IJISEM
INTERNATIONAL JOURNAL OF INNOVATIONS IN SCIENCE ENGINEERING AND MANAGEMENT

OPEN ACCESS

Volume: 4

Issue: 3

Month: September

Year: 2025

ISSN: 2583-7117

Published: 2.09.2025

Citation:

Mudita Sharma "The Role of Machine Learning in Enhancing Data Science Workflows: A Systematic Review" International Journal of Innovations in Science Engineering and Management, vol. 4, no. 3, 2025, pp. 301–306.

DOI:

10.69968/ijisem.2025v4i1392-397

The Role of Machine Learning in Enhancing Data Science Workflows: A Systematic Review

Mudita Sharma[1]

[1]*Research Intern, CSIR-CEERI, Jaipur, Rajasthan*

Abstract

Big data is almost always the foundation of machine learning (ML) models, which have attracted a lot of interest in a range of applications, from computer vision to natural language processing. The intersection of data science, artificial intelligence, and software engineering is shown in the increasing number of products and applications that have integrated machine learning models. This review highlights that machine learning plays a pivotal role in enhancing data science workflows by automating complex tasks, improving predictive accuracy, and enabling data-driven decision-making. Proactive and reactive algorithms, supported by advanced computational power through GPUs and TPUs, allow better forecasting and response in real-world applications. Incorporating data engineering techniques with AI ensures scalability, efficiency, and reduced human errors. Furthermore, machine learning enables the discovery of hidden patterns, handling of massive datasets, and integration of smart computing for decision-making across domains such as business, healthcare, cybersecurity, and urban systems, thereby significantly advancing data science practices.

Keywords; Machine Learning (ML), Data Science Workflows, Natural Language Processing, Artificial Intelligence, Internet of Things (Iot), Data Analytics, Data Mining.

INTRODUCTION

Nearly every element of our daily lives is digitally recorded as data in this age of advanced analytics and data science. As a result, the modern electronic world includes a wealth of data, including social media, commercial, financial, healthcare, multimedia, and internet of things (IoT) information [1]. There is a daily increase in the quantity of structured, semi-structured, and unstructured data. To understand and evaluate real-world events using data, data science is often defined as the combination of statistical methods, data analysis, and related approaches [2]. In addition to data science, the graphic shows the growth in popularity of similar topics including big data, machine learning, data analytics, and data mining. Data science is the use of advanced analytics methods and scientific concepts to get useful business insights from data [3]. Advanced analytics places a greater emphasis on the prediction of future events through the utilisation of data to identify patterns [4]. Advanced analytics contributes to the area by offering a deeper understanding of data and helping with the analysis of granular data, which is of interest to us, while basic analytics gives a broad description of data. There is a high prevalence of descriptive, diagnostic, predictive, and prescriptive analytics in the field of data science [5]. Whereas diagnostic analytics answers the issue of why something happened, descriptive analytics answers the question of what happened. Conversely, prescriptive analytics suggests the best path of action [6]. These ideas are covered in short in Smart Computing and Advanced Analytics Methods. "The Fourth Industrial Revolution (Industry 4.0)" may benefit greatly from sophisticated analytics and machine learning-based decision-making, which are key components of artificial intelligence (AI), because of its learning potential for intelligent computing and automation [7].

Machine Learning

"Machine learning, or ML for short", is a subfield of artificial intelligence (AI) that focusses on developing computer algorithms that improve on their own given experience and data.

To put it simply, machine learning enables computers to learn from data and determine what to do without explicit programming. Machine learning is, at its core, the process of creating and implementing algorithms that facilitate these assessments and forecasts [8]. These algorithms are designed to improve with time, becoming more accurate and efficient as more data is handled. In traditional programming, a computer follows a set of preset instructions to accomplish a task. However, machine learning lets the computer decide how to do a task by utilising a set of samples (data) and a job to perform.

Data science workflow

A systematic framework with several steps that helps data scientists finish a data science project effectively is called a data science workflow. The procedures that must be taken to guarantee a methodical approach to problem-solving are described, including data intake, preparation, integration, analysis, visualisation, and distribution [9]. The following graphic provides a summary of a common process associated with many data science initiatives.

