International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

AUTONOMOUS AI AGENTS FOR DATASCIENCE

[1] K Ravi Naik, [2] K Vivekananda, [3] K Neeraj, [4] K Anil Kumar, [5] M Santhosh
[1]AssistantProfessor, [2345]Students
Department of AIML
Siddhartha Institute of Technology & Sciences, Narapally

ravinaik_cse@siddhartha.co.in, 24tq1a6667@siddhartha.co.in, 24tq1a6678@siddhartha.co.in, 24tq1a6680@siddhartha.co.in, 24tq1a6697@siddhartha.co.in

Abstract

In the modern era of data-driven decision-making, data science has become a crucial field for extracting meaningful insights from large volumes of data. However, traditional data science workflows involve multiple complex and time-consuming steps such as data preprocessing, exploratory data analysis (EDA), model selection, training, and evaluation. These processes require significant human effort, technical expertise, and continuous monitoring, making them challenging for beginners and inefficient for large-scale applications. To address these issues, this project proposes the development of an Autonomous AI Agent for Data Science, which aims to automate the entire data science pipeline.

The proposed system is designed to perform key data science tasks with minimal human intervention. It begins by allowing users to upload datasets through a user-friendly interface. The system then automatically processes the data by handling missing values, removing inconsistencies, and preparing it for analysis. Following this, exploratory data analysis is conducted to identify patterns and relationships within the dataset. Based on the nature of the data, the system intelligently selects appropriate machine learning models, trains them, and evaluates their performance using standard metrics.

The implementation of this system is carried out using Python and various libraries such as Pandas, NumPy, Matplotlib, and Scikit-learn. Additionally, Streamlit is used to develop an interactive web-based interface that simplifies user interaction and enhances accessibility. The modular architecture of the system ensures flexibility, scalability, and ease of maintenance, allowing it to be extended with additional features in the future.

I. Introduction

In the modern digital era, data has become one of the most valuable resources for organizations, businesses, researchers, and industries. The increasing availability of large volumes of data has created a strong demand for efficient data analysis and

intelligent decision-making systems. Data science plays a crucial role in extracting meaningful insights from raw data through processes such as data collection, preprocessing, visualization, machine learning, and predictive analysis. However, performing these tasks manually requires considerable technical expertise, time, and continuous effort. Traditional data science workflows are often complex and challenging, especially for beginners, non-technical users, and organizations with limited resources. To address these challenges, the concept of Autonomous AI Agents for Data Science has emerged as an innovative and efficient solution. Autonomous AI agents are intelligent systems capable of performing tasks independently with minimal human intervention. By integrating Artificial Intelligence (AI), Machine Learning (ML), and automation techniques, these agents can streamline and automate various stages of the data science lifecycle. Such systems reduce manual workload, improve productivity, minimize human errors, and accelerate the process of generating data-driven insights.

This project focuses on the design and development of an Autonomous AI Agent capable of automating key data science operations. The proposed system is designed to perform multiple tasks automatically, including dataset loading, data preprocessing, handling missing values, exploratory data analysis (EDA), feature selection, machine learning model selection, model training, evaluation, and result visualization. The agent intelligently analyzes the dataset and selects suitable machine learning algorithms based on the nature of the problem, thereby simplifying complex workflows into an easy-to-use automated pipeline.

The implementation of the system is carried out using Python and several widely used data science libraries such as Pandas, NumPy, Matplotlib, and Scikit-learn. A user-friendly interface is developed using Streamlit to allow users to interact with the system easily without requiring advanced programming knowledge. This makes the system accessible to students, researchers, beginners, and professionals who want to perform data analysis and machine learning tasks efficiently.

II. Literature Survey

The field of data science has evolved rapidly over the past decade due to significant advancements in Artificial Intelligence (AI), Machine Learning (ML), and large-scale data processing technologies. Traditional data science workflows involve several stages such as data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model selection, training, evaluation, and deployment. These tasks are often performed manually and require extensive technical knowledge, making the process time-consuming, complex, and less accessible to beginners. To overcome these challenges, researchers and organizations have increasingly focused on automation techniques, leading to the development of Automated Machine Learning (AutoML) systems and intelligent AI agents capable of simplifying machine learning workflows.

