International Journal of Technology Engineering Arts Mathematics Science
Vol. 4, No. 2, December 2024, pp. 01~04
ISSN: 2583-1224,

Automating Machine Learning Pipeline with Large Language model

Dr. D. S. Zingade[1], Mr. Saurabh Shahane[2], Aditya Patil[3], Tejas Sakhare[4], Tushar Nepale[5], Omkar Sarshetti[6]

[1,2,3,4,5, 6]AISSMS Institute of Information Technology, Pune, Maharashtra, India

Corresponding Author: Dr. D. S. Zingade (deeplaxmi.zingade@aissmsioit.org)

Article Information

Article history:

Received Dec, 2024
Revised Jan, 2025
Accepted Jan, 2025

AISSMS IOIT RESEARCH
International Journal of TEAMS

ABSTRACT

This paper presents a multi-agent system to simplify data management and automate machine learning tasks and make advanced analytics more accessible. Using Large Language Models (LLMs), the system combines data analysis, preprocessing and model training, converts raw data into insights without users needing technical knowledge. The method automates the entire machine learning workflow from data review to model validation with dedicated AI agents for each step. Users can upload CSV files and get detailed analysis reports, trained machine learning models and performance metrics. This is not only user friendly but makes data driven analytics more accessible so more people can get involved with data science and AI through a no code platform. This opens up data driven analytics to everyone, getting more people involved in data analytics. The research emphasizes the project's uniqueness and practical applicability in democratizing data-driven analytics through a no-code platform that automates the machine learning process. This system bridges the gap between advanced analytics and non-technical users, making it a valuable tool for industries to leverage data without requiring specialized expertise.

KEYWORDS: Multi agent system, Large Language Models, data processing, machine learning automation, data analytics, no code platform.

1. INTRODUCTION

This fast-paced and rapid adaptation of ML in many domains demands for simplicity while developing effective ML models because the complexity often hinders it in multiple domains.[4] The conventional tasks involved in designing ML processes are very sensitive and highly prone to manual intervention, such as data preprocessing, feature selection, and hyperparameter tuning. These can now be approached with a promising new approach called Automated Machine Learning (AutoML), which also enables the automation of the entire ML pipeline, making it hence accessible to non-experts. [4]

Recent breakthroughs in Large Language Models (LLMs) have further refined the capabilities of AutoML systems. [1]. LLMs, which are very skilled in natural language understanding and generation capabilities, are increasingly being utilized in various stages of the ML pipeline - from data preparation to model deployment. They generate application code, develop scripts for data analytics, and enhance data wrangling processes, thus increasing the productivity of scientific research [1][2]. LLM's integration into the frameworks of AutoML contributes not only to automating complicated coding tasks but also improves the interpretability and usability of ML models, rendering it more accessible to a much broader audience [1].

Such innovative approaches, such as the Data Interpreter- an LLM agent for data science thereby exhibit the scope of LLMs in enriching the process of data analysis through dynamic planning and logical reasoning [2] [3]. It thus enhances the agents problem-solving capabilities with such tools, thereby enhancing code execution and logical inconsistency identification that, in turn, escalates performance in several data science tasks. It has been proven that LLMs have made huge contributions to the values of zero-shot and few-shot prediction tasks, especially when working with structured tabular data. This serves as good evidence for the practical utility of these models in AutoML scenarios [5].

The introduction of LLMs is thereby an important opportunity for the evolution of AutoML into the field that efficiently automates the whole ML pipeline.[4]

This paper will give an extensive overview of how LLMs can be used to automatically embrace the pipeline of the machine learning process in a more robust and user-friendly ML solution.

The novelty of the proposed work lies in the joint application of LLMs and AutoML systems to automate the development process of an ML model, which can also adapt to a continuously evolving dataset.[4][5] This approach, minimizes the human interaction at every step, with maximum output integrity; it also helps in the emerging trend of simplifying machine learning pipelines.

