ng
[b] adesuyan0173@pg.babcock.edu.ng
[c] bamideleo@babcock.edu.ng
[d] deogratias.nteziryayo@auca.ac.rw
[c] jmugerwa@bugemauniv.ac.ug

Abstract
Sales forecasting is an essential task for effective business planning and implementation, and also inventory management. Traditional forecasting methods usually fail to identify complex sales patterns, resulting in inaccurate predictions. Therefore, this study examines the accuracy of sales prediction using machine learning models. By leveraging historical sales data, this study estimates multiple machine learning algorithms, such as random forest, neural networks, linear regression, XGBoost, and decision trees. The findings reveal that, while Linear Regression recorded high error rates, MAE (136.07%), RMSE (303.85%), MAPE (302.66, %), Random Forest achieved the lowest errors, MAE (110.10%), RMSE (234.74%), and MAPE (38.61%), showing its outstanding forecasting performance. With a better predictive accuracy, Random Forest gives businesses better insights for decision-making concerning inventory management, efficiency of operations, and meeting market demands. The study emphasizes the increase of data-driven strategies and business analytics in sustaining competitive advantage.

Keywords: Decision trees, Machine Learning, Predictive Analytics, Regression models, Sales Forecasting

1 Introduction

Sales forecasting is crucial for proper inventory management, financial planning and the allocation of resource. Traditional approaches, such as Autoregressive Integrated Moving Average (ARIMA) and exponential smoothing, were employed before now, but they were not efficient enough in managing the intricacies of the present sales data. These conventional methodologies did not adequately identify the complex, nonlinear interactions that drive sales, such as market trends, promotional activity, and seasonal variations [1]. In business management, strategic decision-making and other business

processes are directly impacted by sales forecasting. Basically, product availability, consumer demand, streamlining production schedules, optimizing stock levels, reduced costs, increased customer satisfaction, lowering superfluous inventory, and maximizing resource allocation were regulated by correct sales projections [2].

These conventional sales forecasting methods cannot be used in today's dynamic marketplaces due to the complication of modern transactional record that is dependent on market trends, seasonality, and promotions. Basically, market intricacy, inflexibility, and few features are some of the limitations they have [3]; and they cause overstocking, stock out, and poor resource allocation, causing harm to profit levels and competitiveness [4].

The application of machine learning (ML) algorithms to process of huge data made possible to detect hidden patterns in the dataset in real time [5]. Also, they efficiently uncover complex data relationships with high precision. Due to their nature to adaptively learn with changing economic instability; ML can accurately outperform traditional sales forecasting methods as they can process large datasets and detect intricate, nonlinear patterns. Basically, these algorithms can greatly improve forecasting ability [6], and they are particularly good at finding complex links in the data, and they produce projections that are more accurate and useful.

Techniques such as random forests incorporate features like product characteristics, marketing efforts, and external economic factors, greatly improving forecasting capabilities [6]. As an ensemble classifier algorithm, it has the capacity to solve the problems of overfitting and local minima [4]. This study addresses these limitations by analysing machine learning models explicitly planned for sales forecasting, such as Random Forest, XGBoost, and neural networks. Despite advancements in sales forecasting using machine learning models, a notable limitation is inevitable in complex datasets having high seasonality and varieties. Hence, the need to evaluate the performances of these models would help to advance the accuracy and dependability of sales forecasts, enhancing decision-making in financial planning, resource allocation, and inventory management. This study is expected to help businesses navigate the competitive market landscape.

2 Literature Review

2.1 Sales Forecasting

Sales forecasting, according to [7], is important for managing workforce, cash flow, and resource allocation in companies. Also, [8] added that it enables businesses to effectively plan their inventory, production, and marketing strategies. Accurate sales forecasting will enable strategic planning and fostering market growth and revenue generation [7]; [8] supported that it will allow businesses to optimize their operations, reduce waste,

and increase profits. The authors explored the optimal approach for highly precise sales forecasting, with a focus on its importance in operational, marketing, sales, production, and finance for businesses seeking investment capital.

Retail sector has experienced a notable increase in sales since the emergence of e-commerce sites like Jumia, according to [9]. However, [10] argued that traditional sales forecasting methods like linear regression struggle with the intricacy of sales figures, including varieties and seasonality, despite their successes. Hence the need to explore optimized techniques to forecast sales to improve productivity.

