



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14

Issue: IX

Month of publication: September 2026

DOI:

www.ijraset.com

Call:  08813907089

E-mail ID: ijraset@gmail.com

Generative AI-Powered Intelligent Data Analytics Using AutoML and LLMs

Nirlipta Pallabee Samapika¹, R. Rajesh²

¹Assistant Professor, ²M.Tech in Computer Science and Technology (CST) Sanketika Vidya Parishad Engineering College,
Vishakhapatnam, Andhra Pradesh, India

Abstract: *The Gen AI Powered Data Analyst System is a web-based intelligent analytics platform designed to automate the complete data analysis pipeline using Machine Learning and Large Language Models (LLMs). The system allows users to upload datasets and automatically performs data preprocessing, feature engineering, exploratory data analysis, visualization, and AutoML model selection. It integrates LLM-powered capabilities to generate dataset insights, explain analytical results, interpret model performance, and assist users with natural language understanding of their data. The application provides interactive visualizations, feature importance analysis, prediction results, and downloadable reports through a userfriendly Streamlit interface. By combining Machine Learning, Generative AI, and intelligent automation, the system reduces manual effort, accelerates data analysis, and supports accurate, explainable, and data-driven decision-making.*

Keywords: *Generative AI, Large Language Models (LLMs), AutoML, Machine Learning, Data Preprocessing, Feature Engineering, Streamlit, Predictive Analytics, Data Visualization, Explainable AI.*

I. INTRODUCTION

The Gen AI Powered Data Analyst System is a web-based application developed to simplify and automate the complete data analysis process using Machine Learning and Large Language Models (LLMs). Organizations generate large volumes of data, making manual analysis time-consuming and prone to errors. Traditional data analysis requires expertise in preprocessing, feature engineering, visualization, and model selection, which can be challenging for many users. To address these challenges, the proposed system provides an intelligent, automated solution for end-to-end data analytics. The application is developed using the Streamlit framework and offers an interactive and user-friendly interface. Users can upload datasets, perform intelligent preprocessing, generate advanced visualizations, engineer features, and automatically train multiple machine learning models through an AutoML pipeline. The integrated LLM generates dataset insights, explains model results, and provides business-oriented recommendations in natural language. The system also allows users to download processed datasets, trained models, and prediction reports for future use. Overall, the project provides an efficient, explainable, and AI-powered solution for intelligent data analysis and decision-making.

II. LITERATURE SURVEY

Recent research in intelligent data analytics has shifted from traditional statistical methods to automated machine learning frameworks and AI-driven data analysis techniques^{[1], [8]}. Conventional manual preprocessing and feature selection methods are often time-consuming and prone to human bias. Automated preprocessing and feature engineering techniques improve data quality and model performance by handling missing values, outliers, and redundant features efficiently^{[13], [14]}. Studies show that AutoML frameworks outperform manual model selection by automatically identifying the best-performing algorithms and optimizing hyperparameters^{[7], [8]}. Dimensionality reduction techniques such as Principal Component Analysis (PCA) further enhance computational efficiency while preserving essential information^[11]. Modern visualization libraries and interactive dashboards simplify exploratory data analysis and improve decisionmaking^{[9], [15]}. Recent advancements in Explainable AI and Generative AI have also enhanced transparency and user interaction in intelligent analytics systems^{[16], [19]}. This project builds upon these advancements by developing a Gen AI Powered Data Analyst System that integrates intelligent preprocessing, feature engineering, AutoML, and interactive visualization for efficient data analysis.

III. CHALLENGES

Developing the Gen AI Powered Data Analyst System involves several challenges related to data quality, automation, and intelligent decision-making.