2. LITERATURE SURVEY

LLM, especially after the introduction of ChatGPT, have become powerful tools in scientific and data science fields, as they can write code as well as analyze data. Though impressive, they face the challenge of hallucinations that affects reliability. The trend of using LLM for tasks such as text refinement indicates a shift towards more AI-driven scientific practices.[1]

In data science, LLMs are strong in interpretability, especially for code reasoning and decomposition tasks. Coupled with tools and dynamic planning, problem-solving in multimodal tasks improves, hence not only are automated verification mechanisms crucial in accuracy, but also the verification process ensures accuracy in the insights derived from LLMs.[2]

Two of the most commonly used predictive modeling techniques are XG-Boost and random forests. A newer deep learning technique, NODE, is promising. Transformer-based architectures such as TaBERT and TUTA are well-suited for tabular data processing but are limited in this scenario, as comprehensive quality training datasets are rare.[3]

Automated Machine Learning (AutoML) systems aim at making machine learning more streamlined and automated in certain phases and tasks such as the selection of the algorithm or setting of the hyperparameters. Although they do lessen the work in terms of manual labor, full automation leads to inconsistency, and semi-automated techniques make an instant call.[4]

Recent research posits that, while improving analytics on the vast amounts of information, such LLMs fail to ensure any accuracy with insights derived. In agent-based approaches, a delegation of tasks amongst specialized agents provides for a better efficiency and effect through collaboration and knowledge sharing.[5]

Synergies of LLMs and AutoML are evident as LLMs have been integrated into AutoML systems for the automation of ML program synthesis. Whereas "Text-to-ML" frameworks hold promise at initial stages to fully automate complex tasks, current systems depend on manual input for such tasks. The direction is toward maximizing workflow efficiency, besides democratizing machine learning.[6]

The work of this proposed AUTOML-AGENT framework addresses all the aforementioned challenges as it attempts to fill the gaps in existing AutoML systems by comprehensive support through a pipeline formed by ML. Retrieval augmented planning enhances task automation with high accuracy, and ML will be open for laypersons not having much technical expertise.[7]

The work of LLMs on qualitative data analysis also in fields like software engineering depicts the possibility of using them to perform extremely complex things such as extracting and verifying data. This automatically expands their utility beyond mere numerical analysis, indicating potential in understanding context and meaning.[8]

Most current systems of AutoML, which include Auto-sklearn, leverage automation to optimize a workflow but have put huge demands concerning technical expertise and oversight manual. Progress like Alpine Meadow combined methods to be used toward improved performance but leaves the gaps open in interactivity as well as feedback mechanisms.[9]

The STREAMLINE tool helps in constructing end-to-end AutoML pipelines that are able to perform effective data analysis and algorithm comparison. In that regard, it does provide an open and reproducible method of analyzing datasets through inclusion of a number of ML algorithms that contribute to the improvement of the workflow of data science in its entirety.[10]

LLMs speed up tasks in data wrangling by producing code that facilitates transformations of the data set. This is made easier through methods like programming by example, but LLMs represent a different class of techniques that really demand semantic understanding for some of the things they are required to do, so efficient routing mechanisms have been introduced.[11]

Pre-trained transformers have been found to optimize AutoML pipelines for multimodal data. Such capabilities reflect how important LLMs are in auto-marginalizing and making complex workflows relatively free of intervention by manual means.[12]

The integration of LLMs into the AutoML system can enable user interaction and feedback within the ML pipeline. Even though current frameworks are mostly hands-off, their efforts to bring end-to-end interaction between users and the systems is a good starting point.[13]

LLMs are a very prominent area for marketing for content generation and translation, although phenomena such as hallucinations affect their reliability. Research continues with overcoming the computational challenges within Transformer architectures so that multimodal capabilities can be integrated into different kinds of analytical tasks.[14]

LLMs improve the AutoML process by making better configuration suggestions in the pretraining, fine-tuning, and inference stages. Most issues occur in handling the entire lifecycle of an LLM, but promising techniques for resource allocation and performance optimization include multi-fidelity techniques.[15]

3. METHDOLOGY

This methodology implements a LLM-driven multi-agent system that automates Machine Learning pipeline and automatically executes advanced data analysis and machine learning tasks for a non-technical user.

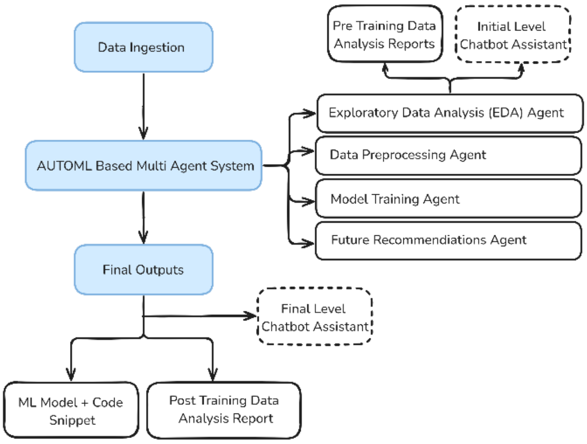
Data Ingestion
Pre Training Data Analysis Reports
Initial Level Chatbot Assistant
Exploratory Data Analysis (EDA) Agent
AUTOML Based Multi Agent System
Data Preprocessing Agent
Model Training Agent
Future Recommendiations Agent
Final Outputs
Final Level Chatbot Assistant
ML Model + Code Snippet
Post Training Data Analysis Report

Fig.1. System Block Diagram

Data Ingestion: The user uploads the dataset in the form of CSV. After that, the data is centralized into a database and then transmitted into AI agents for processing.

