iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84087

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Artificial Intelligence Based Business Data Analysis and Visualization

K. Vennela[1], Dr. B. Kranthi Kiran[2]
[1]*M.Tech Student, Department of Information Technology, University College of Engineering, Science and Technology, JNTU Hyderabad, Telangana*
[2]*Professor, Department of Computer Science and Engineering, JNTU Hyderabad, Telangana*

Abstract: ***Business organizations generate large volumes of data from sales, customers, transactions, marketing activities and online platforms. Traditional business analysis methods are often slow and unable to identify hidden patterns in high-dimensional datasets. This study presents an Artificial Intelligence based business data analysis and visualization framework using Principal Component Analysis (PCA), Autoencoder, K-Means clustering, t-SNE and XGBoost. The proposed system first preprocesses the ecommerce business dataset, performs feature extraction using encoding and scaling, reduces dimensionality using PCA, and learns compressed nonlinear representations using an Autoencoder. K-Means clustering is used for customer segmentation, t-SNE is applied for two-dimensional visualization of customer groups, and XGBoost is used for business risk prediction. Experimental analysis on the ecommerce_business_dataset containing 35,000 records shows that PCA retained 95.04% information, the K-Means model obtained a silhouette score of 0.2166, and XGBoost achieved 99.97% accuracy in the current experimental setup. The system supports customer segmentation, cluster profiling, feature importance analysis and actionable business insight generation.***

Keywords: Artificial Intelligence, Business Data Analytics, PCA, Autoencoder, K-Means Clustering, t-SNE, XGBoost, Feature Importance.

I. INTRODUCTION

In the present digital business environment, data has become one of the most important assets for decision-making. Ecommerce organizations collect data from customer purchases, product categories, sales channels, marketing campaigns, website visits, discounts, customer satisfaction scores, purchase frequency and profit values. When this data is analyzed properly, it can reveal customer behavior patterns, market trends, profit-related factors and risk indicators.

Traditional business intelligence methods mainly use manual reporting, simple statistical analysis and basic visualization. These approaches are useful for small datasets, but they become less effective when the dataset contains many records and a large number of dimensions. High-dimensional data makes analysis complex, increases computation time and reduces the clarity of visualization. Therefore, artificial intelligence and machine learning techniques are required to process and interpret business data efficiently.

The proposed project develops an AI-based business data analysis and visualization system. The framework combines dimensionality reduction, deep feature learning, unsupervised clustering, visualization and supervised classification. PCA reduces data complexity while retaining major information. Autoencoder learns compressed hidden representations from PCA features. K-Means identifies customer groups. t-SNE converts complex feature space into a visual two-dimensional representation. XGBoost performs business risk prediction and feature importance analysis. The main contribution of this work is the integration of multiple artificial intelligence techniques into a single business analytics pipeline. The system not only predicts risk but also provides customer segmentation, cluster profiling and visual insights, making the output useful for business decision support.

II. RELATED WORK

Business data analysis and visualization have gained importance because of the rapid growth of ecommerce and digital transaction platforms. Traditional systems depend on statistical models and dashboard-based reporting, but they often fail to discover hidden patterns from high-dimensional datasets. Recent studies have explored artificial intelligence assisted business analysis methods to improve data understanding and decision-making. Dimensionality reduction techniques such as PCA are commonly used to reduce the number of variables while preserving the maximum possible information. PCA is suitable for business datasets because it transforms correlated features into principal components, reduces redundancy and improves computational efficiency. Deep learning based Autoencoders have also been applied for compressed feature representation because they can learn nonlinear relationships in

data. Visualization techniques such as t-SNE are widely used for representing high-dimensional data in two-dimensional space. This helps analysts visually understand hidden patterns and segment separation. K-Means clustering is frequently used in customer analytics to group customers with similar behavior. For prediction tasks, XGBoost is considered effective due to its high accuracy, gradient boosting mechanism and ability to provide feature importance scores.