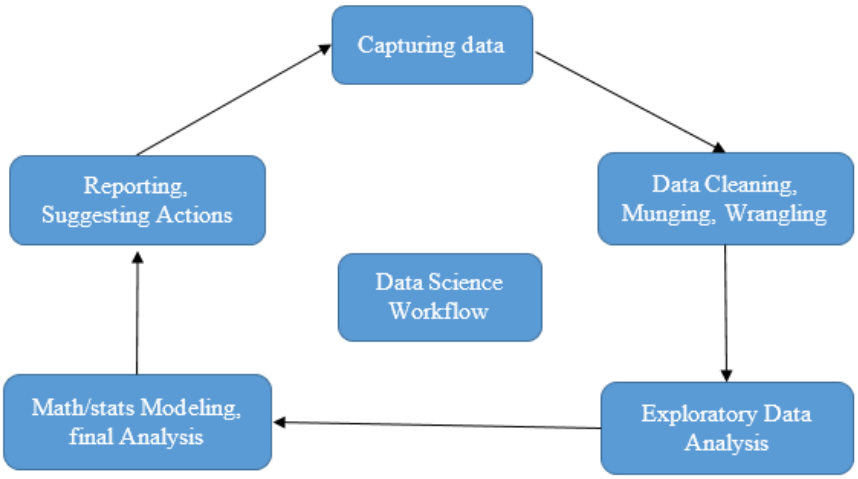

Figure 1 Workflow in data science projects

1. Capturing Data and Data Cleaning, Munging, Wrangling

On its own, gathering a lot of data is difficult. It might include requesting information from databases, online repositories, web servers, APIs, etc. Data are almost never clean and are often noisy. Inaccurately recorded or stored data, or missing data, are common. Furthermore, not all data may be utilised immediately, and not all data are equally valuable. It could be necessary to alter and convert them into datasets that are more beneficial. Before cleaned and altered data may be utilised as an input in statistical and mathematical models, these issues must all be resolved. At least in its broadest sense, data science now includes addressing these issues.

2. Exploratory Data Analysis (EDA)

Once the data has been cleaned and modified, we want to utilise it in statistical models to forecast the company or domain from where it originated. However, it is crucial to first have an understanding of the data and evaluate what it tells us in order to decide which statistical models are suitable to use or which business hypotheses are plausible to assume and verify. We do exploratory data analysis for that reason. Plotting different graphs, histograms, bar charts, pie charts, etc. using multiple characteristics, variables, or columns is a common way that we visualise data. In addition to finding indications of intriguing occurrences that the data is revealing, this aids in our first impression and intuition.

3. Mathematical and Statistical Modelling; Final Analysis

The process of visualising data and gaining an intuitive understanding of its contents frequently leads to the formulation of queries that require answers and the formulation of hypotheses regarding the data that require verification. Additionally, we want to forecast future data. All of this is accomplished via statistical and mathematical modelling. In other words, using the cleansed data as inputs and using mathematical and statistical tools and techniques. This often calls for extensive coding and programming.

4. Reporting; Suggesting Actions

The report and/or presentation should be generated after the implementation of statistical methods and the generation of results and predictions. Conclusions are made, and

sometimes some recommendations for further action are included. This directs the associated business's decision-making process. Graphs, animations, and other visual aids are often used in reports. A dynamic report can be generated by a code, allowing the report's output to be automatically altered and modified when the entire code is executed in the future with new data. Interactive applications may also be included in reports, allowing readers and decision-makers to experiment with the apps and get a deeper understanding of the findings and outcomes.