Real-world datasets often contain missing values, outliers, inconsistent formats, and high-dimensional features, making preprocessing and feature engineering complex. Automatically selecting appropriate preprocessing techniques, identifying the best machine learning model through AutoML, and generating meaningful insights using Large Language Models (LLMs) require careful system design and optimization. Ensuring accurate model performance while avoiding overfitting, handling diverse dataset structures, and maintaining scalability are also significant challenges. Additionally, integrating machine learning, AutoML, data visualization, and LLM-based explanations into a single interactive Streamlit application requires efficient system integration. Therefore, developing a reliable, explainable, and userfriendly AI-powered data analytics platform demands robust data processing, intelligent automation, and seamless integration of multiple technologies.

IV. PROPOSED METHODOLOGY

The proposed system follows a structured approach for intelligent data analysis using Machine Learning and Large Language Models (LLMs). Initially, users upload a dataset, which is preprocessed by handling missing values, treating outliers, detecting data types, and performing intelligent feature engineering. The processed dataset is then passed to an AutoML pipeline, where multiple machine learning algorithms are trained and evaluated to automatically select the best-performing model. The system generates interactive visualizations, feature importance analysis, and prediction results to help users understand the data. Finally, the integrated LLM provides dataset insights, explains model performance, and generates business-oriented recommendations through a Streamlit-based web interface for user interaction.

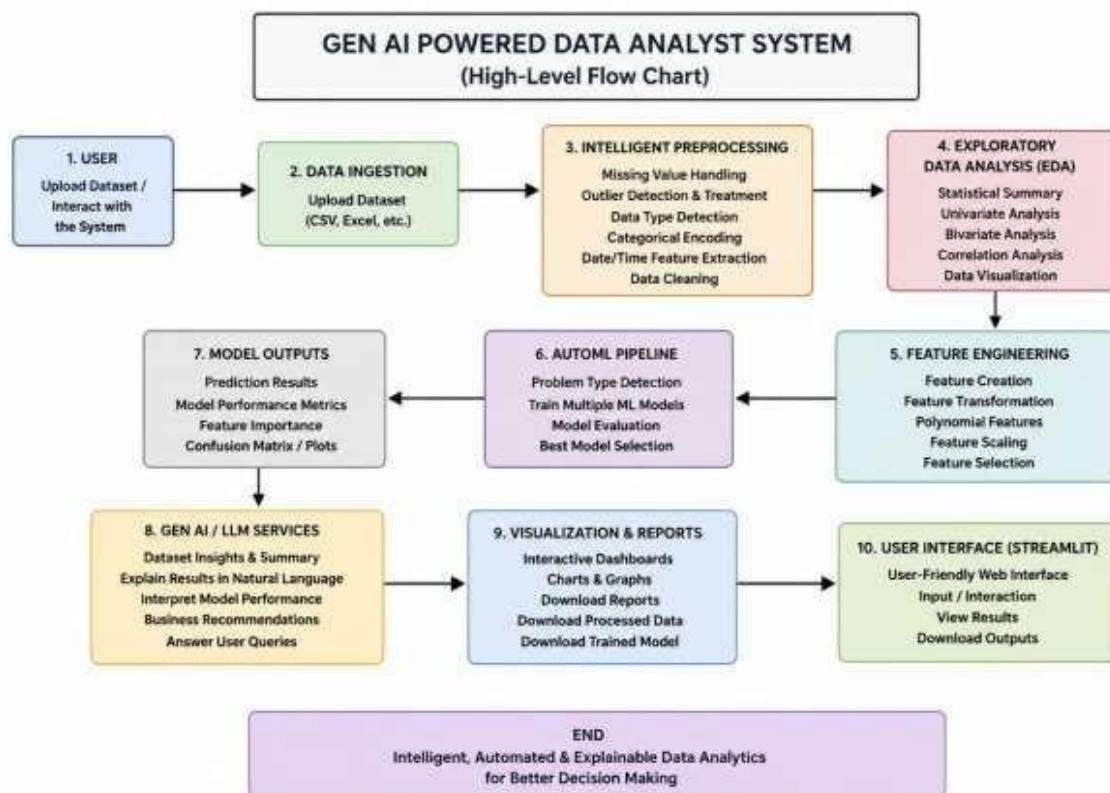


FIG1: Flow Chart Of The Proposed Methodology

V. ALGORITHMS AND TECHNIQUES

The project utilizes a combination of Machine Learning, Large Language Models (LLMs), and modern web technologies to perform intelligent data analysis and predictive modeling:

- 1) AutoML Pipeline: The AutoML module automatically detects the problem type (classification or regression), trains multiple machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Linear Regression, evaluates their performance, and selects the best-performing model based on appropriate evaluation metrics.

- 2) Feature Engineering Techniques: Intelligent feature engineering techniques such as missing value handling, categorical encoding, datetime feature extraction, interaction feature generation, polynomial features, feature scaling, feature selection, multicollinearity removal, and optional PCA are applied to improve model performance and reduce data complexity.
- 3) Large Language Models (LLMs): LLMs are integrated to generate dataset summaries, explain preprocessing and model results, provide business insights, and answer user queries in natural language, making data analysis more understandable and interactive.
- 4) Streamlit Web Framework: Streamlit serves as the frontend application, enabling users to upload datasets, perform automated analysis, visualize results, train models, and download processed datasets, trained models, and prediction reports through an interactive interface.
- 5) Data Visualization: Interactive visualizations such as statistical summaries, correlation heatmaps, distribution plots, feature importance charts, confusion matrices, ROC curves, and model comparison graphs are generated to help users understand data patterns and evaluate model performance effectively.

VI. ARCHITECTURE

The proposed Gen AI Powered Data Analyst System follows a multi-layer architecture consisting of data, processing, AutoML, LLM, and application layers. Users upload datasets through a Streamlit-based interface, where the processing layer performs data cleaning, preprocessing, and intelligent feature engineering. The AutoML layer automatically detects the problem type, trains multiple machine learning models, evaluates their performance, and selects the best model. The LLM layer generates dataset summaries, explains model results, and provides business insights in natural language. Finally, the presentation layer displays interactive visualizations, prediction results, performance metrics, and downloadable reports. This architecture ensures an efficient, automated, and user-friendly data analytics workflow.