2.2 Machine Learning Algorithms

[11] defined machine learning as the different algorithms by which computer systems provide accurate answers to complex questions by observing patterns and insights in datasets, and this can be done without the use of explicit programming [12]. Machine learning is utilised in prediction systems to discovery the forms and other similarities in nearby data points that are connected by substantial weight edges that are similar [13]. According to [1], some machine learning algorithms include Linear Regression, Neural Networks, XGBoost, Support Vector Machines (SVM),) Decision Trees, Random Forest, AutoRegressive Integrated Moving Average (ARIMA), Gradient Boosting Machines (GBM), and Long Short-Term Memory Networks (LSTM) to mention a few.

2.2.1 Types of Machine learning Algorithms

According to [14], the differences between machine learning and traditional programming are shown in Table 1 as follows:

Table 1: Difference between Machine Learning and Traditional Programming

Machine Learning	Traditional Programming	Artificial Intelligence
It is a subset of artificial intelligence (AI) that focuses on learning from data in order to create an algorithm capable of making predictions.	Developers write rule-based code in traditional programming based on the problem statements.	Artificial intelligence is the process of making a machine as capable as possible so that it can accomplish jobs that would normally require human intelligence.
It takes a data-driven method, often training on historical data before making predictions on fresh data.	Conventional programming is usually dependent on rule and deterministic. It lacks self-learning capabilities such as AI and ML.	AI can use a variety of techniques that include Machine Learning, classic rule-based programming, and Deep Learning
ML can automatically detect forms and insights in vast datasets that may be missed by humans.	It depends on the inventors' intelligence. As a result, its capabilities are extremely limited.	AI sometimes employs a amalgamates predetermined rules and data that provides a significant advantage in accomplishing complex tasks with high precision that people believe are impossible.

ML is a subset of AI. It is now employed in a variety of AI-based jobs, including chatbot question answering, self-driving cars, and so on.	Basic programming is commonly used to create software with specialized functions.	AI is a vast field with many uses, including natural language processing, computer vision, and robots.

There are many benefits to incorporating machine learning models into sales forecasting. Through their ability to recognize intricate, nonlinear patterns and relationships in data, machine learning models increase prediction accuracy. Conversely, to implement ML models, advanced computational resources and knowledge are necessary understand these models due to their intricacies. Also, the used of huge datasets may cause privacy and security issues [6].

2.3 Sales Forecasting using Machine Learning Model

Many studies have proven that machine learning approaches are efficient in sales forecasting in a range of businesses. By leveraging machine learning (ML) according to [15], there has been the implementation of composite models for sales forecasting that have tremendously increased prediction accuracy and reliability. These machine learning models can scan massive, complex datasets, find complex patterns, and produce incredibly accurate predictions. These models employ different features, such as external economic data, marketing activities, and product qualities to produce more accurate predictions.

Categorically, [16] showed how retail sales forecasting was done using LightGBM that showed a better outcome than conventional techniques in terms of accuracy and computing economy. Also, [2] employed Recurrent Neural Networks (RNNs) in sales prediction that proved their capacity to handle sequential data and identify long-term relationships. Consequently, a study by [17], showed that ML systems could perform better than conventional methods in processing multidimensional sales data. In addition, Long Short-Term Memory (LSTM) in particular and RNNs are models that have demonstrated greater performance in identifying temporal relationships in sequential data, making them compatible for time series forecasting task [18]. Importantly, ML approaches have been used to transform sales forecasting; offering sophisticated tools capable of analyzing complicated datasets, spotting minute trends, and producing incredibly precise predictions.

Many attributes, such as external economic data, marketing activities, and product qualities, have brought about more accurate and useful forecasts using ML [6]. For example, gradient boosting techniques such as LightGBM have been used to capture nonlinear interactions among features and handle huge datasets efficiently, improving forecast accuracy [16]. Furthermore, these hybrid models have shown increasing forecast precision by fusing classic time series techniques to capture temporal dependencies needed to manage complicated, high-dimensional data [2]. Their ability to endlessly update their models based on fresh data, ML have what it takes to adjust to shifting market conditions [6]. For example, ensemble approaches, such as gradient boosting, bagging, and stacking, have the potential to yield more accurate predictions [17]. Furthermore, the authors argued that continued development of advance models for tasks including sales forecasting will birth more advanced prediction models needed to make data-driven decisions.