The Applications of Machine Learning in Data Science

In data science, the following are a few of the most popular machine learning applications:

- **Real-Time Navigation:** Google Maps is one of the most popular real-time navigation tools. In spite of the fact that you are in the typical traffic, have you ever considered why you are on the quickest route? The Historical Traffic Data database and the data collected from users of our service today are the reasons behind this. Every user of this service helps to improve the accuracy of this application. Data is transmitted to Google on a continuous basis upon application launch, which furnishes traffic patterns and route information at any given hour.
- **Image Recognition:** Image recognition is one of the most widely used applications of machine learning in data science. Identification of objects, people, locations, etc. is accomplished by image recognition.
- **Product Recommendation:** Companies in the eCommerce and entertainment sectors, such as Amazon, Netflix, Hotstar, and others, heavily rely on product recommendations. They use a range of "Machine Learning algorithms" to recommend products and services that you may find interesting based on the data they have collected from you.
- **Speech Recognition:** The process of turning spoken words into text is called speech recognition. Words, syllables, sub-word elements, or even characters may be used to describe this text. Youtube Closed Captioning, Siri, and Google Assistant are among the most well-known examples.

The Challenges of Machine Learning in Data Science

In the realm of data science, machine learning has transformed the way industries operate. It has aided businesses in making wise choices that have led to company expansion. However, there are still a few issues that a data scientist has to take into account [10]. The Top 3 Machine Learning Challenges in Data Science are listed below:

- **Lack of Training Data:** Data is at the core of all machine learning models. Nevertheless, there is a significant cost and difficulty associated with obtaining labelled data. Without a lot of data, every data scientist is concerned about how to train a machine learning model.
- **Discrepancies between Data:** The second difficulty is that the training and production data often differ in some ways. Occasionally, the model performs well in the prototype setting but is unable to generalise in practical situations. For instance, the model may function well in one nation but not in another because of geographic differences; it might function well in winter but not in summer because of seasonal variations; it might function well on mobile devices but not on desktops because of user differences; etc.
- **Model Scalability:** One of the biggest problems facing industry is this. It is essential for a data scientist to ensure that their model is both quick and lightweight. Quantisation after training is one way to solve this issue. This conversion technique reduces the size of the model while simultaneously enhancing the latency of the CPU and hardware accelerators, albeit with a slight decrease in model accuracy.

The Role of Machine Learning in Data Science

Data science relies heavily on the quickly expanding subject of machine learning. Computers can learn from data in this branch of artificial intelligence and use that knowledge to predict or take action. Data may be mined for insights and predictions using machine learning, which can then be used to inform choices or actions.

Reinforcement learning, unsupervised learning, and supervised learning are the three categories of machine learning. Supervised learning uses labelled data, such data that has been tagged or classified, to train a model. Using fresh information to generate predictions or take action is the aim of supervised learning.

Creating a model using unlabelled data such as unclassified or unlabelled data is known as unsupervised learning. Finding structures or patterns in the data is the aim of unsupervised learning. Reinforcement learning is the process of teaching a model to function in a particular context, such a game or a robot, by using feedback in the form of rewards or penalties [11].

To assess consumer data and forecast customer behaviour, including product suggestions and customer attrition, businesses utilise machine learning. Machine learning is used in healthcare to evaluate medical data and forecast

patient outcomes, including the course of a disease or the efficacy of a therapy. Machine learning is used in the financial industry to assess market data, forecast stock values, and identify fraudulent transactions [12].

Beyond these instances, machine learning plays a larger role in data science. Machine learning is used, for example, in self-driving vehicles to identify objects and forecast their motion, resulting in safer and more effective driving. Machine learning is employed in the manufacturing sector to optimise production processes and anticipate equipment malfunctions.

LITERATURE REVIEW

(Andersen & Reading, 2024) [13] The speed at which these data are being generated has exacerbated bottlenecks in study design and data analysis approaches because conventional methods that depend on traditional statistical tests and assumptions are inadequate or inappropriate for highly dimensional data (i.e., more than 1,000 variables). Analysing large amounts of data using machine learning techniques is an interesting and more common choice. A significant obstacle, however, is the lack of experience necessary to evaluate machine learning model findings in a way that provides valuable biological insight. To address this challenge, "a general overview of data analysis and machine learning approaches and considerations" is provided, along with an easy-to-use machine learning workflow that can be applied to a variety of data types. This workflow reduces the large amounts of data to those variables (attributes) that are most likely to determine experimental and/or observed conditions. The procedure outlined here has been successfully beta-tested as a standardised way to reduce the dimensionality of data, and it is recommended that big data analysis pipelines include it. Additionally, the methodology is adaptable, and the fundamental concepts and procedures can be adjusted to accommodate the study parameters, objectives, and user requirements.

