International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME094
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Generative AI for Automated Data Analysis and Insight Generation

1ST Sathvika A
Assistant professor of department AI&DS
Annamacharya institute of technology and sciences Tirupati,India
andesathvika2602@gmail.com
2th P Joseph Prasanth
Dept of AIDS Annamacharya institute of technology and sciences Tirupati,India
josephprasanthreddy@gmail.com
4th Indravathi E
Dept of AIDS Annamacharya institute of technology and sciences Tirupati,India
emaniindira6@gmail.com
5th Somasankar Naik A
Dept of AIDS Annamacharya institute of technology and sciences Tirupati,India asankarnaik921@gmail.com
3th Charan Kumar S
Dept of AIDS Annamacharya institute of technology and sciences Tirupati,India
sriramulacharankumar24@gmail.com

Abstract--Data analysis constitutes a significant part of data science and machine learning with the rapid expansion of digital data to assist in decision making particularly. Data exploration and selection of appropriate machine learning models can however be tedious and require high level of expertise. This study suggests the Automated Data Analyzer which is an easy tool of Insight Creation on the dataset. The system is coded in Python and Streamlit framework and enables Google Gemini to form conclusions and suggest machine learning algorithms. The application is easy to use since all one has to do is input a dataset and the data is subsequently preprocessed, visualized, and analyzed. When dealing with large number of data, Principal Component Analysis (PCA) is used in order to compress the number of dimensions in preserving important information and hence making the analysis faster and more efficient. Statistical analysis of the system is also done to comprehend distribution of data. The insights and the visualizations are stored in an SQLite database, and a final report is created that contains insights, the visualizations and model suggestions. This method assists in the automation of the Exploratory Data Analysis (EDA).

Keywords— Exploratory Data Analysis (EDA), Generative AI, Machine Learning, Data Visualization, Anomaly Detection, PCA, Statistical Analysis, AI-driven Insights.

I. INTRODUCTION

In the current era of data-driven decision making, both organizations and individuals

increasingly depend on data analysis to support effective decisions. Despite its importance, performing data exploration and analysis manually can be complicated, time-consuming, and Automating the EDA process, the users may not consume so much money of their time on the manual data processing and interpretation. Rather, they will be able to spend more time in making informed and strategic decisions on the insights created. The other major obstacle to data analytics is that automation enhances the efficiency, as well as the fact that a user needs to select the right machine learning algorithm to use on a specific dataset. Most of the traditional methodologies rely on knowledge of experts to complete tasks like algorithm choice, data pre processing and methods of visualization. Consequently, the non-experts might be frustrated in achieving successful analysis. Hence, it is increasingly necessary to create a system where automated analysis of

datasets can be carried out as well as decisions and suggestions generated intelligently and easily accessible to many users. This type of system may facilitate the data analysis process and make it more available, efficient, and practical to a greater number of users. The main goal of the given project is to overcome these issues by developing a smart automated system that will help its users get a better insight and comprehend their data with ease.In this case, the Gemini API will also be critical to facilitating AI-driven data processing. It offers a sensible and organized method of data set analysis and appropriate choice of classifiers. Using Exploratory Data Analysis (EDA), the system will be able to identify problems like outliers, missing values, feature correlations. On these findings, it is able to suggest the right data preprocessing methods, such as scaling, encoding, or data imputation which can be used to prepare the data before further analysis and extraction of models.

II. LITERATURE SURVEY

Kothai et al. [1] also develop this idea by discussing AI-aided data visualization and proving how automation can streamline the process of performing complex analytics and enable data insights to be easier to absorb in business settings. In their work, they derive the importance of interactive visualizations to enhance the decision- making process, which is also taken into account in the proposed system. Equally, Manatkar et al. address similar models in automated analytics, which note the significance of increasingly intelligent machines that can guide their users in comprehending and interpreting information in a better manner.

Manatkar et al. [2] suggested ILAEDA, an automated exploration of data analysis that relies on imitation learning, which suggests that AI-based solutions can make the process of analyzing data simpler. Their study supports the use of automation as a way of obtaining improved decision making on the basis of the data and lessening the use of domain knowledge.