Automated Machine Learning, commonly known as AutoML, is an emerging approach designed to automate the end-to-end process of applying machine learning to real-world problems. AutoML systems automatically select suitable machine learning algorithms, optimize hyperparameters, and improve model performance with minimal human intervention. Platforms such as Google AutoML and H2O.ai have gained popularity for simplifying complex machine learning tasks and making them accessible to non-experts. Google AutoML provides cloud-based services for tasks such as image classification, natural language processing, and structured data analysis, while H2O.ai offers open-source AutoML capabilities including automatic feature engineering and model interpretability. Although these platforms improve efficiency and reduce manual effort, most AutoML systems primarily focus on model selection and optimization while lacking complete automation of the entire data science pipeline, particularly in data preprocessing and domain-specific customization.

Research on AI agents and intelligent systems has further expanded the capabilities of automation in data science. AI agents are intelligent systems capable of perceiving their environment, making decisions, and taking actions autonomously to achieve specific objectives. These systems are increasingly used in domains such as healthcare, finance, robotics, recommendation systems, and intelligent automation. In the context of data science, autonomous AI agents can independently analyze datasets, identify patterns, choose suitable machine learning algorithms, and evaluate model performance without continuous human intervention. Researchers have focused on integrating machine learning with intelligent decision-making mechanisms to create adaptive and self-improving systems that function as intelligent assistants for data scientists. Concepts such as reinforcement learning, reasoning, and knowledge representation further enhance the ability of AI agents to learn from experience and improve performance over time.

Several frameworks and libraries have been developed to support machine learning and data science tasks efficiently. Scikit-learn is one of the most widely used Python libraries for implementing algorithms related to classification, regression, clustering, and model evaluation. Its simplicity and ease of use make it highly popular among beginners and professionals. Pandas is extensively used for data manipulation, cleaning, transformation, and structured data analysis through DataFrame-based operations. Similarly, NumPy provides support for numerical computing and multi-dimensional array operations, forming the backbone of scientific and analytical applications in Python. While these libraries provide flexibility, scalability, and powerful functionalities, they still require significant manual coding and technical expertise for effective usage.

Despite the progress made in AutoML tools and intelligent data science frameworks, several limitations remain unresolved. One major limitation is the lack of flexibility and transparency in many AutoML systems, which often function as black-box models with limited customization options. This becomes problematic when dealing with domain-specific problems or highly complex datasets that require tailored

solutions. Another limitation is partial automation, as many tools automate model training and hyperparameter tuning but still require manual intervention for preprocessing, feature engineering, and visualization tasks. Additionally, many current systems are not beginner-friendly because they require prior knowledge of machine learning concepts and programming skills. Scalability and integration challenges also exist when handling large-scale or real-time data processing across multiple tools and platforms.

The literature survey concludes that although AutoML platforms and machine learning frameworks have significantly improved the efficiency of machine learning workflows, they do not yet provide a completely autonomous and intelligent data science solution. Autonomous AI agents present a promising approach by combining automation, reasoning, adaptability, and decision-making capabilities into a single framework. This project aims to address existing limitations by developing an intelligent system capable of integrating data preprocessing, exploratory analysis, machine learning model training, evaluation, and visualization into a seamless automated pipeline. The proposed system seeks to improve usability, efficiency, scalability, and accessibility for both technical and non-technical users in modern data science applications.