Based on these existing techniques, the proposed system integrates PCA, Autoencoder, K-Means, t-SNE and XGBoost to create a complete artificial intelligence based business data analysis and visualization framework.

III. METHODOLOGY

A. System Overview

The proposed framework performs business data analysis through a sequential AI pipeline. The ecommerce dataset is uploaded into Google Colab and preprocessed. Relevant features are extracted using one-hot encoding and feature scaling. The high-dimensional encoded data is reduced using PCA and compressed further through an Autoencoder. The learned features are used for K-Means customer segmentation and t-SNE visualization. Finally, XGBoost classification is applied for business risk prediction and feature importance analysis.

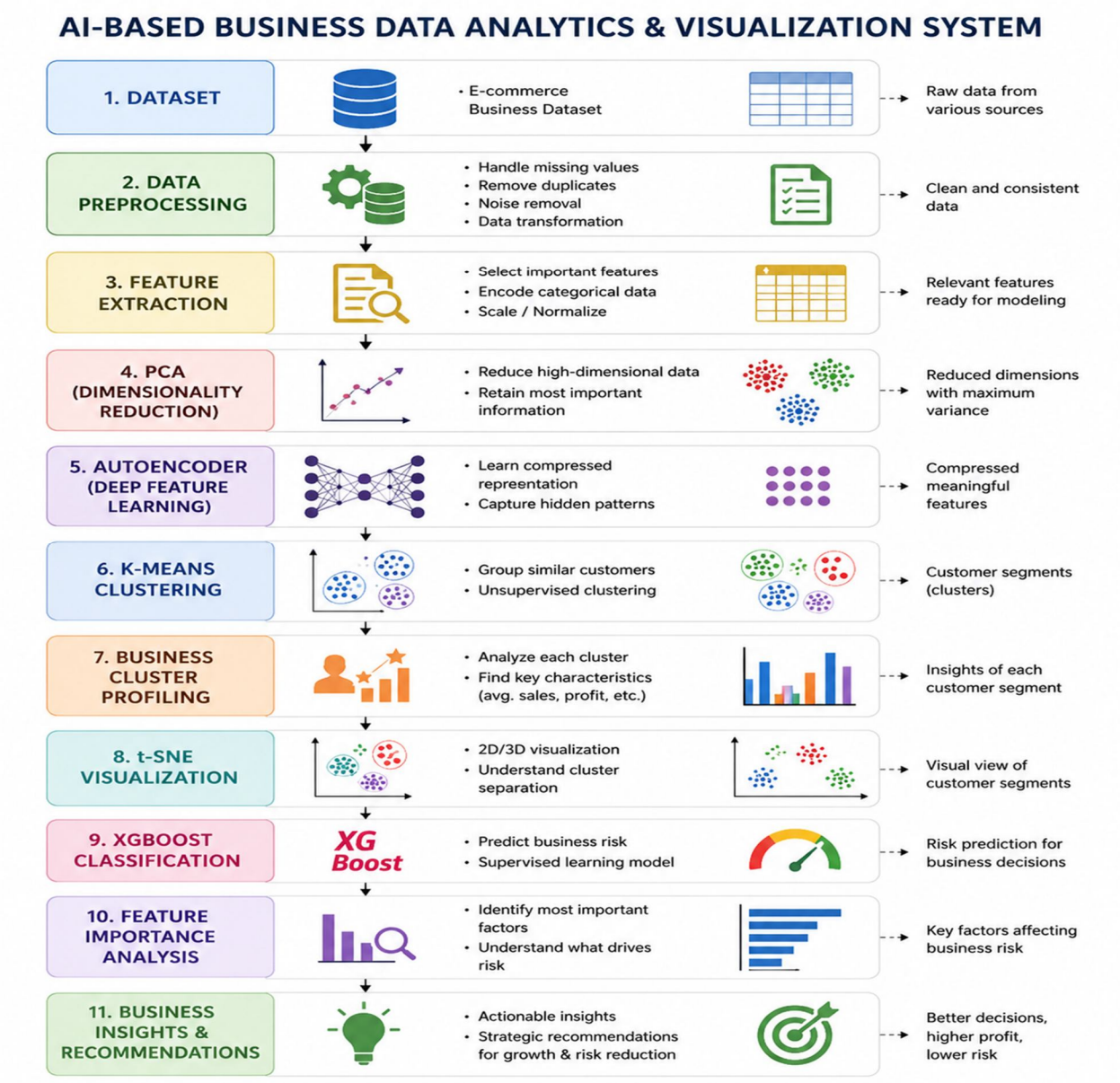


Fig.1 System Architecture of AI-Based Business Data Analytics and Visualization System

B. Data Acquisition

The ecommerce_business_dataset was used for experimentation. The dataset contains 35,000 records and initially includes 22 business-related columns such as order details, region, customer segment, product category, sales channel, unit price, quantity, discount, marketing spend, website visits, customer age, customer rating, sales, profit, return flag, satisfaction, recency, purchase frequency, monetary value and churn flag. During processing, additional columns such as Business_Risk and Cluster were generated, resulting in 24 final columns in the experimental output.

C. Preprocessing and Feature Extraction

Data cleaning was performed by removing duplicate records and replacing missing numerical values using median imputation. This step improves the quality and consistency of the dataset before model training. The target variable Business_Risk was created when it was not available in the dataset. The input features were separated from the target column. Unnecessary ID-based columns were removed to avoid misleading model training. Categorical features such as region, customer segment, product category and sales channel were converted into numerical features using one-hot encoding. StandardScaler was applied to scale all features to a common range. After feature extraction and encoding, the final feature shape was approximately 35,000 x 763.

D. Dimensionality Reduction and Deep Feature Learning

