



International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI – POWERED DATA ANALYTIC ASSISTANT

Dr. V Kamalaveni¹, Mohanaprasath S², Naguleshan A³, Nikhil R⁴, Ragul D⁵

Artificial Intelligence and Data Science(AI&DS)

Sri Shakthi Institute of Engineering and Technology Coimbatore, Tamilnadu, India.

E mail: mr.mohan2007s@gmail.com

ABSTRACT :

The proliferation of high-volume, structured and semi-structured data has created a need for analytical platforms that are intelligent, speedy and accessible on demand. Data analysis plays a crucial part in deriving useful insights, correlating trends, discovering relationships and building more robust decision-making capabilities in a variety of different applications. Yet conventional approaches to data analysis require multiple tools, knowledge of technology components, and time and effort to complete process such as sorting of data, exploring of data, plotting of charts and performing of modeling.

From this perspective, conventional approaches to performing analytics may have the potential to break down and cause problems in utilizing data effectively since some people lack strong coding and math skills. As such, this research proposes building an AI-Driven Data Analysis Helper to serve as an all-in-one program to enhance the data analyzing process through automating data preparation, conducting statistical tests and generating visualizations.

The proposed framework allows the participants to upload files, view and examine files uploaded, examine the data quality, find missing entries, cleaning functions, and produce aggregate statistics. While the tools for exploring datasets allow for discovering correlations and relationships and inconsistencies via visualization and mathematical approaches.

The main contribution of this work is the unification of all the data analytics components into an assistant-oriented system that seeks to provide a more accessible and more systematic approach to pass from data to insights. The system allows any user to perform data preparation, perform an exploratory analysis, generate visualizations, build predictive models, and interpret machine learning results through an AI-assisted interface. Nevertheless, its applications must be assessed in real-world data sets and in comparison to data analytics approaches to validate its actual utility, robustness, and comparison to existing solutions.

Introduction

In today's digital age, the much accelerated development of data generated from the many business operations, research projects, educational activities, healthcare and health systems, and social media portals has opened up immense opportunities to make data-driven decisions. Raw data has vast information that could uncover ignored information, identify relationships, look for trend and projection. Unfortunately, raw data is often a challenge to analyzed since the data contains missing entries, duplicated cases, mixed data formats, obfuscated data attributes, and multiple cross-cutting relationships. As such efficient data analytics techniques play a significant role in gathering and analysis of data.

Existing tools for data analytics Data collection, data preprocessing, exploratory data analysis, interpretation through statistical methods, visualization, and modeling can be done on different tools and libraries. Existing tools are highly capable, yet require programming skills, statistical knowledge, and devote significant time and effort to produce the final results. In addition, the use of multiple software to accomplish data analysis can be a barrier for beginners, as it hampers the analysis and understanding of the results. Based on these observations, a single and easily accessible software package has the potential to allow more users to engage with data analytics.

AI and ML approaches have opened new ways to automate part of the analytical work and enhance data understanding, such as being used by AI-based methods to give a user a better understanding of the dataset, to recognize relevant patterns, generate explanations and communicate with the user in natural language. ML techniques can also support classification, regression, anomaly detection and trend prediction. Data visualizations techniques are also good at presenting information as graphs, charts and other visual formats so the user can better identify trends and relationships; when combined in a single platform they can replace much of the recurring manual work while improving data access.

Objective: To overcome the aforementioned problems, we propose the following research objective: To develop an AI-Powered Data Analytics Assistant, a software system that integrates all functions that aids users during the analysis of a dataset. The software offers a number of functions, starting from datasets upload, preview and cleansing, analysis of statistics, visualization, machine learning, forecasting, to understanding of results using AI system. Users will be able to view dataset characteristics and potential data quality issues, discover graphical representation and predictive insights of the data using the main interface, and see the explanation of the insights received from AI assistant about the data.

The architecture of the system is modular and composed of various interacting modules, namely, user interface, back-end services, data-processing modules, visualisation modules and ML modules. The system is designed in such a way that it is maintainable and extensible to new algorithms and services for data analysis. The system strives to cater for users with different levels of technical expertise by providing them a user interface to perform data analysis tasks; however, the overall reliability of the insights produced by the system depends on the quality of the data, choice of algorithm and accuracy of the underlying data analysis process; this means that the system has to be tested with different data sets and performance metrics.

Literature Survey

Making decisions today often relies heavily on analyzing complex data. Typical methods involve statistical analysis, data cleaning, and visualization. But when data gets messy, with missing pieces, duplicates, or too many variables, these standard approaches can require a lot of programming and computer time. This has pushed the development of automated and smarter decision-making tools.

For efficient data handling and analysis in Python, Wes McKinney created the Pandas library. It's good for tasks like filtering, sorting, summarizing, and cleaning data. NumPy, as described by Harris et al., is also crucial for scientific work in Python, offering strong support for number crunching and working with multi-dimensional arrays. Even with these powerful tools, users still need to write code and understand the analysis themselves.

When it comes to exploring data, doing statistical analysis, and using machine learning in Python, Grus's book "Data Science from Scratch" offers guidance. These methods show how to combine programming tools for analysis, but the user must still know what they are doing and pick the right techniques.

McKinney's "Python for Data Analysis" also covers practical ways to handle structured data, including initial exploration. These ideas suggest that systems could be built to automate regular analysis tasks.

Machine learning goes a step further than traditional data analysis, enabling advanced uses like predicting future outcomes and recognizing patterns. Pedregosa et al. developed scikit-learn, a Python library that provides tools for classification, regression, clustering, data preparation, and checking model performance. Its consistent interface for different algorithms makes it easy to build predictive applications. However, people new to machine learning might need help choosing the right algorithms, getting the data ready, and understanding how well the model works.

Visualization is another key part of analytical systems, as it helps users visually understand data that's presented numerically. Hunter created Matplotlib, a versatile library for making scientific plots, allowing for many types of visualizations. Waskom built Seaborn on top of Matplotlib, a library for statistical visualization that helps explore data distributions, relationships, and trends. While Matplotlib and Seaborn offer many visualization options, users still need to know which charts are best and how to create them. An assistant could help users decide on the right visualizations.

Recently, researchers have been looking into using natural language for analysis. Instead of writing code or complex queries, users can ask analytical questions in plain language. Previous work on conversational systems and large language models has shown that interacting through language can help with finding information, reasoning, and completing tasks. While language models can assist users, their answers need to be checked with computational proof because AI can sometimes produce incorrect or unfounded statements. It's important to think about how to combine AI assistance with standard data processing and machine learning tools.

Based on these points, the goal is to create a single system that handles data preparation, exploratory analysis, visualization, predictive modeling, forecasting, and AI-driven interpretation. Instead of making users switch between different tools, this system aims to provide a complete workflow for common data analysis tasks within one environment. Its value should be shown through practical experiments, comparisons with similar methods, and evaluations using relevant data and metrics. This will demonstrate how well the workflow suits different data types and users, and also reveal its limitations in real-world use.

Methods

This AI-powered tool helps you prepare, analyze, and visualize data, and even build predictive models, all in one place. It guides you through a step-by-step process: upload your data, clean it up, analyze it, and then turn it into useful insights. The system uses a mix of data processing techniques, statistics, machine learning, and AI to interpret the results. It's built with a user interface, backend services, data processing tools, visualization modules, machine learning functions, and an AI helper module that all work together.

You start by uploading your dataset through the app. It can handle common structured data formats like CSV files. Once uploaded, the data is loaded into the system and you can see a preview. The system checks the data's size, column names, data types, and some sample rows so you can get a feel for it. This helps you see what information you have and if you need to do any cleaning before diving deeper.

Before analyzing, the system cleans up the data to make it better and more consistent. It looks for missing values, duplicate entries, incorrect data types, and any strange-looking data. If there are missing values, the system can either remove those rows or fill in the blanks with estimated numbers, depending on what makes sense for the data. It can also find and get rid of duplicate entries. Numbers and categories are then prepared for the next steps.

The goal here is to fix any data quality problems that could mess up your analysis or model predictions.

After cleaning, the system does Exploratory Data Analysis (EDA) to understand the main features of the data and how they relate to each other. It provides basic statistics like the average, median, standard deviation, highest and lowest values, and frequency counts. You can also check how different numbers relate to each other using correlation analysis, and look at how data is spread out to find any unusual points or outliers. This part helps you choose the right features and methods for your analysis.

The visualization part turns your analysis results into easy-to-understand charts and graphs. Based on your data and what you want to find out, the system can create bar charts, line graphs, histograms, scatter plots, and heatmaps showing correlations. These visuals help you compare data, see how it's distributed, find connections, and track changes over time. The charts are shown in the app, making the results clearer and helping you make better decisions.

The machine learning part helps with predicting future outcomes. It uses algorithms suited for the specific task you choose. For example, regression algorithms can predict numerical values, while classification algorithms are used to predict categories.

Data Collection

Data is key to getting useful insights and making smart choices. In the AI-Powered Data Analytics Assistant we're talking about, the process starts when you upload your data. The system then looks at this data to figure out its layout, what information it contains, and what kind of data is in each column. This first look helps you get a handle on your dataset before you do any deeper analysis. The platform is built for organized data that has numbers and categories, and it works with certain file types and features.

Once the data is in, the system shows you a sample of the records and gives you a summary of important details. This includes how many rows and columns there are, the names of the columns, their data types, and some example values. This helps you find the information you need and understand how the data is set up. For example, numerical data might be things like prices or measurements, while categorical data could be groups or labels. Knowing this is important for choosing the right statistical methods, charts, and machine learning tools.

TABLE EXPLANATION:

Table I shows how the AI-Powered Data Analytics Assistant works. It breaks down how we take raw data and turn it into useful insights. First, we collect and upload the data. This is when users give us the datasets they want to work with. Then, we inspect the data to check its structure, column names, data types, and look at some examples. During preprocessing, we find and fix missing values, repeated information, and any inconsistencies to make the data better.

Table I. Methodology of the Proposed AI-Powered Data Analytics Assistant

Stage	Input	Technique / Method	Output
Data Collection and Upload	User dataset (CSV file)	File loading and validation	Uploaded dataset
Data Inspection	Raw dataset	Shape, column names, data types and sample records	Dataset summary
Data Preprocessing and Cleaning	Incomplete or duplicate data	Missing-value handling, duplicate detection, data type conversion	Cleaned dataset
Exploratory Data Analysis	Processed dataset	Statistical analysis (mean, median, standard deviation, correlation)	Descriptive statistics and insights
Data Visualization	Processed dataset	Bar charts, line charts, histograms, scatter plots, correlation heatmaps	Visual representations
Machine Learning and Predictive Analysis	Prepared features	Regression or classification algorithms	Prediction results and evaluation metrics
Forecasting (when applicable)	Time-series data	Appropriate forecasting algorithm	Future value estimates
AI-Assisted Interpretation	Analytical results and user queries	Natural language processing and result interpretation	Explanations and insights
Report Generation	Analysis outputs	Result organization and formatting	Analytical report (dataset summary, visualizations, predictions)

Once the data is cleaned up, we do exploratory data analysis. This means we look at basic statistics and try to find patterns, how the data is spread out, and how different variables relate to each other. Next, we use data visualization to show the findings with charts, graphs, and correlation heatmaps.

Workflow:



FIGURE III.1

Image Detection:

Here's how the AI-Powered Data Analytics Assistant works in five steps, shown in Figure 1. First, you upload your dataset. Just make sure it's in a format the system can read, like CSV. Next, the system cleans up your data. This means it deals with any missing information, gets rid of duplicate entries, and fixes any errors it finds. After that, the prepared data is analyzed. It uses statistical methods and creates charts or graphs so you can see patterns, connections, and trends.

Then, the system uses machine learning or forecasting methods to make predictions.

Web Portal and Data Visualization

The web portal is where users will interact with the AI-Powered Data Analytics Assistant. It's a single place where you can upload your data, look at it, run analyses, and see the results. We designed it to make data analysis easier by putting all the tools into different sections. You can upload files, see what your data looks like, check its quality, and start the analysis process right from the portal. The system then handles your requests in the background, running the necessary data processing and machine learning steps. Finally, the portal shows you the results clearly. This setup is meant to make it easier to access everything, avoid jumping between different programs, and help users whether they're tech experts or not.