III. System Analysis

In the modern digital era, organizations and industries generate massive amounts of data every day, making data science an essential process for extracting meaningful insights and supporting decision-making. Traditional data science workflows involve multiple complex stages such as data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model selection, training, evaluation, and deployment. These tasks often require strong technical expertise, significant manual effort, and considerable time to complete. Many beginners and non-technical users face difficulties in performing these operations efficiently due to the complexity of machine learning tools and programming requirements. Existing systems also require continuous human intervention during different stages of analysis, reducing productivity and scalability. Autonomous AI agents provide a promising solution by automating repetitive and complex tasks using Artificial Intelligence (AI), Machine Learning (ML), and intelligent decision-making techniques. The proposed system is designed to automate the complete data science workflow while reducing manual effort and improving efficiency. The system analyzes datasets, performs preprocessing, selects suitable machine learning models, trains and evaluates them automatically, and generates meaningful insights. The analysis focuses on improving automation, scalability, accuracy, user accessibility, and ease of use while minimizing human intervention and operational complexity. Overall, the system demonstrates the growing importance of intelligent AI-driven automation in modern data science applications.

Existing System

The existing data science workflow mainly depends on manual processes and traditional machine learning tools that require extensive programming knowledge and technical expertise. Data scientists and analysts must perform several tasks manually, including data cleaning, handling missing values, exploratory data analysis, feature engineering, model selection, hyperparameter tuning, training, evaluation, and visualization. Although machine learning libraries such as Scikit-learn, Pandas, and NumPy simplify certain operations, users still need to write large amounts of code and understand machine learning concepts thoroughly. Existing AutoML platforms automate some aspects of model selection and tuning, but they often lack complete workflow automation and flexibility. Most systems also function as black-box models, limiting transparency and customization. Beginners may find these systems difficult to use because they require technical knowledge of data science, statistics, and programming. Existing systems are also time-consuming when handling large datasets and complex workflows. Additionally, integrating preprocessing, visualization, training, and deployment across multiple tools can create fragmented and inefficient workflows.

Disadvantages of Existing System

- Requires strong programming and technical expertise
- High dependency on manual data preprocessing and analysis
- Time-consuming machine learning workflow
- Existing AutoML tools provide only partial automation
- Difficulty for beginners and non-technical users
- Limited flexibility and customization in AutoML platforms
- Fragmented workflows across multiple tools and libraries
- High computational and operational complexity
- Manual model selection and hyperparameter tuning required
- Limited scalability for large and complex datasets
- Increased chances of human errors during analysis
- Lack of intelligent decision-making and adaptive automation

Proposed System

The proposed system is an Autonomous AI Agent for Data Science designed to automate the complete machine learning and data analysis workflow with minimal human intervention. The system integrates Artificial Intelligence (AI), Machine Learning (ML), and intelligent automation techniques to perform data science tasks automatically and efficiently. The AI agent is capable of loading datasets, cleaning and preprocessing data, handling missing values, performing exploratory data analysis (EDA), selecting suitable machine learning algorithms, training models, evaluating performance, and generating visual insights. Unlike traditional systems, the proposed solution intelligently analyzes the dataset and automatically determines the best processing and modeling strategies. The system uses popular Python libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib to support efficient data analysis and

machine learning operations. A user-friendly interface is developed using Streamlit to enable users to interact with the system without requiring advanced coding knowledge. The proposed system improves efficiency, reduces manual effort, enhances accessibility, and simplifies complex workflows into a seamless automated pipeline. Overall, the system provides an intelligent, scalable, and user-friendly solution for modern data science automation.