PCA was applied with a 95% variance retention setting. This means the system automatically selected enough principal components to preserve 95% of the information from the original encoded features. In the experiment, PCA retained 95.04% information and reduced the feature space for further processing. The PCA output was passed into an Autoencoder. The Autoencoder architecture contained encoder layers, a bottleneck layer and decoder layers. The bottleneck layer generated a compressed feature representation. This representation was used for clustering and visualization because it captures hidden nonlinear patterns in the business data.

E. Customer Segmentation, Visualization and Prediction

K-Means clustering was applied on the Autoencoder output to divide customers into four segments. These groups were interpreted as Regular Customers, Premium Customers, High Value Customers and Low Value Customers. Business cluster profiling was used to analyze the average values of sales, profit, customer satisfaction, purchase frequency and other numerical attributes for each cluster. t-SNE was used to transform the compressed feature representation into a two-dimensional graph. This helped visualize customer groups and understand whether the clusters were separated or overlapping. XGBoost was then trained using the extracted features for business risk classification. The feature importance scores generated by XGBoost were used to identify the business factors that strongly influenced prediction.

F. Algorithms

1) PCA: Principal Component Analysis is used to reduce the dimensionality of high-dimensional business data while retaining maximum information.
2) Autoencoder: Autoencoder is a neural network used to learn compressed nonlinear representations from PCA features.
3) K-Means Clustering: K-Means groups similar customers based on their business behavior and purchase characteristics.
4) t-SNE: t-SNE converts high-dimensional customer data into two-dimensional visualization for easier pattern understanding.
5) XGBoost: XGBoost is a gradient boosting algorithm used for business risk classification and feature importance analysis.

IV. EXPERIMENTAL RESULTS

The proposed framework was evaluated using standard machine learning and clustering evaluation measures. PCA information retention was used to measure dimensionality reduction quality. Autoencoder retention was used to represent compressed reconstruction quality. Silhouette score was used to evaluate clustering quality. XGBoost accuracy, precision, recall, F1-score and confusion matrix were used to evaluate classification performance.