By taking it a step further, deep learning models, that outperform conventional techniques in forecasting sales trends over extended time horizons, are better able to adjust to shifting customer behaviour patterns and market conditions. For instance, integrating recurrent neural networks and exponential smoothing has shown better outcome in raising forecasting accuracy [18]. Additionally, [15] conducted a study that showed that ensemble learning approaches, like gradient boosting and stacking, that integrate the powers of different algorithms that are used in a variety of forecasting settings, indicating their potential to increase sales prediction accuracy that are more reliable. Also, a study by [19] showed an effective use of ML models to enhance the predictive capabilities that are necessary to recognize distinct trends within the data through clustering algorithm. However, the study indicated that the exactness of the predictions deeply depended on the worth and comprehensiveness of the input data as errors or omitted values can pointedly affect outcomes.

A study by [20] suggested that ensemble learning approaches, such as decision trees and random forest algorithms outperform other individual algorithms, with marginal superiority in their performances. A study by [9] found that improved models, particularly the gradient boosting machine, outperformed the

baseline linear regression model in predicting sales on Jumia, achieving the low Mean Absolute Error and Mean Squared Error, demonstrating superior prediction accuracy. However, [19] argued that the intricacy of the ML algorithm may require substantial resources and expertise, posing a challenge for smaller businesses with limited resources.

Based on the foregoing, it is necessary to say that the black-box nature of certain sophisticated algorithms may complicate the interpretation and communication of results to non-technical stakeholders.

3 Methodology

This study discusses the existing machine learning models for sales forecasting. The conceptual model was designed using various algorithms, including traditional methods like Linear Regression and more complex techniques like Decision Trees, Random Forests (a homogeneous ensemble learning algorithm), Gradient Boosting Machines, and Neural Networks. In essence, the accuracy and timeliness of these five classification algorithms' processing of huge datasets, as well as their capacity to handle widely distributed data points within a dataset, were used to compare their performances. The process includes three stages: the preparation stage, the model selection and training stage, and the validation and evaluation stage. Figure 1 presents the conceptual diagram of the proposed model. It outlines the sequential steps involved, starting from data assemblage, cleaning and normalization, model training, validation, and deployment for sales forecasting.

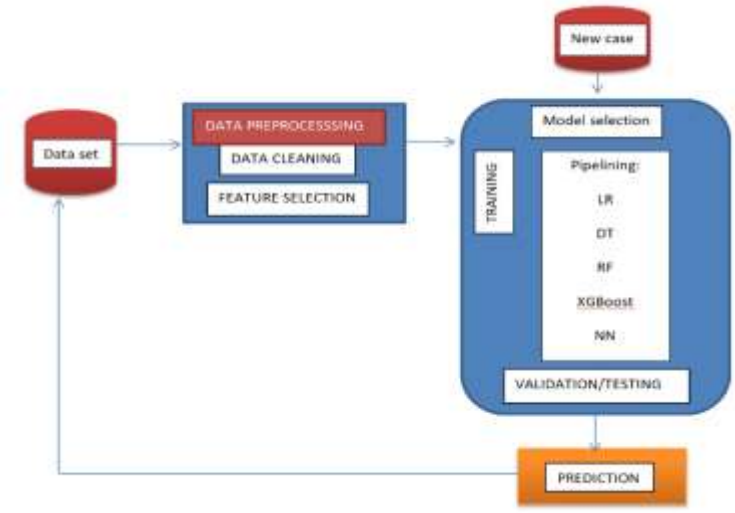

Figure 1: Conceptual Diagram of the Model

3.1 Design of the Machine Learning Model for Sales Forecasting

The models used for sales forecasting in this study consolidated a dataset with different dimensions to make accurate forecasts for informed decisions on sales forecasting. This design prioritizes scalability, reliability, and actionable insights for enhanced sales performance. The evaluation of these algorithms is necessary to ensure reliability, while a feedback loop enables iterative improvement. Figure 2 highlights the logic flow of the model from the input stage through the processing stage to the output stage.