Mumuni et al. [3] have thoroughly reviewed automated feature engineering and data preprocessing methods in application in the big data and deep learning scenarios. In their paper, the authors focus on the significance of AI-supported automation when managing large volumes of data and enhancing the process of extracting features. In the same manner, Santoshkumar et al. [4] used AI and machine learning tools in predicting energy consumption, and it shows how much automated data analysis can be used in various sectors. These articles indicate that there is necessity of intelligent and non-domain specific systems of EDA which can adapt to various datasets.

Usha et al. [5] explored the elements of databases that can be improved through the use of Large Language Models (LLMs) to improve data access and analysis. Their contribution aligns with the goal of using generative AI to make the automated data insights and effective exploratory data analysis. Wu et al. [6] also talked about using AutoML methods to ease data pipelines and even enhance the efficiency of machine learning models, especially automated data preprocessing that is an essential part of the suggested system.

Zhu et al. [7] proposed Chat2Query, an automated EDA system that lacks any manually formulating queries by relying on zero-shot LLM. The way they do it shows how generative AI could allow a person with weak technical skills to analyze data more readily. Jadhav et al. [8] used a program of automated EDA to investigate the role of social media on mental health and emphasized the potential use of AI analytics to reveal the valuable findings in complex data.

Snyder et al. [9] suggested DIVI, which is an interactive contents of AI visualization capable of real-time exploration of data using dynamically updated visual representations. Their conclusions are significant because they emphasize that generative AI should be used together with interactive analytics, a significant idea in the given system. Zhao et al. [10] investigated the application of LLMs in keypoint detecting in the qualitative data analysis to enhance the procurement of relevant information in textual data and enable researchers in the interpretation process.

Sharma et al. [11] studied the application of GPT-3 to statistical data analysis and showed that language models that are large can create insightful recommendations and summaries. On the same note, Mahendru et al. [12] examined the application of AI automation in various contexts and revealed that automated data analysis can ease the manual load but enhance the accuracy and efficiency in the process of providing insights.

Chauhan et al. [13] talked about the future of AI and machine learning in the field of data science to enhance automation in data exploration. Ferreira et al. [14] contrasted the various AutoML algorithms, among which are deep learning and XGBoost, and came to the conclusion that automated model selection and hyperparameter optimization can further enhance predictive performance and minimise human effort.

Lundgard et al. [15], came up with a way of producing natural language description of visualization, which assists the user who may lack good statistical background to comprehend analytical results.

On the whole, the studies that were reviewed imply that the automated EDA, AI- controlled visualization, and machine learning model selection make considerable strides. Nevertheless, the majority of current strategies take into consideration the narrowly scoped aspects, as opposed to offering an integrated and easy-to-use environment. The proposed Automated Data Analyzer with Generative AI is an extension of all these improvements as it is proposed to combine data exploration, visualization, and machine learning recommendations to one system, which would allow closing the gap between the knowledge of data science and its availability.

III. PROPOSED WORK

The proposed work includes automatically analysing the dataset starting from preprocessing it to extracting the deeper insights. In this task, Generative AI. plays a vital role in analysing the data. The proposed workflow includes the following tasks to extract insights, generate the data visualizations, suggest the ML Algorithms, Generate the Reports etc.

A. Data Acquisition and Preprocessing

The first step of the system is to gather the datasets of users. The platform accepts various data types, and these are Comma-Separated Values (CSV), Excel files, JSON files, and database connections, such as SQLite. JSON is a data exchange format that is very lightweight and formats information in pairs that include key and value hence making it easy to store and transfer structured information. Once the dataset has been posted, the system enters the data preprocessing stage. In this step, the values that are missing are identified and processed with the help of suitable methods like imputation or dropping so that the information is complete and can be considered in further analysis. Outlier detection is also offered by the system through statistical methods, such as Z- score analysis and Interquartile Range (IQR) to detect abnormal companions and implement the required corrective actions. Also, the data can be normalized or standardized based on the need of analysis so that all the features are defined within similar range. The system is also capable of feature engineering and specifically feature extraction in order to detect the most relevant features that can enhance the performance of machine learning models. Moreover, Generative AI methods are also implemented at this point, to help in automatizing the process of feature extraction. The system is also able to suggest possible combinations and transformations of features and hence enhance the general performance of the data preparation process by examining trends in such datasets.