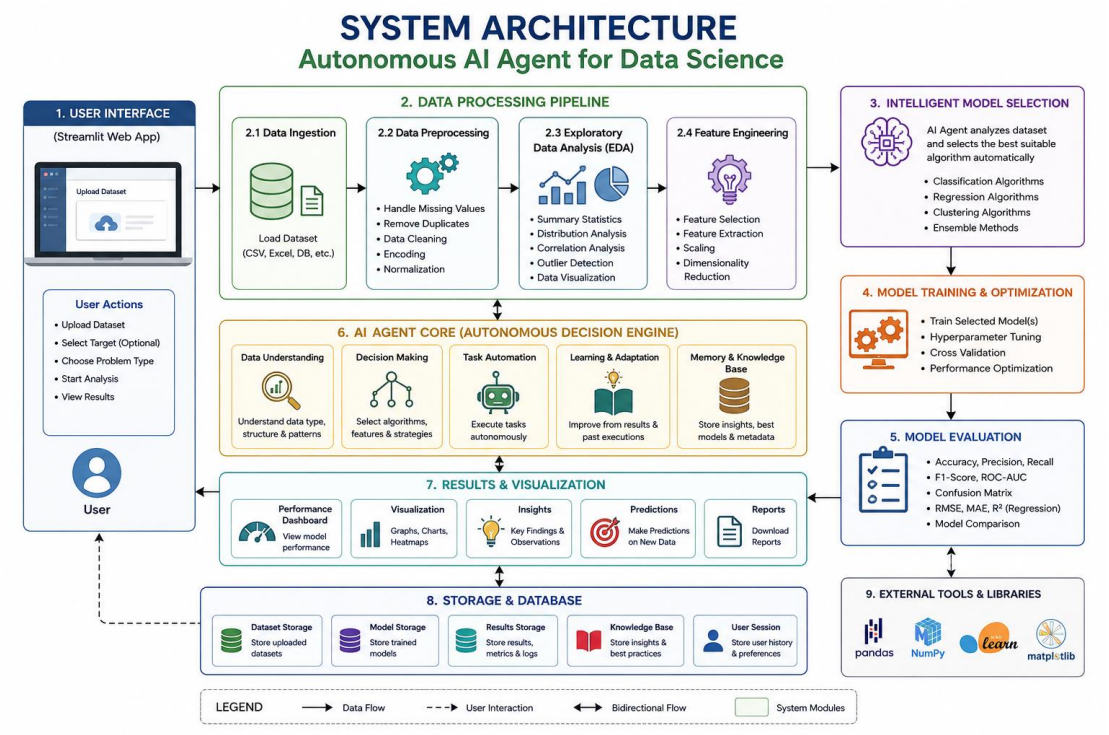
Advantages of Proposed System

- Automates the complete data science workflow
- Reduces manual effort and human intervention
- Saves time in data preprocessing and model training
- Beginner-friendly and easy-to-use interface
- Automatically selects suitable machine learning models
- Improves productivity and workflow efficiency
- Reduces human errors during analysis
- Supports intelligent decision-making and automation
- Provides automated visualization and performance evaluation
- Scalable for handling large datasets and workflows
- Integrates preprocessing, training, and evaluation into one system
- Accessible to students, researchers, and professionals