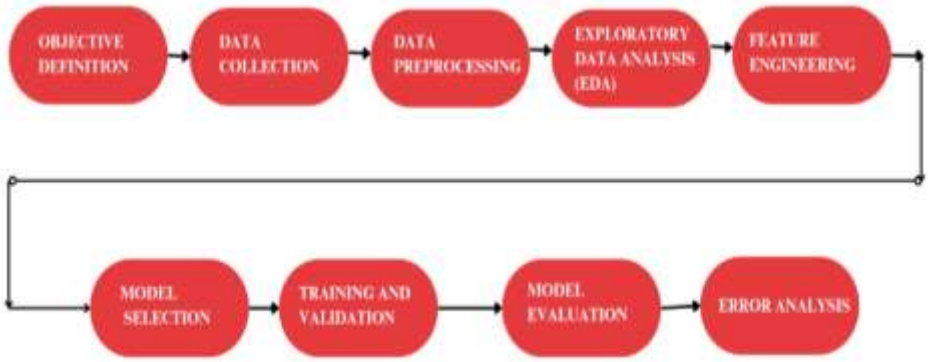

Figure 2: Logic Flow Diagram

3.2 Data Collection and Preprocessing

This study employed the Store Sales Forecasting obtained from Kaggle [21], and contains the records of past sales transactions of businesses operating in the retail furniture sector. This dataset includes information such as order date, sales amount, quantity sold, discounts applied, customer details, and product information.

3.3 Cleaning and Preprocessing

For data cleaning and pre-processing, the missing values in the attributes of the dataset, such as Sales, Quantity, Discount, and Profit were imputed using the mean value of the data points. This was done to ensure that the missing values were replaced with representative values, preserving the overall distribution and statistical properties of the data. Also, for the missing features in categorical features that cannot be imputed using mean or median, such as Customer Information and Product Details, often contain missing values that cannot be imputed using mean or median values, replacing the omitted values with the most recurrent category within each feature. By so doing, we maintained the categorical structure of the data while filling in missing entries with plausible values. Thereafter, duplicate records were systematically removed from the dataset, retaining only unique instances of each transaction or event.

3.4 Feature Engineering

In machine learning, the necessary features needed for training the model is selected automatically. Basically, categorical encoding, one-hot encoding, to be precise, was used for the nominal categorization of the dataset order was not important.

3.5 Model Development and Evaluation

Each of the models was trained on the pre-processed dataset with an 80-20 train-test split to allow for a better training process of the model using sufficient data points. Randomized search was used for the hyper parameter tuning to enhance model performance. Figure 3 shows the training process of the developed model based on the 80-20 train-test split.

```
import pandas as pd
from sklearn.model_selection import train_test_split

# Assuming 'df' is your DataFrame and 'Sales' is the target variable
X = df.drop('Sales', axis=1)
y = df['Sales']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print("X_train shape:", X_train.shape)
print("X_test shape:", X_test.shape)
print("y_train shape:", y_train.shape)
print("y_test shape:", y_test.shape)
```

```
X_train shape: (1695, 447)
X_test shape: (424, 447)
y_train shape: (1695,)
y_test shape: (424,)
```

Figure 3: Training of the Model

3.6 Model Evaluation

The evaluation of the models through the following performance metrics:

1. **Mean Absolute Error (MAE):** Measures the average absolute difference between actual and predicted sales values.
2. **Mean Squared Error (MSE):** Measures the average squared difference between actual and predicted sales values.
3. **Root Mean Squared Error (RMSE):** The square root of the MSE, providing a measure of the model's prediction error.

4 Results and Discussions of Findings

The discoveries from the study provided understanding of the strengths and weaknesses of the different models used, highlighting their common challenges, and the best practices for developing accurate sales forecasting models. The machine learning models demonstrated superior performance when compared with each other. The outcome of the models used were evaluated through Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and R-squared. The evaluation highlights the best-performing model based on these metrics. With the use of MAPE and RMSE for instance, Random Forest, had the smallest error rate (38.61 and 234.74 respectively) with Neural Network being the highest for RMSE and Linear Regression being the highest for MAPE respectively. Also, with MAE, XGBoost had the smallest error rate (110.02), indicating their superior ability to predict sales accurately while Decision tree had the highest rate (153.86). The outcome provided tremendous progresses in forecasting sales. Table 1 below contains the evaluation of each metric used and Table 2 contains the summary of the results of each of the models based on the observed evaluation scores.