B. Exploratory Data Analysis and Visualization

However, once the preprocessing phase is completed, the system will execute Exploratory Data Analysis (EDA) in order to derive the meaningful information out of the dataset. The system will first compute the summary statistics including the mean, median, variance, and skewness that will assist in getting the general idea about the central tendency and the overall distribution of the data in this step. Also, a correlation table is created to evaluate the dependences between various numerical variables included in the data set. The other key step is the review of the data distribution to identify whether it is of common statistical trends like the Gaussian (normal) or the Poisson distributions. The system generates different types of visual representation to allow easier data exploration, such as histograms, boxplot, scatter plot, and heatmap. The visualizations enable the user to perceive patterns, relationships, and anomalies in the data easily. To make accessibility even more accessible, the system also uses Generative AI via the Google Gemini API to automatically generate natural language explanations to the produced visualizations. This assists non technical users to be in a better position to comprehend the insights better. Besides making the interpretation process easier, the suggestions to conduct further examination on the basis of the observed patterns

and results are also presented in the AI-generated explanations.

C. Machine Learning Model Recommendations According to the information gathered on the Exploratory Data Analysis (EDA) step, the system suggests suitable machine learning algorithms to be utilized in the further analysis. This starts with analysis of the kind of dataset to be working with; it can be classified, regression or clustering and then different modeling can be selected that will be good in that task.

The system is also good at assessing the significance of various elements in the dataset and implementation of additional data preprocessing, which can potentially improve the overall performance of the models. The portal suggests algorithms that include decision trees, support systems, neural networks, and ensembles, depending on the features of the dataset. Moreover, the system uses AutoML methods to automatically run a number of machine learning models and compare them with each other. This automated experimenting allows the user of the system to discover the model which gives a good balance between accuracy and efficiency. Besides that, Generative AI is also used to enhance the model selection process offering adaptive recommendations depending on the observed data patterns and statistical distributions in order to optimize the overall analytical process.

D. Dimensionality Reduction and Outlier Detection

In the case of large amount of data with many features, dimensionality reduction techniques, like Principal Component Analysis (PCA) are used to make the data simpler and yet still have the essential patterns and relations in it. The resultant simplification of the dataset by this process facilitates the higher-dimensional data being visualized and analyzed more easily in lower dimensional space. Moreover, the system does the anomaly detection to detect the hacky or abnormal data points in the dataset. This is done by using a combination of both statistical method and machine learning based methods that are helpful in identifying observations that are significantly different compared with the rest of the data. Detecting these anomalies is significant as such data might need to be paid special attention, investigated, or even corrected.

Report Generation and User Collaboration

In order to improve this process, the Generative AI is incorporated into the system which helps to identify the hidden patterns and anomalies that might not be easily identified by traditional statistical methods. The detection of anomalies becomes more accurate and efficient through the analysis of complicated relationships in the data with the help of AI, which will make the process more reliable.

Upon analysis, the system is able to generate a detailed PDF report containing all the important findings, visualization, and recommendations as provided by the machine learning techniques. The report covers a brief overview of the statistics analysis conducted, visualizations that were conducted as part of the exploratory data analysis and machine learning models proposed to the provided data. It also offers the information suggested by the AI where the interesting trends and trends are identified in data. Writing the report is assisted by generative AI, through the creation of intelligent summations and suggestions to dig on, depending on the type of data and the interactions of the users. The site also supports collaborative functionality to ensure people can co-work in real-time on a single dataset, make discoveries and comment on the other person. This improves overall experience of the user and simplifies the task of co-working on data analysis.