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$F1\ Score = 2 \times (Precision \times Recall) / (Precision + Recall)$$

$$Silhouette\ Score = (b - a) / max(a, b)$$

Table.1 Performance Evaluation

Evaluation Component	Metric	Value
Dataset	Total Records	35,000
Dataset	Final Columns	24
Feature Extraction	Encoded Feature Shape	35,000 x 763
PCA	Information Retained	95.04%
Autoencoder	Information Retention	15.78%
K-Means	Number of Clusters	4
K-Means	Silhouette Score	0.2166
XGBoost	Accuracy	99.97%

Table.1 presents the experimental results obtained from the proposed business data analytics framework. The PCA module retained 95.04% of the information, indicating that dimensionality reduction preserved major data patterns. The K-Means model produced four customer clusters with a silhouette score of 0.2166. The XGBoost model achieved 99.97% accuracy in business risk classification for the current experimental dataset.

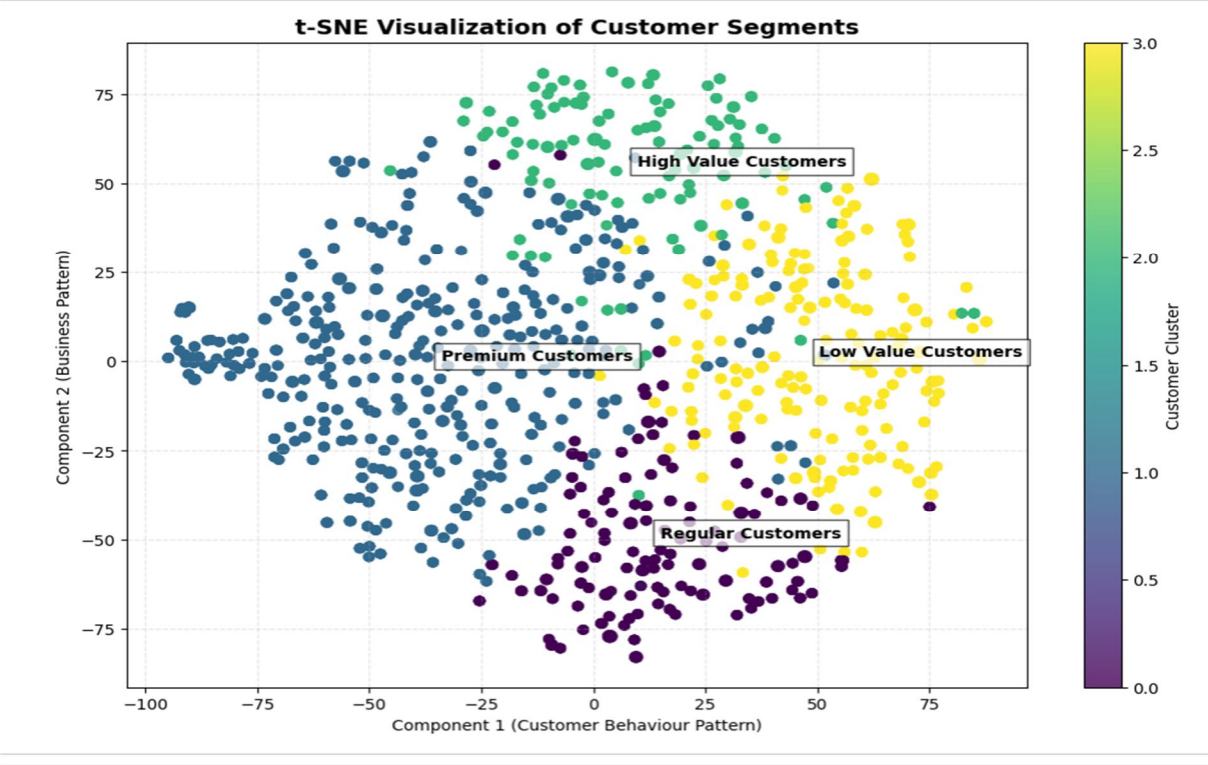


Fig.2 t-SNE Visualization of Customer Segments

Fig.2 shows the customer segments in a two-dimensional space. The visual output helps identify the distribution of Regular Customers, Premium Customers, High Value Customers and Low Value Customers. This type of visualization makes complex high-dimensional business patterns easier to understand.