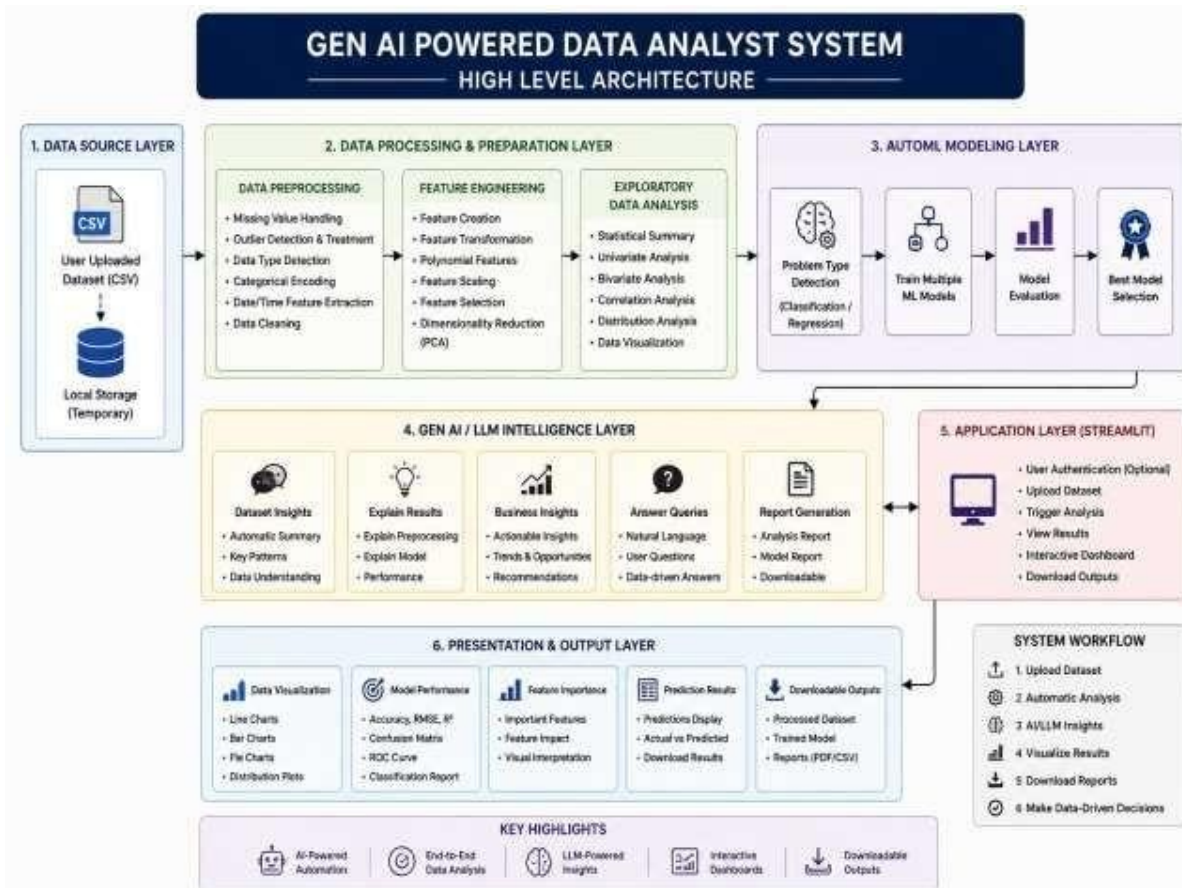


FIG2: System Architecture OfTheGen AI Data Analyst System

VII. OUTPUTS

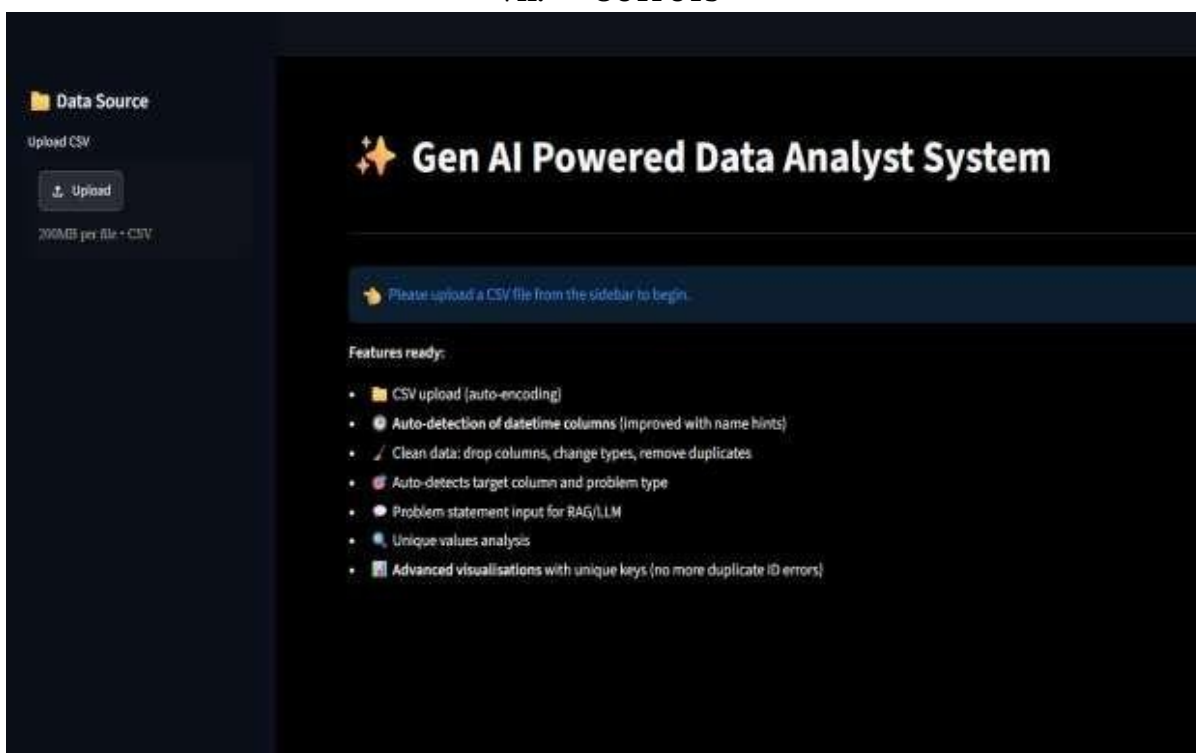


FIG3: Gen AI Powered Data Analyst System – Home Interface

The home interface provides a simple and interactive dashboard for users to upload CSV datasets and begin the data analysis process. It includes features such as automatic CSV loading, datetime detection, data cleaning, target variable identification, unique value analysis, and advanced visualizations. The intuitive design ensures that users can easily access AI-powered analytics and preprocessing tools without requiring extensive technical expertise.



FIG4: Exploratory Data Analysis And Visualization Interface



FIG5: Outlier Detection And Treatment Analysis

The system identifies outliers in the uploaded dataset using box plot visualizations and applies suitable treatment techniques to improve data quality. It displays the distribution of values before and after outlier treatment, enabling users to compare the changes effectively. This process minimizes the impact of extreme values, enhances dataset consistency, and prepares the data for more accurate machine learning model training and prediction.