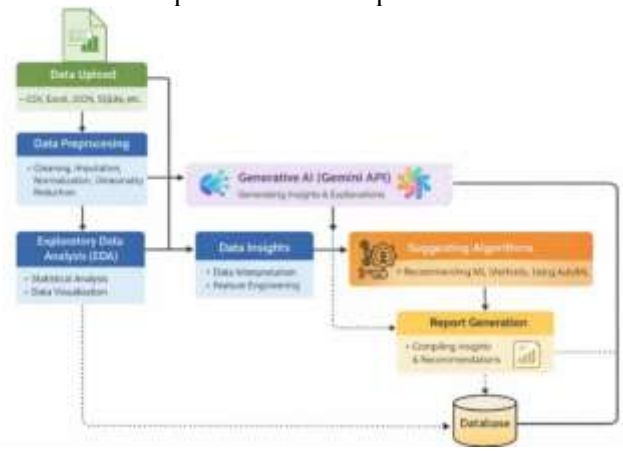

Figure 1. Architecture for Automated Data Analyser.

Figure 1 The figure above shows the design of Automated Data Analyzer system integrating the Gemini LLM. The process starts with the data preprocessing, which includes processes like work on missing values, game of numerical variables and encoding of categorical variables to prepare the dataset to be analyzed further. The Gemini API takes

statistical analysis of the dataset that has been preprocessed after performing statistical analysis on the data, it uncovers various correlations among various features and makes insights that are useful. Depending on the patterns and features identified within the data, the system also suggests appropriate machine learning algorithms which would be suitable in modeling the data. This operation constructs a full data processing loop on automated data analysis with the Gemini-powered data processing system.

IV. EXPERIMENTALRESULTS

To test the system, the system was tested on a number of datasets of varying size, such as small datasets (less than 1,000 records) and large datasets (more than 100,000 rows). The findings have indicated that the platform was in a position to recognize data distribution pattern accurately and recommend appropriate preprocessing approaches. By way of illustration, correct processing of datasets that followed normal distribution was done through the use of standardization techniques and those whose distributions were skewed were advised to go through logarithmic transformation, in order to enhance the quality of analysis. Moreover, Google Gemini API was effective enough to suggest the correct machine learning algorithms depending on the nature of the dataset. The proposed classification models were able to attain a high accuracy of about 95% with clustering models also having good results with a mean Silhouette Score of 0.85. Such results show that the system is able to offer credible analytical value and effective model suggestions. Regarding visualization, the platform created interactive charts including scatter plot, heatmap, and other graphical data visualization, allowing the user to learn the data patterns more intuitively. It was observed that these visualizations reproduced reliably and precisely with datasets of various sizes and these natural language descriptions generated by the Gemini API enhanced even further understanding of the analytical findings by users.

The system was also efficient in processing speeds

in terms of performance. Smaller datasets took less than 3 seconds to analyze, larger datasets took about 1520 seconds, and this guaranteed that users can access it smoothly and responsively. Moreover, the SQLite-based endpoint architecture containing large amounts of data was successfully used to store generated data on insights and visualizations.

Figure 2. Descriptive Statistics

Figure 2 Indicates the Data insights dashboard which presents statistical data and association of data features. The insights are automatically created referring to the dataset and the system determines to the mathematical and nominal columns. The Data Insights module deals with these functions.

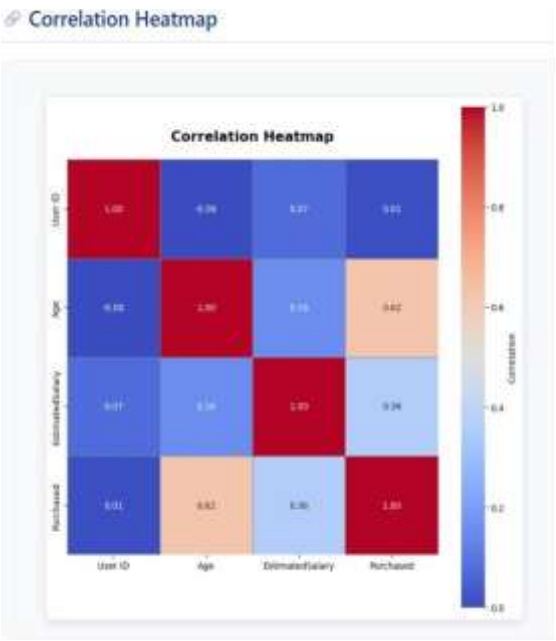

Figure 3. Data Visualization

Figure 3 Demonstrates that the Data Visualization module creates various charts of the type of a heatmap, box plot, histogram, and scatter plot depending on the choices the user made. With the aid of these visualizations, users can easily perceive the data distributions, correlations, and patterns. The system was developed with Plotly, Matplotib, and Seaborn libraries to produce interactive and quality graphs as an excellent tool of exploratory data analysis. There is also the Google Gemini API that creates a descriptive text to each visualization, enabling simple users (non technical) to comprehend complex statistical information and aiding in making decisions based on data.