Exploratory Data Analysis (EDA): This includes EDA, which is conducted by the AI Agent and executed automatically. Included results are distribution of data, descriptive statistics, missing values, and outliers. A user receives an interactive EDA report or PDF with a chatbot-Llama 3.1 available to answer further questions and get deeper insights.

Data Preprocessing: Following EDA, the AI Agent cleans the data and prepares it for use by removing missing values, scaling the numerical features, and encoding categorical variables, thereby making the dataset excellent for any machine learning tasks. This process helps in the elimination of the multicollinearity problem and the inconsistencies present in the data.

AutoML and Model Training: In the last step of the process flow, we pass the preprocessed data into an AI Agent that makes use of AutoML techniques for selecting the algorithms to be used, choosing hyperparameters, so on and so forth to train the models. We can again measure the performance of models through metrics like accuracy and F1-score.

Model Deployment and Recommendations: The system outputs a trained ML model along with a code snippet for user customization, an extensive Data Analysis Report and recommendations from an AI agent for improvement of the model and its deployment.

User Interaction and Result Interpretation: Upgraded AI chatbot can be engaged interactively by the users post-training for inquiring about model performance besides getting more insights. The up gradation not only contributes to the easier comprehension of complex ML concepts by non-expert users but also helps explore results interactively.

Reporting and Feedback Loop: As Final outputs, EDA and model performance is presented as a PDF report, an interactive EDA Report, the ML model and code snippet, and recommendations going forward. This methodology streamlines the overall process for high-quality outputs and actionable insights with minimum technical expertise.

4. SYSTEM ARCHITECTURE

The system architecture consists of several components that overall simplify an automated data analysis and machine learning pipeline.

End User: The end user interacts with the system by uploading a Data.csv file, which contains the dataset they want to analyze or train models on.

Data Storage (Data.csv): This data (in CSV) is uploaded into a central database, which then serves as the basis for all subsequent analysis and machine learning.

Interactive AI Chatbot (Powered by Llama 3.1): This chatbot interacts with the user and assists in answering data-related queries. It allows the user to explore data statistics, which gives an idea about the dataset before doing machine learning.

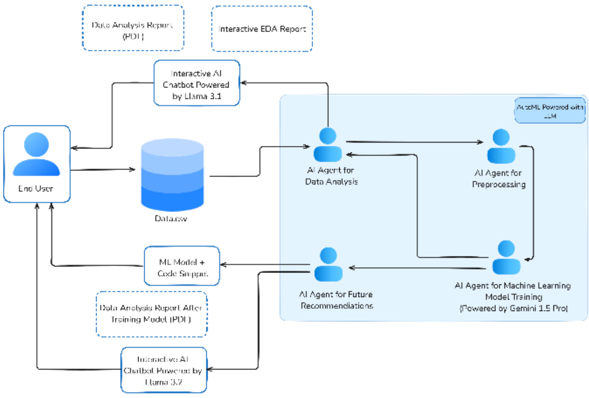
Data Analysis Report (PDF)
Interactive EDA Report
Interactive AI Chatbot Powered by Llama 3.1
AutoML Powered with LLM
End User
Data.csv
AI Agent for Data Analysis
AI Agent for Preprocessing
ML Model + Code Snippet
Data Analysis Report After Training Model (PDF)
AI Agent for Future Recommendiations
AI Agent for Machine Learning Model Training (Powered by Gemini 1.5 Pro)
Interactive AI Chatbot Powered by Llama 3.2

Fig.2. System Architecture

Multi Agent System: It uses AutoML technology driven by LLMs so that the system can do feature selection, model tuning, and evaluation with little user interference. Mainly four AI agents that work together to make an end to end data and ML pipeline.

AI Agent for Data Analysis: Does exploratory data analysis (EDA) and tries to find interesting things about the dataset. Works with the preprocessing agent to make sure the data is all nice and clean and organized.