Table 2: Evaluation Metrics of the Models

Model	MAE	RMSE	MAPE
Linear Regression	136.07	303.85	302.66
Decision Tree	153.86	343.75	48.50
Random Forest	110.10	234.74	38.61
XGBoost	110.02	260.63	53.72
Neural Network	321.62	577.06	89.59

Table 3: Performance Summary

Model	Performance Analysis
Linear Regression	Linear Regression exhibited the highest errors, indicating it is not suitable for this sales forecasting task.
Decision Tree	Improved accuracy over Linear Regression, but significant room for improvement remains.
Random Forest	Best overall performance with the lowest errors, effectively capturing the complexities in sales data.
XGBoost	Competitive performance, slightly higher RMSE and MAPE than Random Forest, but still effective.
Neural Networks	Higher errors compared to Decision Tree, Random Forest, and XGBoost, indicating a need for further tuning.

The performance analysis reveals that the Random Forest model outperforms other models in this study, achieving the highest accuracy and lowest error rates. XGBoost follows closely, confirming the effectiveness of ensemble methods in capturing complex sales data patterns and improving forecasting reliability. Random Forest generally outperforms other models across all metrics, showing the smallest errors in both absolute terms and percentage terms. This suggests it is the utmost robust and reliable model to predict sales. Also, choosing a model should depend on the trade-off between complexity and performance. For instance, Neural Networks might be more complex and resource-intensive without providing better results compared to simpler models like Random Forest. While MAE gave a straightforward view of average error size, RMSE penalized larger errors more heavily, and MAPE provided a percentage-based error metric. Figure 4 show the use of a chart to represent the comparison of the models.

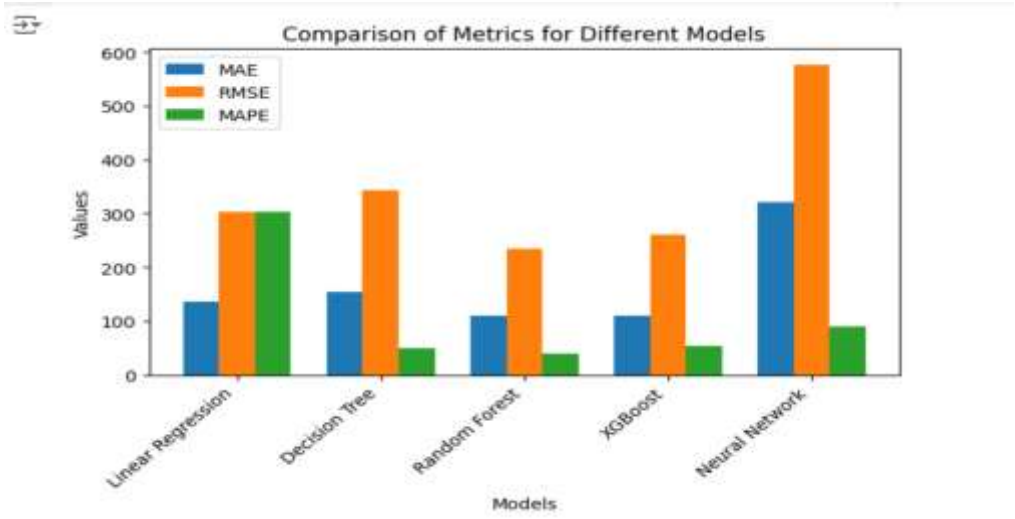

Figure 4: Graphical Representation of Model Comparison

5 Conclusion

This study shows the relevance of advanced ML techniques in predictive analytics as it outperforms traditional methods in sales forecasting. The study boasts of the performance of Random Forests and XGBoost in identifying the hidden complex patterns in sales data. This discovery can improve decision making in businesses with respect to marketing strategies and inventory management, improving operational efficiency and productivity. Importantly, a hybrid model that integrates Random Forest and XGBoost could improve competitive edge.

These results show that ML can identify sales patterns and offer valuable direction for decision making.

However, the study acknowledges limitations like the availability and worth of data, and also, advanced resources having the potential for overfitting in complex models. Further works could explore the use of more advanced techniques like ensemble learning and deep learning, and their application to other domains of forecasting.

References

[1] S. S. Wulff, "Time Series Analysis: Forecasting and Control, 5th edition," *J. Qual. Technol.*, vol. 49, no. 4, pp. 418–419, Oct. 2017, doi: 10.1080/00224065.2017.11918006.