Table.2 Confusion Matrix of XGBoost Classification

Actual / Predicted	Low Risk	High Risk
Low Risk	3498	2
High Risk	0	3500

The confusion matrix shows that the classifier correctly predicted 3,498 low-risk records and 3,500 high-risk records. Only two low-risk records were incorrectly predicted as high risk, while no high-risk records were missed in the current test split. This result indicates strong classification performance for the experimental setup.

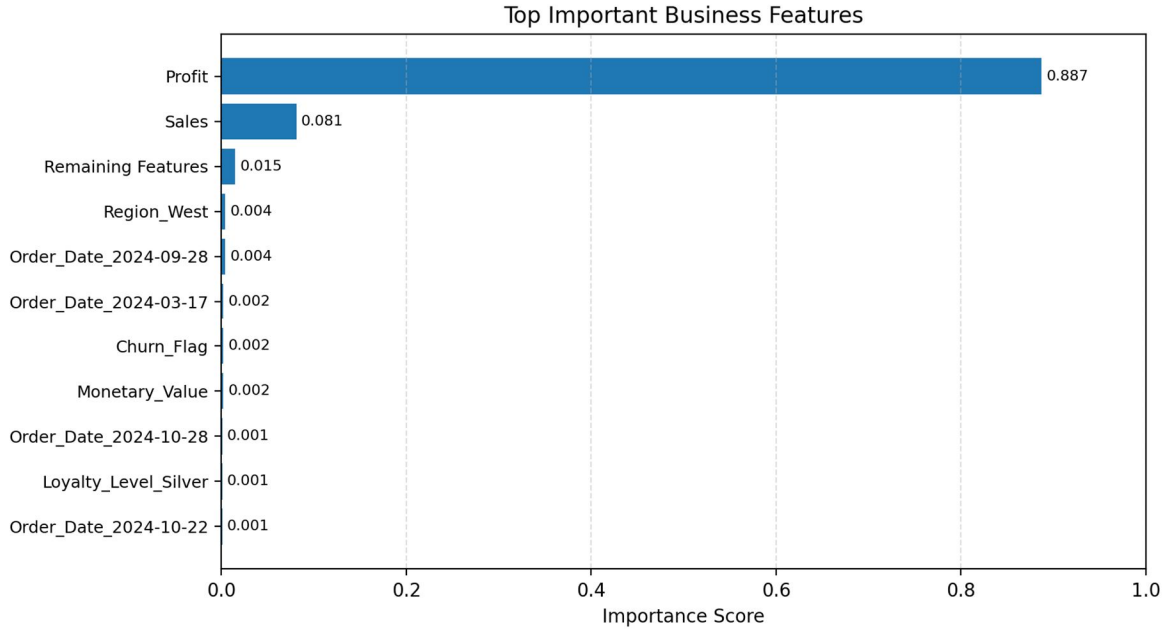


Fig.3 Feature Importance Analysis

Fig.3 shows the most influential business features identified by XGBoost. Profit obtained the highest importance score, followed by Sales and other features such as region, churn flag, monetary value, marketing spend and website visits. Feature importance analysis helps business users understand which factors affect business risk prediction most strongly.