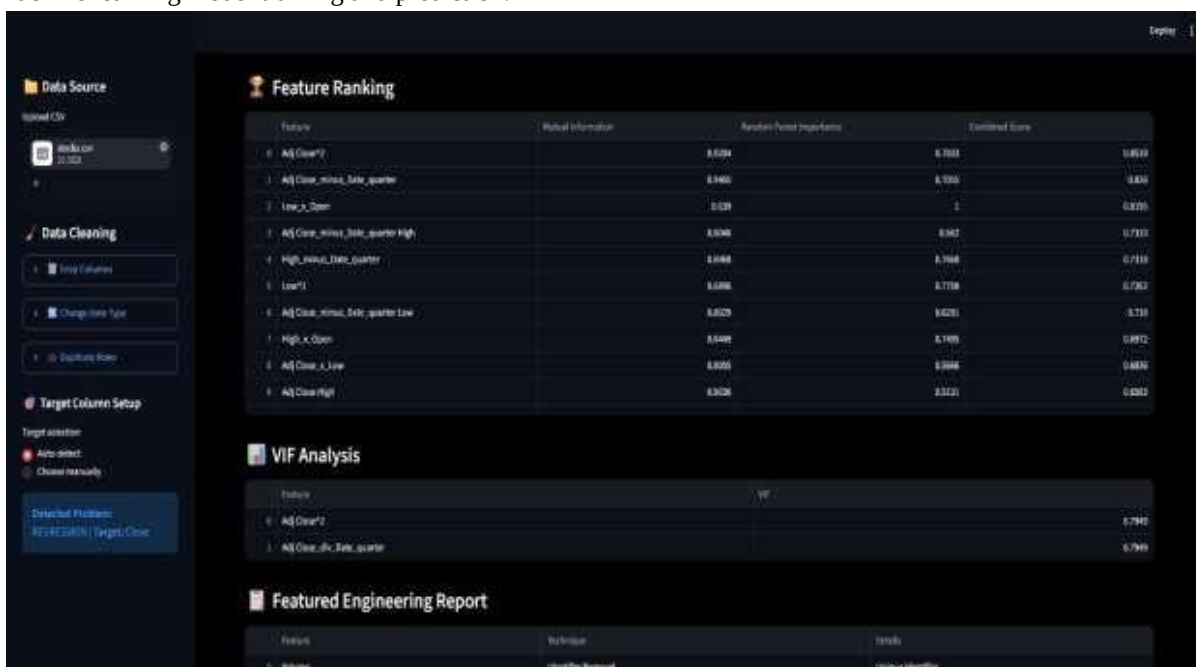


FIG6: Feature Engineering And Feature Selection Analysis

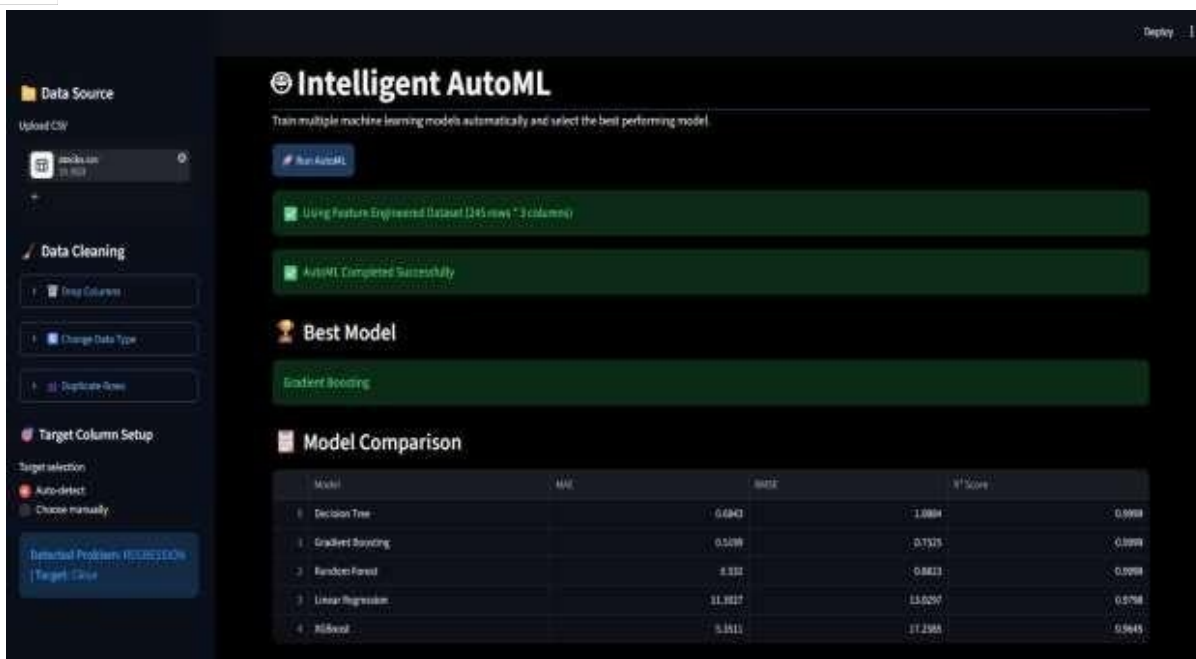


FIG7: Intelligent AutoML Model Training

The final output screen presents the predicted stock prices along with visual graphs for better interpretation.

The system enables users to observe future trends and make informed decisions based on data-driven insights. The integration of machine learning with visualization enhances usability and provides a clear understanding of stock behavior. This output demonstrates the effectiveness of the proposed prediction system.



FIG8: Performance Comparison Of Various Models

VIII. CONCLUSION

The proposed Gen AI Powered Data Analyst System provides an intelligent and efficient solution for data analysis using machine learning techniques. It automates data preprocessing, feature engineering, visualization, and model selection through AutoML, reducing manual effort and improving accuracy. Interactive visualizations enable better understanding of data patterns and model performance. The system offers a user-friendly Streamlit interface for seamless analysis and decision-making. Its modular architecture ensures scalability and flexibility for various datasets. Future enhancements can include real-time data integration, LLM-based insights, and advanced explainable AI techniques.