[2] S. Elsworth and S. Güttel, "Time Series Forecasting Using LSTM Networks: A Symbolic Approach." arXiv, 2020. doi: 10.48550/ARXIV.2003.05672.

[3] J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, "Deep Learning for Time Series Forecasting: A Survey," *Big Data*, vol. 9, no. 1, pp. 3–21, Feb. 2021, doi: 10.1089/big.2020.0159.

[4] U. Brunelli, V. Piazza, L. Pignato, F. Sorbello, S. Vitabile, "Two days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy," Atmos. Environ. Vol 41, pp. 2967-2995.

[5] A. A. Udosen, O. F. Ajayi, and O. Adelowo, "Voting Ensemble Learning Model (VELM) for Harmful Gas Detection Systems: A Proposed Model," *Current Trends in Information Communication Technology Research (CTICTR)*, vol. 2, no. 1, pp. –, Jun. 2023.

[6] Samek, W., Wiegand, T., & Müller, K. R. (2019). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. Proceedings of the IEEE, 107(3), 441-465.

[7] H. S. M. Hitesh, A. Yukthi, and B. N. Ramya, "Sales Prediction Using Machine Learning Techniques," International Journal of Novel Research and Development, vol. 9, no. 1, pp. a443–a451, Jan. 2024.

[8] H. Jain, V. Dattpalsinh, S. K. Ray, and Dr. Vishal, "Sales Prediction using Machine Learning," *School of Computer Application, Lovely Professional University*, Punjab, India.

[9] E. K. Akinyemi, A. T. Audu, E. P. Akubo, O. A. Ogunsola, and D. O. Ighawho, "Machine Learning Techniques in Predicting Sales: A Case Study of Jumia," International Journal of Research and Innovation in Applied Science (IJRIAS), vol. 9, no. 12, pp. 53–60, 2024. doi: 10.51584/IJRIAS.2024.912053.

[10] P. Ganguly and I. Mukherjee, "Enhancing Retail Sales Forecasting with Optimized Machine Learning Models," arXiv preprint arXiv:2410.12345, Oct. 2024, revised Dec. 2024. [Online]. Available: https://arxiv.org/abs/2410.12345

[11] D. Stenvatten, "A Comparative Study for Classification Algorithms on Imbalanced Datasets. An Investigation into the Performance Of RF, GBDT And MLP," Examensarbete inom huvudområdet, 2020.

[12] Z, Mohammed, "Artificial Intelligence Definition, Ethics and Standards," Electronics and Communications: Law, Standards and Practice, British University in Egypt

[14] J. R. Kumar, R. K. Pandey, B. K. Sarkar, "Pollutant Gases Detection using the Machine learning on Benchmark Research Datasets," International Conference on Pervasive Computing Advances and Applications, Procedia Computer Science, issue. 152, pp. 360–366, 2019.

[14] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[15] Machine Learning Mastery With Python Understand Your Data, Create Accurate Models and Work Projects End-To-End.

[16] T. Zhou, "Improved Sales Forecasting using Trend and Seasonality Decomposition with LightGBM," in *2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, China: IEEE, May 2023, pp. 656–661. doi: 10.1109/ICAIBD57115.2023.10206380.

[17] K. Saraswathi, N. T. Renukadevi, S. Nandhinidevi, S. Gayathridevi, and P. Naveen, "Sales prediction using machine learning approaches," Proceedings of the 4th National Conference on Current And Emerging Process Technologies E-Concept-2021, Erode, India, 2021, p. 140038. doi: 10.1063/5.0068655.

[18] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," *Int. J. Forecast.*, vol. 36, no. 1, pp. 75–85, Jan. 2020, doi: 10.1016/j.ijforecast.2019.03.017.

[19] Lim, T. S., Loh, W. Y., & Shih, Y. S. (1999). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning.

[20] O. F. Ajayi, A. A. Udosen, W. Ajayi, B. O. Ohwo, and A. I. Amusa, "Voting Ensemble Learning Model (VELM) for Harmful Gas Detection in Environmental Applications," Asian Journal of Electrical Sciences, vol. 13, no. 2, pp. 45–50, 2024. doi: 10.70112/ajes-2024.13.2.4252.

[21] https://www.kaggle.com/datasets/tanayatipre/store-sales-forecasting-dataset