IV. Methodology

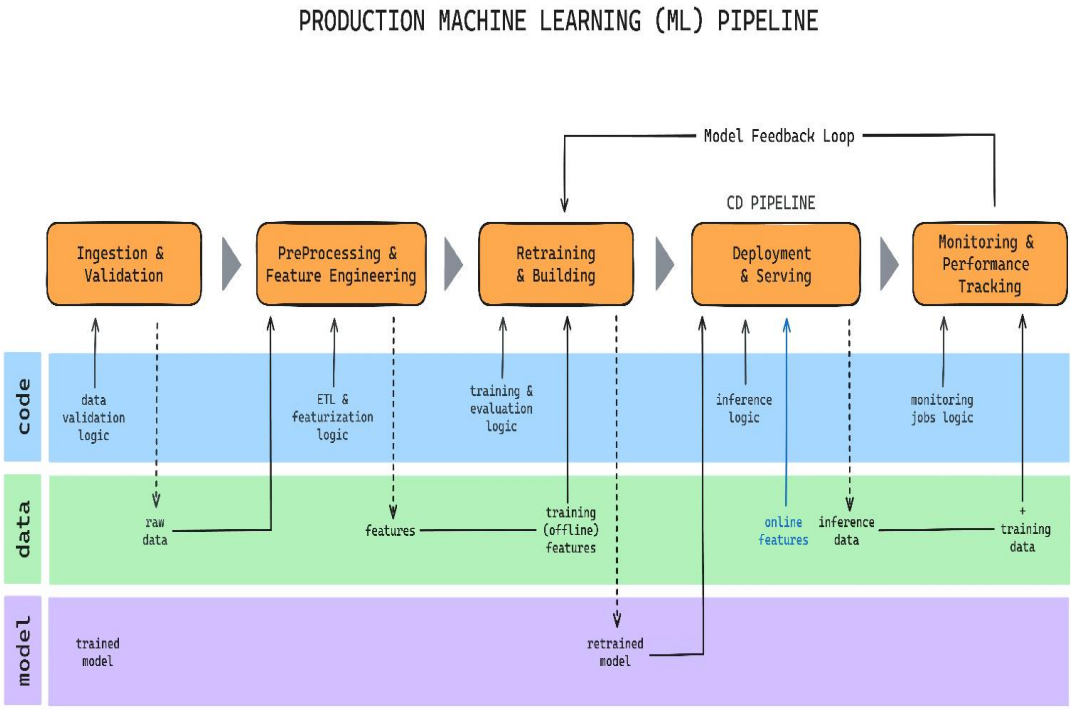
The development of the Autonomous AI Agent for Data Science follows a structured methodology involving data collection, preprocessing, intelligent automation, model training, evaluation, and visualization. Initially, datasets are collected from CSV files, databases, or external sources and loaded into the system using Python-based data processing libraries. The preprocessing module automatically cleans the data by handling missing values, removing duplicates, normalizing data, and converting categorical values into machine-readable formats. Exploratory Data Analysis (EDA) is then performed to analyze patterns, relationships, and statistical summaries using visualization techniques. The intelligent AI agent analyzes the dataset characteristics and automatically selects suitable machine learning algorithms for classification, regression, or clustering tasks. The selected models are trained using machine learning frameworks such as Scikit-learn, and hyperparameters are optimized automatically for better performance. Model evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix are generated to measure performance. Visualization libraries are used to create graphs and charts for better understanding of analytical results. A user-friendly frontend interface is developed using Streamlit to enable interactive user communication with the system. Finally, the entire workflow is tested using different datasets to ensure accuracy, reliability, scalability, and automation efficiency. The methodology ensures the development of a smart, efficient, and fully automated data science assistant system.