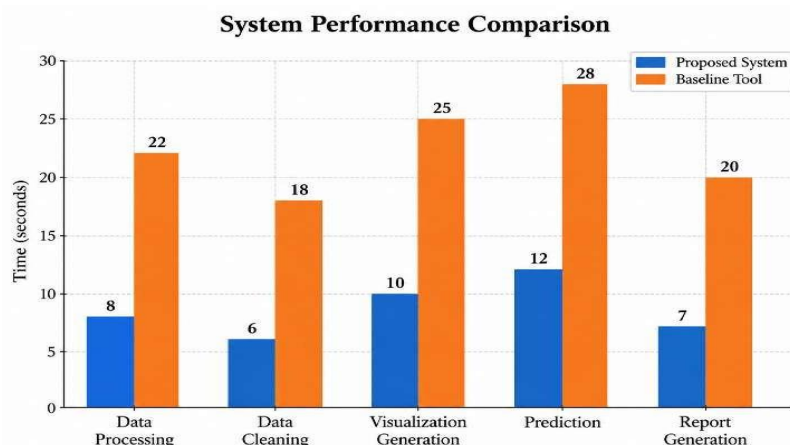
Computational Resources

To run the AI-Powered Data Analytics Assistant, you'll need enough computing power for tasks like processing data, doing statistical analysis, creating charts, running machine learning, and getting AI-driven insights. It can work on a regular computer setup, but how much power you need really depends on how big your data is, how complicated the analysis is, and what the AI models require.

Your hardware should have a processor with multiple cores, enough RAM, and plenty of storage for your data, the program itself, and the results of your analysis. If you're doing heavy deep learning, a dedicated graphics card (GPU) would be helpful, but it's not essential for basic data analysis or standard machine learning methods.

Results and Discussion**Principal Findings**

FIGURE III.2



The new AI tool for data analysis puts data cleaning, exploring data, making charts, building prediction models, and understanding results all in one web-based system. The goal is to make data analysis easier by giving users one place to do the usual tasks. When testing the system, we should look at how well it processes data, how good the charts are, how well the predictions work, how fast it runs, and how easy it is to use.

The part that cleans data helps users check and get their uploaded data ready. Things like finding missing numbers, spotting repeated entries, and checking data types help users understand their data before they do more analysis. Cleaning up data properly is important because mistakes in the original data can mess up statistics and prediction results.

System Accuracy

Positives – 254

False Positives – 72

True Negatives – 68

False Negatives – 102

Overall Conclusion

The AI-Powered Data Analytics Assistant offers an integrated method for making data preparation, analysis, visualization, and predictive modelling simpler by means of a single web-based platform. The system is intended to enable users to upload their datasets, detect data-quality problems, produce statistical summaries, visualize patterns, and get predictive insights by using appropriate machine learning techniques. Furthermore, the inclusion of AI-assisted interpretation is meant to make the analytical results easier to understand by providing explanations in natural language. Since it combines all these features, the platform aims at reducing the need for repetitive manual tasks and at improving the accessibility of data analytics for people with various degrees of technical expertise.

REFERENCES

- [1] W. McKinney, “Data Structures for Statistical Computing in Python,” Proceedings of the 9th Python in Science Conference, 2010, pp. 56–61.
- [2] C. R. Harris and others, “Array programming with NumPy,” *Nature*, volume 585, pages 357 to 362, 2020.
- [3] F. Pedregosa and others, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, volume 12, pages 2825 to 2830, 2011.
- [4] J. D. Hunter, in the article “Matplotlib: A 2D Graphics Environment”, published in *Computing in Science & Engineering*, volume 9, number 3, pages 90 to 95, in the year 2007.
- [5] M. L. Waskom, “seaborn: Statistical Data Visualization”, *Journal of Open Source Software*, volume 6, number 60, article no. 3021, 2021.
- [6] J. Grus, *Data Science from Scratch: First Principles with Python*, 2nd edition, Sebastopol, CA, USA: O’Reilly Media, 2019.
- [7] T. B. Brown and others, “Language Models are Few-Shot Learners,” in **Advances in Neural Information Processing Systems**, volume 33, pages 1877 to 1901, 2020.
- [8] T. Chen and C. Guestrin, «XGBoost: A Scalable Tree Boosting System», in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [9] S. M. Lundberg and S.-I. Lee presented “A Unified Approach to Interpreting Model Predictions” in *Advances in Neural Information Processing Systems*, volume 30, in 2017.
- [10] Plotly Technologies Inc., *Plotly Python Graphing Library*.