(Gouthami et al., 2024) [14] Intends to illustrate the importance of machine learning as a key element in supporting data discovery and efficient administration. The meticulous analysis of extensive data sets is a notable feature that enables the generation of valuable forecasts that can inform "improved decision-making and prompt intelligent actions in real-time, without the need for human involvement". This research is noteworthy for its comprehensive examination of numerous proposed frameworks in the field of data science, as well as its emphasis on the significant influence of machine learning methodologies, including pipeline development, algorithm development, and model evaluation and selection. I also point out potential misconceptions that can result from ignoring machine learning's reasoning component.

(Kilaru, 2024) [15] The goal of this project is to increase the speed, size, and repeatability of data-driven jobs by using data engineering techniques in data science. The study investigates the use of WMS to integrate ADP and AFT across the data science process, from data collection to final model deployment. This paper demonstrates the efficacy of automation in addressing problems such as integration, processing time, and dependence on human efforts for improving organisational and decision-making processes via the analysis of simulation reports and real-life examples. According to the major points, using data engineering techniques enhances the quality and dependability of analytics findings and outputs, saves time and resources when doing data pre-processing and analysis, and is a crucial part of modern analytical pipelines.

(Ogrizović et al., 2024) [16] There are several challenges when a system must operate in a real-world environment, such as how to handle inaccurate predictions the model may make, how to maintain safety and security despite possible errors, which features are more important than a model's prediction accuracy, and how to identify and measure crucial quality requirements like "learning and inference latency, scalability, explainability, fairness, privacy, robustness, and safety". In an effort to ascertain their potential capabilities and deficiencies, it is imperative to conduct comprehensive testing of these models. In this study, all answers to the aforementioned problems are given methodologically uniformly, along with a taxonomy and conclusions on potential future development tendencies. The main contributions of this paper are a classification that closely reflects the structure of the ML-pipeline, a thorough explanation of the role of each team member in that pipeline, and a summary of the challenges and trends in integrating "ML and big data analytics, with applications in both industry and education".

(Hassani & Silva, 2023) [17] Describes how data scientists may use ChatGPT to automate several processes in their workflow, including cleaning and preparing data, evaluating findings, and learning models. Based on the overall results of the research, ChatGPT has the ability to greatly increase the precision and effectiveness of data science procedures and is anticipated to gain prominence as a data science intelligence augmentation tool. ChatGPT can assist with a range of natural language processing tasks related to data

science, including sentiment analysis, language translation, and text classification. ChatGPT may not function effectively on other tasks if it has not been trained for them, despite the fact that it can be optimised for specific use cases and save time and resources compared to training a model from the ground up. Furthermore, ChatGPT's output could be hard to understand, which might make data science applications' decision-making more challenging.

(Sarker, 2021) [18] Give a comprehensive account of data science, which encompasses a variety of advanced analytics techniques that can be employed to enhance the intelligence and capabilities of an application by making informed decisions in a variety of scenarios. In light of this, we conclude by outlining the difficulties and possible lines of inquiry within the parameters of our investigation. The primary contribution of this research, which takes into account "data-driven smart computing and decision making", has been identified as the elucidation of diverse advanced analytical techniques and their suitability for a range of real-world data-driven application domains, including cybersecurity, healthcare, business, rural and urban data science, and such.

RESEARCH OBJECTIVE

- To study the machine learning, and its application and challenges.
- To study the role of machine leaning in data science workflow.
- To study the various literature's perspective on role of machine learning in data science workflow.

RESEARCH GAP

Although machine learning has significantly advanced data science workflows by automating tasks, improving predictions, and enabling large-scale data analysis, several gaps remain unaddressed. Current studies largely emphasize technical advancements but lack focus on standardizing integration frameworks that combine machine learning with data engineering practices for scalable, automated workflows. Moreover, limited attention has been given to challenges such as interpretability, ethical concerns, and adaptability of models in dynamic, real-world environments. Research on domain-specific optimization, energy-efficient computation, and handling unstructured or heterogeneous data is also scarce, highlighting the need for deeper exploration to fully harness machine learning's potential.