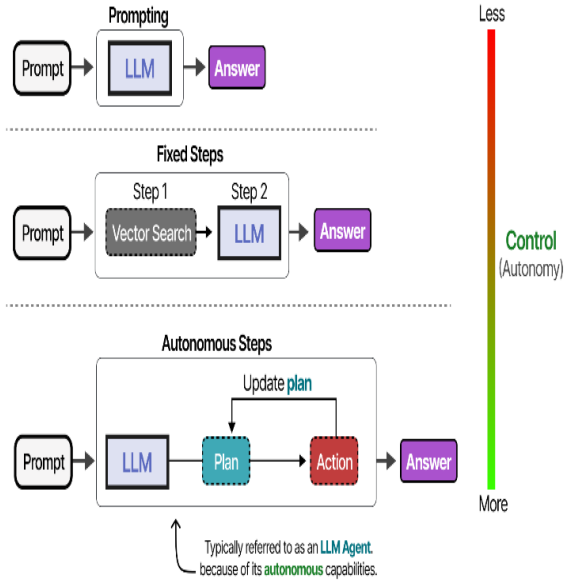
System Architecture

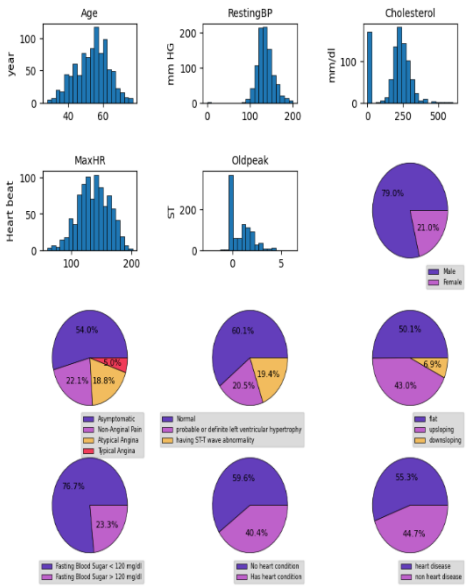


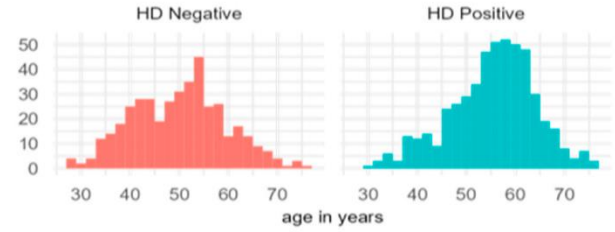
The system architecture of the Autonomous AI Agent for Data Science consists of multiple interconnected modules that work together to automate the complete data science process. The architecture begins with the user interface developed using Streamlit, where users upload datasets and interact with the system. The uploaded data is passed to the preprocessing module, which performs tasks such as data cleaning, handling missing values, normalization, encoding, and transformation automatically. The processed dataset is then forwarded to the Exploratory Data Analysis (EDA) module, which generates statistical summaries, visualizations, and pattern analysis. After analysis, the intelligent AI engine examines the dataset characteristics and selects appropriate machine learning algorithms automatically based on the problem type. The model training module trains the selected machine learning models using Scikit-learn and optimizes performance through automated parameter tuning. A model evaluation module calculates performance metrics such as accuracy, precision, recall, and F1-score to compare different models. The visualization module generates charts, graphs, and analytical reports for better result interpretation. A backend server manages workflow execution, data handling, and communication between different modules efficiently. Databases and storage systems are used to save datasets, trained models, and analytical results securely. Finally, the generated insights, predictions, and visual outputs are displayed to the user through the frontend interface, enabling a seamless and automated data science experience.

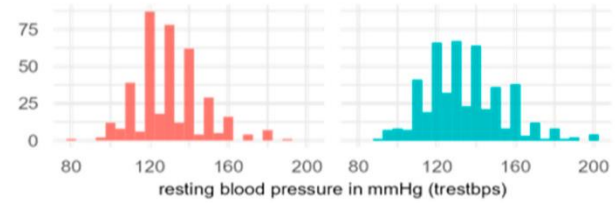
V. Result and Output

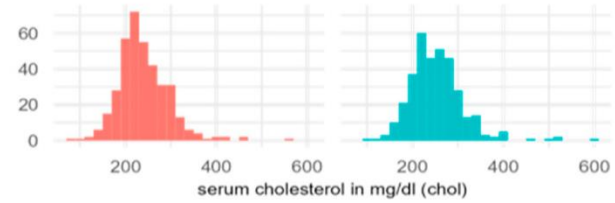

PRODUCTION MACHINE LEARNING (ML) PIPELINE
Model Feedback Loop
CD PIPELINE
Ingestion & Validation
PreProcessing & Feature Engineering
Retraining & Building
Deployment & Serving
Monitoring & Performance Tracking
code
data
model
data validation logic
ETL & featurization logic
training & evaluation logic
inference logic
monitoring jobs logic
raw data
features
training (offline) features
online features
inference data
+ training data
trained model
retrained model

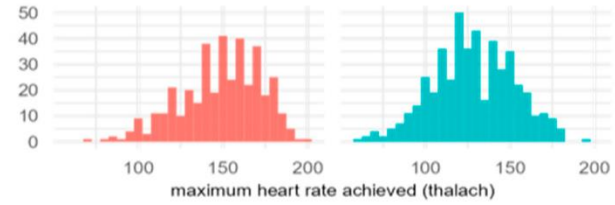

Prompting
Prompt
LLM
Answer
Fixed Steps
Step 1
Step 2
Vector Search
Autonomous Steps
Update plan
Plan
Action
Typically referred to as an LLM Agent.
because of its autonomous capabilities.
Less
Control
(Autonomy)
More