AI Agent for Preprocessing: Handles data preprocessing tasks such as cleaning, handling missing values, scaling, and feature engineering. It ensures that the data is in suitable format for machine learning. AI Agent for Machine Learning Model Training (Powered by Gemini 1.5 Pro): The PyCaret library is going to be utilized to automate the workflows of the machine learning processes. This would enable training and evaluation of hundreds of models applying many performance metrics with minimal user intervention. and thus obtains the best ML model results in no time. AI Agent for Future Recommendations: This agent gives future recommendations based on data analysis and model prediction, which includes model improvement and deployment strategies after evaluating models. Outputs Generated by the System: The system will generate an output data analysis report in the form of a PDF summarizing initial findings, an interactive EDA report, a trained ML model with a code snippet to implement later on further tailoring it, and a post-training data analysis report in the form of PDF detailing the performance of the model. Interactive AI Chatbot (Powered by Llama 3.2): This chatbot after training the model allows users to ask questions about the results and helps them to have a better understanding and analysis of the outputs. This architecture represents automated, and user-friendly system that takes care of the entire data analysis and machine learning pipeline, through the use of multiple AI agents, large language models, and AutoML techniques.

6. CONCLUSION

The paper aims to implement a multi-agent system in the LLMs that automate the machine learning pipeline, thus opening access towards further advanced data analysis even for non-technical users. It points out how AI agents enhance the overall process of data ingestion, exploratory data analysis, data preprocessing, and training better by using AutoML for optimal algorithm selection and adjustment. The interactive AI chatbot enhances user engagement and simplifies complex machine learning concepts, thereby ensuring high-quality output with actionable insights. As a result, the potential for more research into enhancing the effectiveness and usability of automated machine learning workflows will surely enhance the democratization of machine learning by using automation and AI-powered solutions.This research will help users across multiple domains, allowing them to take full advantage of data-driven decision making and machine learning applications.

REFERENCES

[1] M. Nejjar, L. Zacharias, F. Stiehle, and I. Weber, “LLMs for Science: Usage for Code Generation and Data Analysis,” *arXiv.org*, 2023.

[2] S. Hong *et al.*, “Data Interpreter: An LLM Agent For Data Science,” *arXiv.org*, Mar. 12, 2024.

[3] Y. Yang, Y. Wang, S. Sen, L. Li, and Q. Liu, “Unleashing the Potential of Large Language Models for Predictive Tabular Tasks in Data Science,” *arXiv.org*, Apr. 16, 2024.

[4] T. Nagarajah and G. Poravi, “A Review on Automated Machine Learning (AutoML) Systems,” *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Mar. 2019.

[5] T. Chugh, K. Tyagi, R. Seth, and P. Srinivasan, “Intelligent agents driven data analytics using Large Language Models,” vol. 30, pp. 152–157, Nov. 2023.

[6] J. Xu *et al.*, “Large Language Models Synergize with Automated Machine Learning,” *arXiv.org*, 2024.

[7] P. Trirat, W. Jeong, and S. J. Hwang, “AutoML-Agent: A Multi-Agent LLM Framework for Full-Pipeline AutoML,” *arXiv.org*, 2024.

[8] Z. Rasheed *et al.*, “Can Large Language Models Serve as Data Analysts? A Multi-Agent Assisted Approach for Qualitative Data Analysis,” *arXiv.org*, Feb. 02, 2024.

[9] Z. Shang *et al.*, “Democratizing Data Science through Interactive Curation of ML Pipelines,” *Proceedings of the 2019 International Conference on Management of Data*, Jun. 2019.

[10] R. J. Urbanowicz, R. Zhang, Y. Cui, and P. Suri, “STREAMLINE: A Simple, Transparent, End-To-End Automated Machine Learning Pipeline Facilitating Data Analysis and Algorithm Comparison,” *arXiv.org*, 2022.

[11] X. Li and Till Döhmen, “Towards Efficient Data Wrangling with LLMs using Code Generation,” May 2024.

[12] A. Moharil, J. Vanschoren, P. Singh, and D. Tamburri, “Towards efficient AutoML: a pipeline synthesis approach leveraging pre-trained transformers for multimodal data,” *Machine Learning*, vol. 113, no. 9, pp. 7011–7053, Jul. 2024

[13] R. Barbudo, A. Ramírez, and J. R. Romero, “Evolving machine learning workflows through interactive AutoML,” *arXiv.org*, 2024.

[14] Q. Yang, S. Nikolenko, M. Ongpin, I. Gossoudarev, Y.-Y. Chu-Farseeva, and A. Farseev, “SOMONITOR: Explainable Marketing Data Processing and Analysis with Large Language Models,” *arXiv.org*, 2024.

[15] A. Tornede *et al.*, “AutoML in the Age of Large Language Models: Current Challenges, Future Opportunities and Risks,” *arXiv.org*, Feb. 21, 2024.