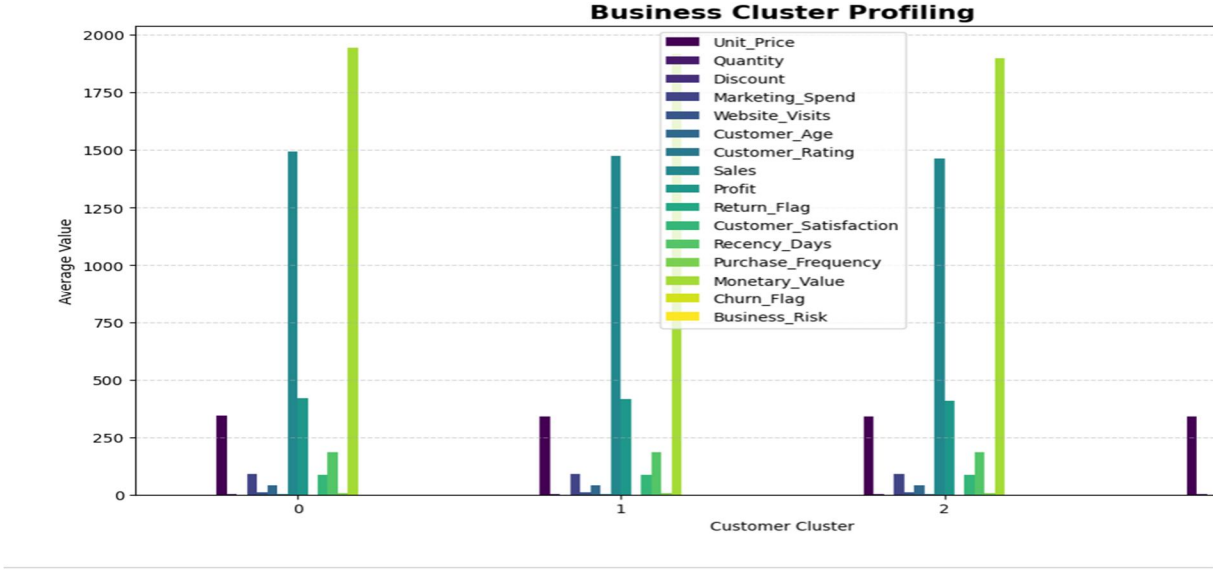


Fig.4 Business Cluster Profiling

Fig.4 represents the average business values across different customer clusters. Cluster profiling helps compare business characteristics such as sales, profit, quantity, customer satisfaction, purchase frequency and monetary value. This helps managers design different strategies for each customer group.

V. CONCLUSION

This work presented an Artificial Intelligence based business data analysis and visualization framework for ecommerce business datasets. The proposed system integrates PCA, Autoencoder, K-Means clustering, t-SNE visualization and XGBoost classification into a single analytical pipeline.

The PCA module reduced dimensionality while retaining 95.04% of important information. The Autoencoder learned compressed nonlinear features. K-Means clustering generated four meaningful customer groups, and t-SNE provided an understandable visual representation of customer segments. XGBoost achieved 99.97% accuracy for business risk prediction in the current experimental setup.

The system supports business decision-making by providing customer segmentation, cluster profiling, feature importance analysis and predictive risk evaluation. Overall, the proposed framework improves the ability to analyze complex business data and convert it into actionable insights.

VI. FUTURE SCOPE

The proposed system can be extended by using larger real-world business datasets collected from multiple industries. Future work can include real-time business analytics, dashboard integration and cloud-based deployment.

Advanced deep learning models and explainable AI techniques can be incorporated to improve prediction transparency. Integration with Power BI or web-based dashboards can make the results more interactive for business users.