REFERENCES

- [1] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. [Online]. Available: <https://scikit-learn.org>
- [2] W. McKinney, "Data Structures for Statistical Computing in Python," Proceedings of the Python in Science Conference, 2010. [Online]. Available: <https://pandas.pydata.org>
- [3] C. R. Harris et al., "Array Programming with NumPy," Nature, vol. 585, pp. 357–362, 2020. [Online]. Available: <https://numpy.org>
- [4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the ACM SIGKDD Conference, 2016. [Online]. Available: <https://xgboost.readthedocs.io>
- [5] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: <https://link.springer.com/article/10.1023/A:1010933404324>
- [6] H2O.ai, "H2O AutoML Documentation," 2025. [Online]. Available: <https://docs.h2o.ai>
- [7] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, 2001. [Online]. Available: <https://projecteuclid.org>
- [8] Microsoft, "Automated Machine Learning (AutoML)," Azure Machine Learning Documentation, 2025. [Online]. Available: <https://learn.microsoft.com/azure/machine-learning>
- [9] Streamlit Inc., "Streamlit Documentation," 2025. [Online]. Available: <https://streamlit.io>
- [10] J. VanderPlas, Python Data Science Handbook, O'Reilly Media, 2016. [Online]. Available: <https://jakevdp.github.io/PythonDataScienceHandbook>
- [11] I. T. Jolliffe, Principal Component Analysis, 2nd ed., Springer, 2002. [Online]. Available: <https://link.springer.com/book/10.1007/b98835>
- [12] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2009. [Online]. Available: <https://hastie.su.domains/ElemStatLearn>
- [13] M. Kuhn and K. Johnson, Feature Engineering and Selection: A Practical Approach for Predictive Models, CRC Press, 2019. [Online]. Available: <https://bookdown.org/max/FES>
- [14] J. Brownlee, "Feature Engineering for Machine Learning," Machine Learning Mastery, 2023. [Online]. Available: <https://machinelearningmastery.com>
- [15] [15] Plotly Technologies Inc., "Plotly Python Graphing Library," 2025. [Online]. Available: <https://plotly.com/python>
- [16] [16] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed., Packt Publishing, 2019. [Online]. Available: <https://www.packtpub.com>
- [17] [17] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," Proceedings of the ACM SIGKDD Conference, 2016. [Online]. Available: <https://arxiv.org/abs/1602.04938>
- [18] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras& TensorFlow, 3rd ed., O'Reilly Media, 2022. [Online]. Available: <https://www.oreilly.com/library/view/hands-onmachinelearning/9781098125967/>

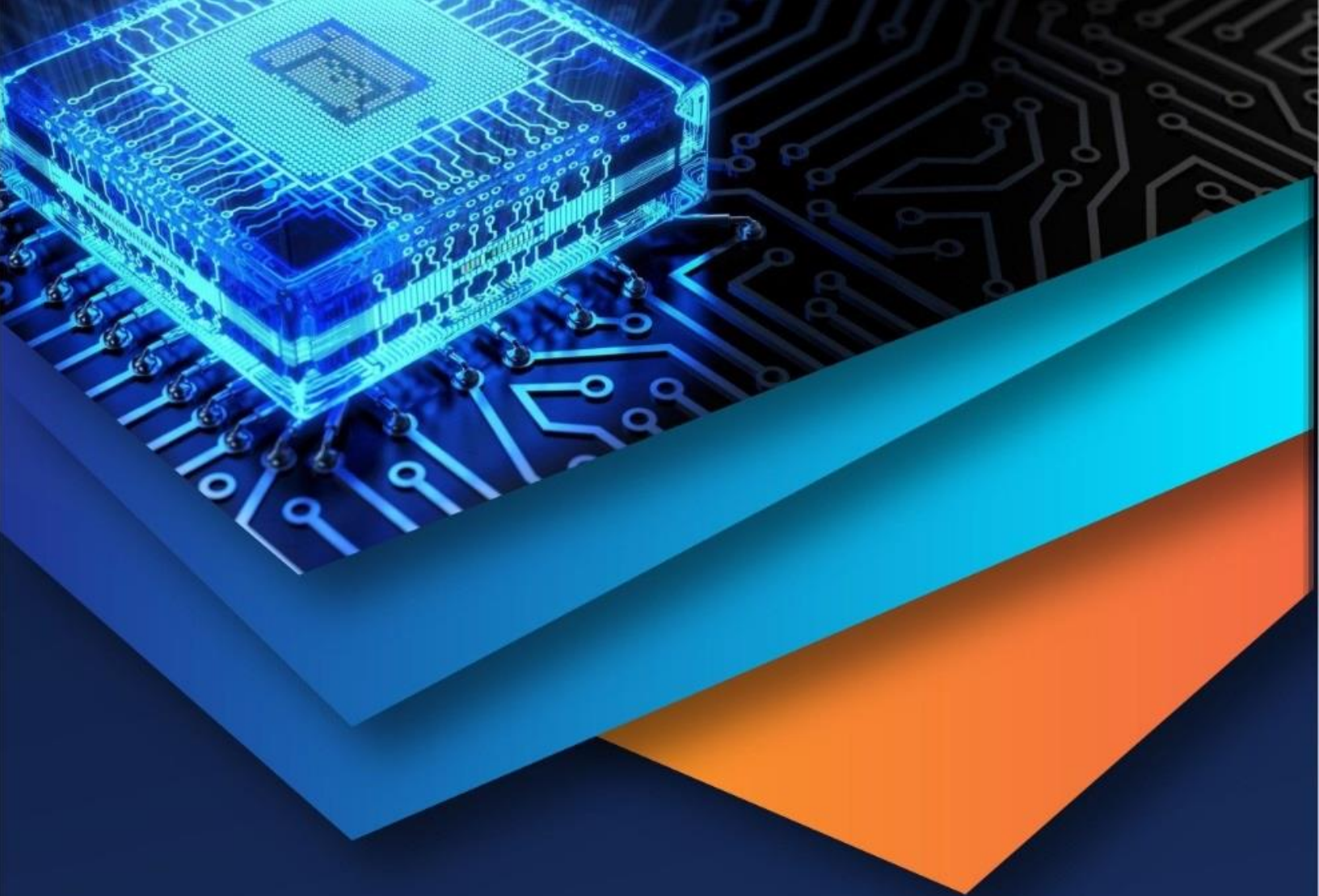
BIBLIOGRAPHY

NirliptaPallabeeSamapika is working as an Assistant Professor and holds B.Tech, M.Tech, and (Ph.D). qualifications in Computer Science. She is associated with the Alwardas Group of Technical Institutes and Sanketika Vidya Parishad Engineering College. Her areas of interest include Artificial Intelligence, Machine Learning, Generative AI, and Data Analytics. She has guided the project titled "Generative AI-Powered Intelligent Data Analytics Using AutoML and LLMs."



R. Rajesh is an M.Tech student in Computer Science and Technology (CST). His academic interests include Artificial Intelligence, Machine Learning, Generative AI, and Data Analytics. As part of his M.Tech academic work, he has undertaken the project titled "Generative AI-Powered Intelligent Data Analytics Using AutoML and LLMs" under the guidance of NirliptaPallabeeSamapika, Assistant Professor. The project focuses on integrating AutoML techniques and Large Language Models for intelligent and automated data analysis. His research work emphasizes the practical application of modern AI technologies for data-driven insights and decision-making.





10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)