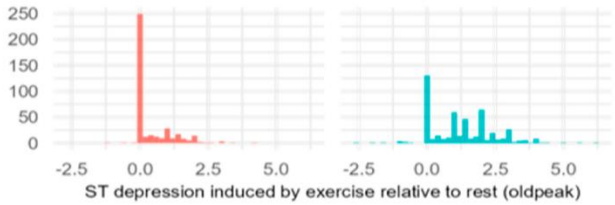

Age
RestingBP
Cholesterol
MaxHR
Oldpeak

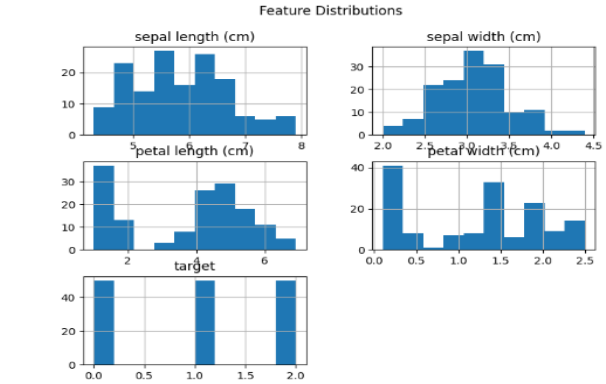

HD Negative
HD Positive
50
40
30
20
10
0
30 40 50 60 70
30 40 50 60 70
age in years

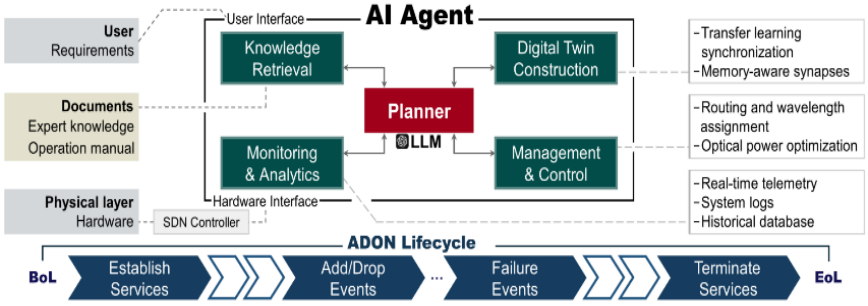

75
50
25
0
80 120 160 200
80 120 160 200
resting blood pressure in mmHg (trestbps)


60
40
20
0
200 400 600
200 400 600
serum cholesterol in mg/dl (chol)


50
40
30
20
10
0
100 150 200
100 150 200
maximum heart rate achieved (thalach)


250
200
150
100
50
0
-2.5 0.0 2.5 5.0
-2.5 0.0 2.5 5.0
ST depression induced by exercise relative to rest (oldpeak)





VI. Conclusion

The Autonomous AI Agent for Data Science project demonstrates how Artificial Intelligence and Machine Learning can be effectively used to automate complex data science workflows. Traditional data science processes often require significant manual effort, technical expertise, and time for tasks such as data preprocessing, exploratory data analysis, model selection, training, and evaluation. The proposed system addresses these challenges by providing an intelligent and automated solution capable of performing these tasks with minimal human intervention.

The developed AI agent successfully integrates various stages of the data science lifecycle into a single streamlined framework. By utilizing technologies such as Pandas, NumPy, Scikit-learn, and Matplotlib, the system can automatically clean datasets, analyze data patterns, select suitable machine learning models, train algorithms, evaluate model performance, and generate visual insights efficiently. The integration of a user-friendly interface using Streamlit further improves accessibility, allowing beginners, students, and non-technical users to interact with the system easily.

The project highlights the importance of intelligent automation in reducing complexity, minimizing human errors, saving time, and improving productivity in data science applications. The autonomous AI agent not only simplifies the machine learning workflow but also enhances scalability and efficiency by automating repetitive tasks. Additionally, the system demonstrates the potential of AI-driven

decision-making in creating adaptive and intelligent analytical solutions for real-world problems.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.