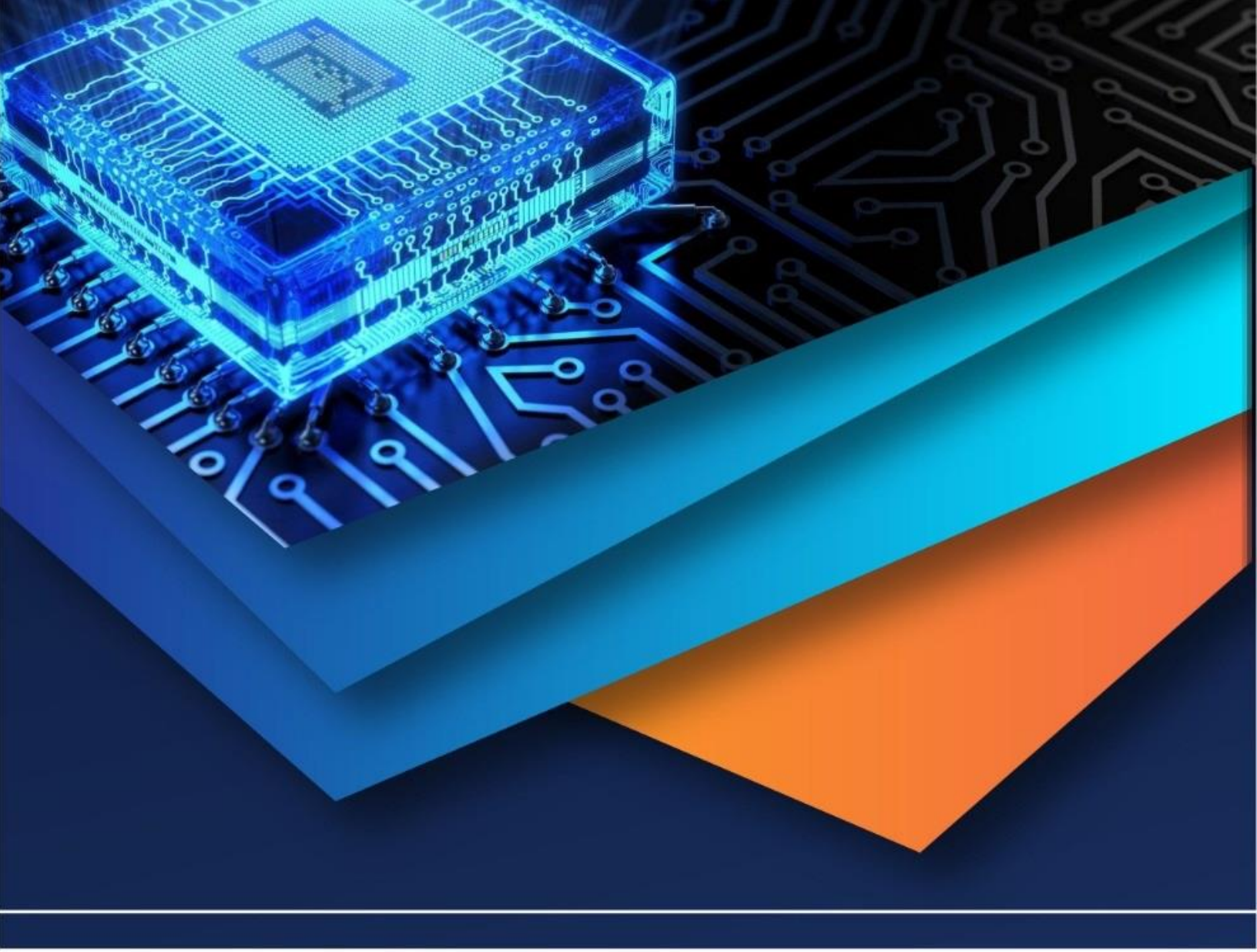
The framework can also be modified for industry-specific applications such as retail analytics, customer churn prediction, financial risk prediction, supply chain analysis and marketing recommendation systems.

VII. ACKNOWLEDGEMENT

I express my sincere gratitude to my guide, Dr. B. Kranthi Kiran, Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University Hyderabad, for his valuable guidance, continuous support and constructive suggestions throughout the development of this research work.

REFERENCES

[1] Kun Liang and Lili Zhang, “Research on New Business Data Analysis and Visualization Technology Assisted by Artificial Intelligence.”
[2] I. T. Jolliffe and J. Cadima, “Principal Component Analysis: A Review and Recent Developments,” Philosophical Transactions of the Royal Society A, 2016.
[3] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks,” Science, vol. 313, no. 5786, pp. 504-507, 2006.
[4] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.
[5] J. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.
[6] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[7] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[8] F. Chollet, “Keras,” 2015.
[9] W. McKinney, “Data Structures for Statistical Computing in Python,” Proceedings of the 9th Python in Science Conference, 2010.
[10] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY