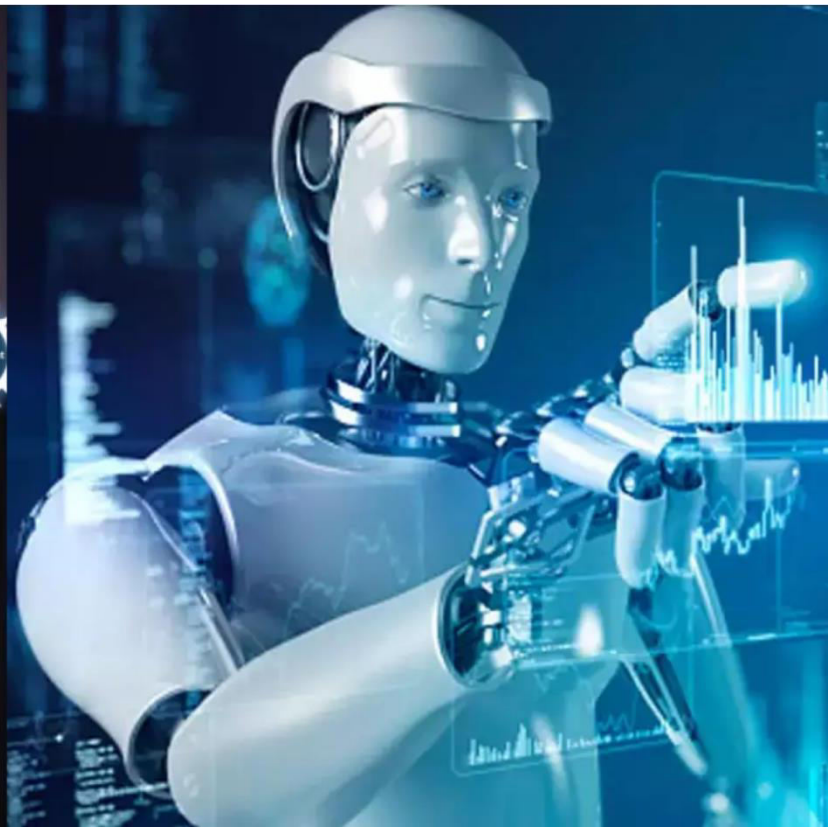
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 5, May 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

A Machine Learning and LLM-Driven Approach for Automated CSV Data Analysis using Random Forest and Local AI Models

Rutuja Bobade[1], Shaheen Khan[2], Vaishnavi Gurumukhi[3], Rachana Rahane[4], Prof. Akshada Patil[5], Prof. Suvarna Bahir[6]

U.G. Student, Department of Computer Science, Padmabhooshan Vasantdada Patil Institute of Technology, Bavdhan, Pune, Maharashtra, India

Professor, Department of Computer Science, Padmabhooshan Vasantdada Patil Institute of Technology, Bavdhan, Pune, Maharashtra, India

ABSTRACT: The proposed CSV AI Analyst is an AI-powered data analysis system designed to simplify CSV dataset analysis for both technical and non-technical users. The system enables users to upload datasets, view statistical summaries, interact with data using natural language queries, and generate machine learning predictions without requiring programming knowledge. It integrates data preprocessing, visualization, conversational AI, and predictive analytics into a single platform using React and Tailwind CSS for the frontend, FastAPI for the backend, Pandas and NumPy for data processing, Scikit-learn for machine learning, and Ollama-based local Large Language Models (LLMs) for intelligent analysis. The system runs completely on the user's local machine without requiring paid APIs, making it a cost-effective, privacy-friendly, and user-friendly solution for intelligent data analysis.

KEYWORDS: CSV Data Analysis, Machine Learning, Predictive Analytics, Exploratory Data Analysis, Anomaly Detection, Trend Analysis, Large Language Models, AI-Based Insight Generation, Data Visualization, Intelligent Analytics System.

I. INTRODUCTION

In recent years, the rapid growth of data across various domains such as healthcare, finance, education, business, and research has increased the need for efficient data analysis systems. Large amounts of structured data are commonly stored in CSV files, making data analysis an essential process for extracting useful insights and supporting decision-making. However, traditional data analysis methods often require knowledge of programming, statistics, and machine learning, which makes them difficult for non-technical users.

Existing data analysis tools provide features such as visualization, statistical analysis, and predictive modeling, but they usually require users to manually preprocess datasets, select suitable methods, and interpret results. This process can be time-consuming and complex, especially for beginners. In addition, many available tools focus only on specific functionalities rather than providing a complete integrated solution for intelligent data analysis.

Recent advancements in Artificial Intelligence (AI), Natural Language Processing (NLP), and Machine Learning (ML) have made it possible to automate several data analysis tasks. AI-powered systems can understand natural language queries, generate insights, and provide interactive communication with datasets. These technologies help users perform analysis more efficiently without requiring advanced technical skills [3] [8] [16].

To address these challenges, this paper proposes CSV AI Analyst, a full-stack AI-powered CSV analysis system that combines data preprocessing, visualization, conversational AI, and machine learning prediction into a single platform. The system allows users to upload CSV files, analyze datasets, ask questions in natural language, and generate predictions using machine learning models. The frontend is developed using React and Tailwind CSS, while the backend uses FastAPI, Pandas, NumPy, and Scikit-learn. The system also integrates Ollama-based local Large Language Models (LLMs) to provide intelligent analysis without requiring paid APIs or cloud-based services.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

The main objective of the proposed system is to make data analysis simpler, faster, and more accessible for both technical and non-technical users through an interactive and user-friendly interface.

II. RELATED WORK

In recent times, various research works have been conducted to simplify data analysis through natural language processing, machine learning, and intelligent systems [3],[16]. The main objective of these works is to minimize the complexity involved in handling data.

2.1 Liu and Xu (2025):

The authors in this study have provided an in-depth survey on "Natural Language Interfaces for Databases." In this study, they have explained how user queries can be processed through various stages, such as preprocessing, understanding, and translation into SQL form. They have also discussed different approaches to improve user query accuracy and usability [4].

2.2 Srinivas & Kumar (2025):

This research aimed at designing a speech-based database querying system. The system allowed the user to query the database using voice commands. The system processed the voice commands as well as extracted the relevant data from the databases. The research improved the system based on the ability of the system to enhance the accessibility of the system, especially for those who are not conversant with the query languages [5].

2.3 Nooralahzadeh et al. (2024):

The authors in this study have proposed an intelligent system for multi-modal data exploration through large language models. The system can be used to break down user queries into smaller tasks, such as data retrieval and visualization, as well as to generate explainable output to help users understand the output more effectively [11].

2.4 Wang et al. (2022):

This research aimed at designing intelligent data systems based on the integration of natural language processing as well as machine learning. The system aimed at analyzing the queries of the user, thus providing the relevant insights from the data. The research improved the system based on the ability of the system to enhance the interpretation of the data [8].

2.5 Yu et al. (2021):

This work aimed at improving the text-to-SQL system based on the Spider dataset. The authors proposed methods of handling complex queries as well as improving the accuracy of the system. The research also indicated the difficulties of schema understanding as well as query generalization across various datasets [19].

2.6 Liu et al. (2019):

This research aimed at designing a system based on the understanding of the follow-up queries of the user. The system aimed at maintaining the context of the queries, thus improving the efficiency of the interaction of the user with the data system [18].

III. PROBLEM STATEMENT

The rapid growth of structured data across domains such as healthcare, business, finance, and education has increased the demand for efficient data analysis systems. However, most existing data analysis tools require users to have knowledge of programming, statistics, and machine learning techniques. This makes the process difficult and less accessible for non-technical users who want to analyze datasets and extract meaningful insights.