Figure 4. ML Model recommendation

Figure 4 Illustrate**s** how the system recommends suitable machine learning algorithms based on the characteristics of the dataset. It analyses factors such as dataset size, feature types, data distribution, and class imbalance, while also identifying any constraints within the data. Based on this analysis,the system categorizes and suggests appropriate models for further machine learning tasks

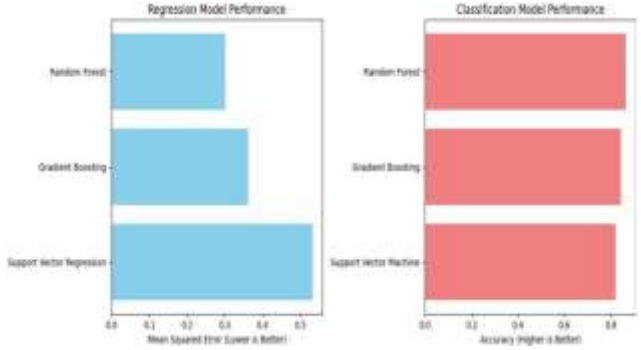

Figure 5 shows the accuracy of the machine learning algorithms recommended by the system for the Wine dataset. The system successfully identifies and suggests algorithms that achieve high performance on the given dataset. It recommends

both classification and regression models, allowing users to evaluate different approaches and select the most accurate model for analysis.

Figure 6 displays the interface of automated data analyses of the Automated Data Analyzer system. It captures the important aspects of the uploaded information such as structuring and quality. The data set in this case has 400 rows and 5 columns,where there are 4 numeric features and one categorical one. The attributes on the show include User ID, Gender, Age, Estimated Salary and Purchased. It is also ability to run a preliminary data quality check and ensure that there are no null values or that the data is balanced as well, meaning it is appropriate to run machine learning analysis on it. This summary will aid users to have a quick overview of the dataset before going on to make use of other steps of analytical task.

V. CONCLUSION

Automated Data Analyzer enhances the decision making process that is data oriented by providing a simple and efficient method of exploring data, the data analysis process (EDA). The system minimizes human work and comes with relevant and valid results because of the combination of the generative AI with machine learning automation. Such a solution illustrates the benefits of AI-related automation that simplifies the process of complex data analysis and can be of service to both researchers and a company.

The site can be used to work with high level of the dataset which becomes one of the advantages of using AI, statistics methods, and automation in the present data science practices. Additionally, the system aids in democratizing the analysis of data in that it makes advanced machine learning and statistical abilities available to users with limited technical knowledge and assists in enhancing decision making in various fields.Comprehensively, the project introduces a united automated framework of data analysis that is fuelled by the Generative AI and makes the process of exploratory data analysis faster and easier.

VII. FUTURE ENHANCEMENT

Future improvements can further enhance the capabilities of the Automated Data Analyzer. The ability to incorporate more sophisticated AI models may help the system to produce more insightful and detailed information. The use of cloud-based solutions would enable the platform to process significant amounts of data effectively, whereas the support of streaming data might provide effective analysis in real-time. Besides, automatic feature selection and model tuning might be used to find the best model and enhance the prediction accuracy with minimum efforts of users. Increased explainability of the models would aid in the transparency of anomaly detection and algorithm suggestions. Analytical reports might be made more interactive and user-friendly by using interactive dashboards and voice-based explanations. The multi-linguistic support would enhance the ease of access to more users of the product, whereas tighter security measures would be established to maintain privacy and safeguard of data. Moreover, the system might improve itself with the experience gained through its interaction with users, which allows them to make better suggestions as time passes. Other progress that can be made includes the inclusion of no-code integration, automatic report generation, and collaborative, where multiple people can collaborate on the same analysis in real time. Increased compatibility with other data and external database links would also add flexibility to the system. These enhancements would enable the technical and non-technical users to analyze and interpret data more efficiently resulting into better decision-making.

REFERENCES

[1] Kothai, G., S. Nandhagopal, P. Harish, S. Sarankumar, and S. Vidhya. "Transforming Data Visualization With AI and ML: Enhancing Business Analytics and Marketing Strategies." In Data Visualization Tools for Business Applications, pp. 125-168. IGI Global, 2025.

[2] Manatkar, Abhijit, et al. "Ilaeda: An imitation learning based approach for automatic exploratory data analysis." arXiv preprint arXiv:2410.11276 (2024).

[3] N Mumuni, Alhassan, and Fuseini Mumuni. "Automated data processing and feature engineering for deep learning and big data applications: a survey." Journal of Information and Intelligence (2024).

[4] Santhoshkumar, T., and S. Vanila. "Exploratory Data Analysis and Energy Predictions With Advanced AI and ML Techniques." In AI Approaches to Smart and Sustainable Power Systems, pp. 336-370. IGI Global, 2024.

[5] Usha, V., Nalagarla Chiru Abhinash, Sakhamuri Nitin Chowdary, V. Sathya, and Eeda Ramakrishna Reddy. "Enhanced Database Interaction Using Large Language Models for Improved Data Retrieval and Analysis." In 2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), pp. 1302- 1306. IEEE, 2024.

[6] Wu, Jiang, et al. "Data Pipeline Training: Integrating AutoML to Optimize the Data Flow of Machine Learning Models." 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE). IEEE, 2024.

[7] Zhu, Jun-Peng, Peng Cai, Boyan Niu, Zheming Ni, Kai Xu, Jiajun Huang, Jianwei Wan et al. "Chat2query: A zero-shot automatic exploratory data analysis system with large language models." In 2024 IEEE 40th International Conference on Data Engineering (ICDE), pp. 5429-5432. IEEE, 2024.

[8] Jadhav, Priya, Shravani Modhave, and Umesh Sangule. "Exploratory Data Analysis for the Survey of Social Media on Mental Health with Automation." 2023 4th International Conference on Computation, Automation and Knowledge Management (ICCAKM). IEEE, 2023.

[9] Snyder, Luke S., and Jeffrey Heer. "DIVI: Dynamically interactive visualization." IEEE Transactions on Visualization and Computer Graphics (2023).

[10]Zhao, Fengxiang, Fan Yu, Timothy Trull, and Yi Shang. "A new method using LLMs for keypoints generation in qualitative data analysis." In 2023 IEEE Conference on Artificial Intelligence (CAI), pp. 333-334. IEEE, 2023.

[11]Sharma, Ashwin, Disha Devalia, Wilfred Almeida, Harshali Patil, and Aniket Mishra. "Statistical Data Analysis using GPT3: An Overview." In 2022 IEEE Bombay Section Signature Conference (IBSSC), pp. 16. IEEE, 2022.

[12]Mahendru, Mansi, and Archana Singh. "Exploratory Analysis of AI Automation in Various Horizons." ICT Systems and Sustainability: Proceedings of ICT4SD 2022. Singapore: Springer Nature Singapore, 2022. 301-311.

[13]Chauhan, Sumika, Manmohan Singh, and Ashwani Kumar Aggarwal. "Data science and data analytics: artificial intelligence and machine learning integrated based approach." Data science and data analytics: opportunities and challenges 1 (2021).

[14]Ferreira, Luis, Andre Pilastri, Carlos Manuel Martins, Pedro Miguel Pires, and Paulo Cortez. "A comparison of AutoML tools for machine learning, deep learning and XGBoost." In 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2021.

[15]Lundgard A, Satyanarayan A. Accessible visualization via natural language descriptions: A four-level model of semantic content. IEEE transactions on visualization and computer graphics. 2021 Sep 30;28(1):1073-83.