Although modern analytical platforms provide functionalities such as data visualization, statistical analysis, and predictive modeling, users are still required to manually preprocess datasets, select suitable algorithms, and interpret results. This increases the complexity and time required for analysis. In addition, many existing systems focus only on specific tasks such as visualization, machine learning, or query processing instead of providing a complete integrated solution for intelligent data analysis.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

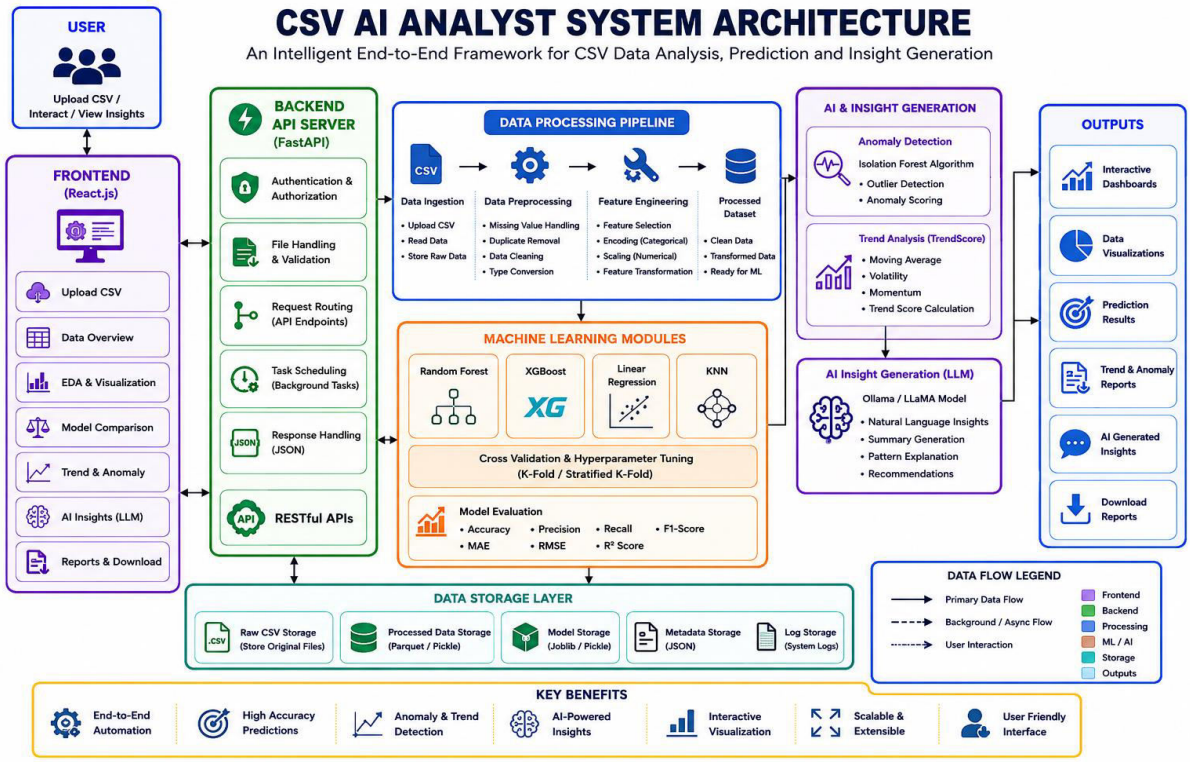
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

Another major limitation of traditional systems is the lack of natural language interaction and dependency on cloud-based APIs. Users often need to write queries or code to perform analysis, which can be challenging for beginners. Furthermore, many AI-powered platforms require paid APIs and internet connectivity, leading to higher costs and privacy concerns [4] [10] [17]. Therefore, there is a need for a unified and user-friendly system that can automate data preprocessing, support natural language queries, generate visualizations and predictions, and provide intelligent insights efficiently. The proposed CSV AI Analyst addresses these challenges by integrating AI analysis, conversational querying, machine learning prediction, and visualization into a single local full-stack platform.

IV. SYSTEM ARCHITECTURE



The architecture of the proposed **CSV AI Analyst System** is designed using a modular and scalable framework that integrates data preprocessing, machine learning, anomaly detection, trend analysis, and AI-based insight generation into a unified analytical platform. The system follows a frontend-backend architecture where the frontend handles user interaction and visualization, while the backend performs data processing and intelligent analytics.

The architecture consists of the following major components:

4.1 User Layer

The user layer represents the interaction point between the user and the system. Users upload CSV datasets, request analytical operations, visualize insights, and download generated reports through a web-based interface.

The user can perform:

- CSV dataset upload
- Exploratory data analysis
- Trend and anomaly analysis
- Machine learning prediction
- AI insight generation
- Report downloading

This layer improves accessibility for both technical and non-technical users.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

4.2 Frontend Layer

The frontend layer is developed using **React.js** and provides an interactive dashboard for visualization and user interaction [6] [14]. It communicates with the backend using RESTful APIs.

The frontend module contains:

- Upload interface for CSV files
- Data overview panels
- EDA visualization components
- Model comparison dashboards
- Trend and anomaly visualization panels
- AI-generated insight display modules
- Report generation interface

Interactive charts and graphs are dynamically rendered to improve user understanding and analytical interpretation.

4.3 Backend API Server

The backend analytical engine is implemented using **FastAPI**. It manages API communication, preprocessing operations, machine learning execution, and AI integration.

The backend performs:

- Authentication and authorization
- CSV file handling and validation
- Request routing through APIs
- Background task scheduling
- JSON response generation
- Communication with AI models

The backend acts as the core computational layer of the system.

4.4 Data Processing Pipeline

The data processing pipeline transforms raw CSV datasets into machine learning-ready structured data.

The pipeline includes:

a) Data Ingestion

The uploaded CSV dataset is read, validated, and stored for further processing.

b) Data Preprocessing

The preprocessing module performs:

- Missing value handling
- Duplicate removal
- Data cleaning
- Data type conversion

c) Feature Engineering

The feature engineering module generates meaningful analytical features using:

- Feature selection
- Categorical encoding
- Numerical scaling
- Feature transformation

The processed dataset is then forwarded to machine learning and analytical modules.

4.5 Machine Learning Module

The machine learning module is responsible for predictive analytics and model training. Multiple algorithms are implemented to improve flexibility and prediction accuracy.

The supported models include:

- Random Forest
- XGBoost
- Linear Regression

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

- K-Nearest Neighbors (KNN)

Cross-validation techniques such as:

- K-Fold Validation
- Stratified K-Fold Validation

are used for model evaluation and generalization [2].

The system evaluates models using:

- Accuracy
- Precision
- Recall
- F1-Score
- MAE
- RMSE
- R^2 Score

The best-performing model is selected automatically based on evaluation metrics.

4.6 Trend Analysis and Anomaly Detection

The analytical intelligence layer includes:

- Trend analysis
- Anomaly detection

Trend Analysis

A custom **TrendScore Algorithm** is implemented to identify:

- Moving trends
- Volatility
- Momentum
- Trend stability

Anomaly Detection

The system uses the **Isolation Forest Algorithm** to detect abnormal patterns and outliers within the dataset [7].
These analytical mechanisms improve intelligent interpretation of structured data.

4.7 AI Insight Generation Layer

The system integrates a Large Language Model (LLM) using **Ollama/LLaMA** architecture for intelligent report generation and natural language insights [1] [11] [21].

The AI module generates:

- Dataset summaries
- Pattern explanations
- Predictive interpretations
- Recommendation-based insights
- Anomaly explanations

This module converts numerical outputs into understandable human-readable insights.

4.8 Data Storage Layer

The storage layer manages different categories of system data.
The storage components include:

- Raw CSV storage
- Processed dataset storage
- Trained model storage
- Metadata storage
- Log storage

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

Processed data and trained models are stored using efficient formats such as:

- Parquet
- Pickle
- Joblib
- JSON

This improves scalability and system performance.

V. PROPOSED METHODOLOGY

The proposed **CSV AI Analyst System** follows a modular and intelligent data analytics pipeline designed to automate data preprocessing, visualization, machine learning prediction, anomaly detection, and AI-based insight generation from CSV datasets. The system is developed using a frontend-backend architecture where the frontend provides interactive visualization and user interaction, while the backend performs analytical and machine learning operations.

5.1 Data Collection and Upload

The system accepts structured datasets in CSV format uploaded by the user through a web-based interface. During the upload process, the dataset is validated to ensure correct formatting, supported file type, and data integrity. Large datasets are handled efficiently using optimized reading mechanisms to reduce memory consumption and improve processing speed.

5.2 Data Preprocessing

After successful upload, the dataset undergoes preprocessing to improve data quality and prepare it for analysis. The preprocessing stage includes:

- Handling missing values using imputation techniques
- Removal of duplicate records
- Detection and correction of inconsistent data
- Identification of numerical and categorical features
- Feature scaling and normalization using StandardScaler
- Encoding categorical attributes using One-Hot Encoding

A reusable preprocessing pipeline is created using Scikit-learn's Pipeline and ColumnTransformer modules to ensure reproducibility and consistency during training and prediction [13].

5.3 Exploratory Data Analysis (EDA)

The system performs automated Exploratory Data Analysis to understand dataset characteristics and discover hidden patterns [16]. Statistical summaries and visualizations are generated dynamically, including:

- Mean, median, variance, and standard deviation
- Correlation analysis
- Distribution plots
- Box plots for outlier detection
- Category-wise comparison charts
- Missing value analysis

For large datasets, sampling techniques are applied to improve responsiveness and computational efficiency.

5.4 Feature Engineering and Trend Analysis

The proposed system generates additional analytical features to improve prediction accuracy and trend interpretation. For time-series or sequential data, the system computes:

- Rolling averages
- Moving trends
- Volatility measures
- Momentum indicators
- Lag-based features

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

A custom **TrendScore Algorithm** is implemented to evaluate overall dataset trends using weighted combinations of slope, momentum, and stability measurements.

5.5 Machine Learning Model Training

The processed dataset is used to train multiple machine learning models for predictive analytics. Depending on the dataset type and prediction objective, the system supports:

- Linear Regression
- Random Forest
- XGBoost
- K-Nearest Neighbors (KNN)

Cross-validation techniques such as K-Fold and Stratified K-Fold validation are applied to improve model generalization and reduce overfitting. Hyperparameter tuning is performed to identify the best-performing model [2].

5.6 Model Evaluation and Comparison

The trained models are evaluated using performance metrics appropriate to the problem domain. The evaluation process includes:

- Accuracy
- Precision
- Recall
- F1-Score
- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- R^2 Score

The system compares multiple models visually and selects the optimal model based on evaluation metrics and prediction performance.

5.7 Anomaly Detection

The proposed framework incorporates anomaly detection capabilities using the Isolation Forest algorithm. The anomaly detection module identifies unusual data patterns, outliers, and abnormal observations within the dataset. Detected anomalies are highlighted visually to help users identify critical insights quickly [7] [12].

5.8 AI-Based Insight Generation

To improve interpretability, the system integrates a Large Language Model (LLM) using Ollama/LLaMA-based architecture for generating natural language analytical reports. Instead of manually interpreting graphs and statistics, users receive AI-generated summaries containing:

- Important trends
- Model performance explanations
- Dataset observations
- Anomaly interpretations
- Predictive insights

This module enhances accessibility for non-technical users by converting complex analytical outputs into understandable textual explanations [21].

5.9 Interactive Visualization Dashboard

The frontend dashboard provides interactive and real-time visualizations for user-friendly analysis. The visualization module includes:

- Dynamic charts and graphs
- Trend analysis panels
- Prediction comparison graphs
- Anomaly visualization
- AI insight display panels

The dashboard is implemented using React.js to provide responsive and smooth user interaction [14] [17].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

5.10 System Deployment and Scalability

The backend analytical engine is deployed using FastAPI/Flask architecture, enabling scalable API-based communication between frontend and backend modules. The system architecture supports future integration with:

- Cloud deployment platforms
- Real-time streaming analytics
- Advanced deep learning models
- Distributed data processing frameworks

The modular design ensures flexibility, maintainability, and scalability for future enhancements and large-scale analytical applications.

VI. EXPECTED OUTCOME

The proposed **CSV AI Analyst System** is expected to provide an intelligent, automated, and user-friendly platform for analyzing CSV datasets using machine learning and AI-based analytical techniques. The anticipated outcomes of the system are as follows:

6.1 Automated Data Analysis

The system will automate the process of data preprocessing, exploratory data analysis, visualization, and predictive analytics, thereby reducing manual effort and analysis time.

6.2 Improved Decision Making

By generating meaningful insights, trends, and predictive outcomes, the system will assist users in making accurate and data-driven decisions.

6.3 Efficient Handling of Large Datasets

The proposed framework is expected to process large CSV datasets efficiently using optimized preprocessing and sampling techniques while maintaining system performance.

6.4 Accurate Predictive Modeling

Machine learning algorithms such as Random Forest, XGBoost, Linear Regression, and KNN are expected to provide reliable prediction accuracy for classification and regression tasks [2] [15].

6.5 Detection of Hidden Patterns and Anomalies

The anomaly detection module using Isolation Forest is expected to identify abnormal patterns, outliers, and unusual observations effectively.

6.6 AI-Generated Analytical Insights

Integration of Large Language Models (LLMs) through Ollama/LLaMA is expected to generate human-readable analytical summaries and explanations automatically, improving interpretability for users [1] [11] [16].

6.7 Interactive Visualization and Reporting

The dashboard will provide dynamic charts, graphs, and trend visualizations that improve understanding of dataset characteristics and analytical results.

6.8 Reduction in Human Errors

Automation of preprocessing, model training, and analysis is expected to minimize errors caused by manual data handling and interpretation.

6.9 Support for Non-Technical Users

The system will simplify complex analytical operations through an intuitive interface and AI-generated explanations, making advanced analytics accessible to users without programming expertise.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

6.10 Cross-Domain Applicability

Integration of Large Language Models (LLMs) through Ollama/LLaMA is expected to generate human-readable analytical summaries and explanations automatically, improving interpretability for users.

VII. ADVANTAGES

1. Business Intelligence and Decision Support
   The system helps organizations analyze CSV datasets automatically, generate insights, identify trends, and support data-driven decision making without requiring advanced technical expertise.

2. Predictive Analytics
   The project can be used for forecasting sales, stock trends, customer behavior, demand prediction, and performance analysis using machine learning algorithms such as Random Forest, and Linear Regression.

3. Automated Data Analysis
   It reduces manual effort in preprocessing, visualization, and model selection by automating exploratory data analysis (EDA), prediction, and reporting tasks.

4. Multifunctional Data Analysis by one platform
   This system combines all features into a single platform, offers a good user experience, for traditional tools which need a separate software for machine learning, statistics, and visualization.

5. Natural Language Insight Generation
   Integration with AI models such as LLaMA enables automatic generation of textual insights and explanations from numerical data, improving interpretability for non-technical users.

6. Improved Decision-Making Process
   AI-generated analytical reports assist organizations in making faster and more informed strategic decisions.

7. Adaptability to Various Datasets
   The system can be used in a various domains because it supports CSV files, including:
   - Analytics for business
   - Healthcare
   - Finance
   - Education

8. Data Democratization
   The system enables non-programmers to perform advanced analytics through a simple interface, making AI-powered data analysis accessible to a wider audience.

9. Multi-Domain Dataset Processing
   The framework can be applied in domains such as e-commerce, banking, education, healthcare, agriculture, and social media analytics due to its generic CSV-based architecture.

VIII. FUTURE SCOPE

The future scope of the CSV AI Analyst project is highly promising, as it can be enhanced with several advanced features to improve both functionality and user experience. In the future, the system can be expanded to support multiple file formats such as Excel, JSON, and database integration, instead of being limited to CSV files only. This enhancement will make the platform more flexible and useful for handling different types of datasets. Additionally, advanced Artificial Intelligence and Machine Learning algorithms can be integrated to improve prediction accuracy and provide more meaningful insights from data [15] [16].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

Furthermore, the project can include interactive data visualization dashboards with charts and graphs to help users better understand trends, patterns, and relationships within datasets. The Natural Language Processing (NLP) capabilities can also be improved, allowing users to ask more complex questions in simple language and receive accurate responses [3] [4] [17]. Features such as automated report generation in PDF or Excel format can further enhance usability and save users time.

In addition, cloud deployment and collaboration features can be introduced, enabling users to access the platform online and work on projects from different locations. Security enhancements, such as user authentication and secure data management, can also make the system more reliable and suitable for professional and enterprise-level applications. Overall, these future developments can transform the CSV AI Analyst into a more intelligent, scalable, and efficient AI-powered data analytics platform.

IX. CONCLUSION

The CSV AI Analyst project successfully demonstrates the practical implementation of Artificial Intelligence and Machine Learning in data analysis. By integrating technologies such as React, FastAPI, Ollama, Pandas, and Scikit-learn, the system provides an efficient platform for users to upload CSV files, analyze datasets, interact with AI, and generate predictions. The project simplifies complex data analysis tasks and enables users to gain meaningful insights from data in an easy and user-friendly manner. Since the system runs entirely on a local machine without requiring API keys, it also ensures privacy, cost-effectiveness, and accessibility.

Furthermore, the project highlights the importance of AI-powered tools in improving decision-making and reducing manual effort in data processing. Features such as AI-based analysis, chat functionality, and machine learning predictions make the application more interactive and intelligent. The integration of local Large Language Models (LLMs) through Ollama enhances the analytical capabilities while maintaining complete control over user data.

Overall, the CSV AI Analyst project serves as an innovative and practical solution for modern data analysis needs. It combines advanced technologies with simplicity, making data analytics more accessible to users with different levels of technical expertise. With future enhancements and additional features, the project has strong potential to evolve into a more powerful, scalable, and comprehensive data analytics platform.

REFERENCES

[1] X. Su, Y. Gu, P. Wang, and W. Gu, “A Robust Natural Language Text-to-SQL Generation Framework with Dynamic Strategies Based on LLMs,” Scientific Reports, 2026. :contentReference[oaicite:3]{index=3}
[2] G. C. Tătaru, “Understanding the Rise of Automated Machine Learning,” Information, 2025
[3] Tsinghua DB Group, “A Survey of Text-to-SQL in the Era of Large Language Models,” 2025.
[4] M. Liu and J. Xu, “NLI4DB: A Systematic Review of Natural Language Interfaces for Databases,” arXiv preprint arXiv:2503.02435, 2025. :contentReference[oaicite:0]{index=0}
[5] M. Srinivas and P. Kumar, “Speech-Based Natural Language Interface for Databases,” International Journal of Creative Research Thoughts (IJCRT), vol. 13, no. 5, 2025. :contentReference[oaicite:2]{index=2}
[6] Q. Chen et al., “How Does Automation Shape the Process of Narrative Visualization: A Survey of Tools,” IEEE TVCG, 2024.
[7] M. S. Desai, “A Review of Predictive Analytics: Techniques, Advantages, and Applications,” 2024.
[8] W. Zhang et al., “Natural Language Interfaces for Tabular Data Querying and Visualization,” IEEE TKDE, 2024.
[9] Lim et al., 2024 “Forecasting with Deep Learning: A Review of Recent Advances,” Pattern Recognition
[10] Narechania et al., 2024 “DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization,” ACM CHIConference.
[11] F. Nooralahzadeh, Y. Zhang, J. Furst, and K. Stockinger, “Explainable Multi-Modal Data Exploration in Natural Language via LLM Agent,” arXiv preprint arXiv:2412.18428, 2024. :contentReference[oaicite:1]{index=1}
[12] Y. Gong et al., “A Survey on Dataset Quality in Machine Learning,” Information and Software Technology, 2023.
[13] S. Mladenovic et al., “Automated Data Preparation for Machine Learning,” 2023.
[14] Moritz et al., 2023 “Automated Visualization Recommendation: State of the Art and Future Directions,” IEEE Computer Graphics & Applications
[15] Hutter et al., 2023 “Automated Machine Learning: Methods, Systems, and Challenges,”Springer Nature

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505183|

[16] Luo et al., 2023 "AI-Augmented Data Analytics: A Survey on Automated Insight Generation," ACM Computing Surveys
[17] Setlur et al., 2023"NL4DV: A Toolkit for Generating Data Visualizations from Natural Language Queries," IEEE TVCG
[18] Q. Liu, B. Chen, J. G. Lou, G. Jin, and D. Zhang, "FANDA: A Novel Approach to Perform Follow-up Query Analysis," arXiv preprint arXiv:1901.08259, 2019. :contentReference[oaicite:4]{index=4}
[19] T. Yu et al., "Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task," Proceedings of EMNLP, 2018.
[20] Chang, S., et al. (2024). Spider 2.0: Evaluating Text-to-SQL Systems in Realistic Settings. ArXiv.
[21] Zhu, C., et al. (2024). A Survey on Employing Large Language Models for Text-to-SQL Tasks. arXiv.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details