RESEARCH METHODOLOGY

This review paper employs a qualitative research methodology grounded in secondary data analysis to explore the role of machine learning in enhancing data science workflows. An extensive literature review was conducted, systematically analyzing peer-reviewed journals, scholarly articles, conference proceedings, government reports, and technical papers published between 2019 and 2024. Relevant case studies were also examined to capture practical insights and applications. Sources were selected based on their relevance, credibility, and contribution to understanding machine learning's integration into data workflows. The collected data was thematically analyzed to identify key trends, advancements, challenges, and future directions in the field.

CONCLUSION

This review highlights that the integration of machine learning within data science workflows has fundamentally transformed how organizations analyze, interpret, and utilize data. Proactive and reactive algorithms have been shown to improve forecasting and response mechanisms by effectively leveraging both processed and unprocessed data. Techniques such as backpropagation through time enable the modeling of sequential dependencies, allowing more accurate predictions from time-series and dynamic datasets. The advent of high-performance computing resources, notably GPUs and TPUs, has further expanded the scope of deep learning applications by facilitating large-scale, complex computations with increased efficiency.

The findings also underscore the growing role of automation in data science through data engineering practices and workflow management systems (WMSs). By integrating these systems with artificial intelligence, organizations can reduce human error, accelerate analytical processes, and enhance scalability in handling massive and ever-changing datasets. This establishes a strong foundation for adaptive, data-driven decision-making in diverse domains.

Moreover, the study emphasizes the versatility of machine learning techniques in real-world applications, ranging from business intelligence and healthcare to cybersecurity, smart urban systems, and rural development. The capacity of machine learning to automate complex analytical tasks, enhance predictive accuracy, and uncover concealed patterns has secured its status as an essential component of data-driven smart computing.

In conclusion, the review affirms that machine learning significantly enhances data science workflows by making them more efficient, scalable, and insightful. Machine learning will play an increasingly important role in data science as AI-driven automation and computing resources develop, resulting in more creative and efficient data-driven solutions.

REFERENCES

[1] S. Umrao, S. Dron, and R. Saxena, "An Examination of the Impact of Artificial Intelligence on Human Resource Management: Improving Efficiency and Employee Experience," Lect. Notes Networks Syst., vol. 928 LNNS, pp. 406–424, 2024, doi: 10.1007/978-3-031-54671-6_30.

[2] A. Nouri, P. E. Davis, P. Subedi, and M. Parashar, "Exploring the Role of Machine Learning in Scientific Workflows: Opportunities and Challenges." 2021.

[3] E. Kesavan, "Internet of Things (IoT): A Review of Security Challenges and Solutions," Int. J. Innov. Sci. Eng. Manag., vol. 2, no. 4, 2023, doi: 10.69968/ijisem.2023v2i465-71.

[4] E. Deelman, A. Mandal, M. Jiang, and R. Sakellariou, "The role of machine learning in scientific workflows," Int. J. High Perform. Comput. Appl., vol. 33, no. 6, 2019, doi: 10.1177/1094342019852127.

[5] M. J. Bdair, "Enhancing Machine Learning Workflows: A Comprehensive Study of Machine Learning Pipelines," Res. gate, 2024.

[6] A. Singh and N. Shanker, "Redefining Cybercrimes in light of Artificial Intelligence : Emerging threats and Challenges," pp. 192–201, 2024, doi: 10.69968/ijisem.2024v3si2192-201.

[7] J. Kumari, E. Kumar, and D. Kumar, "A Structured Analysis to study the Role of Machine Learning and Deep Learning in The Healthcare Sector with Big Data Analytics," Arch. Comput. Methods Eng., vol. 30, no. 6, 2023, doi: 10.1007/s11831-023-09915-y.

[8] R. Pugliese, S. Regondi, and R. Marini, "Machine learning-based approach: Global trends, research directions, and regulatory standpoints," Data Sci. Manag., vol. 4, 2021, doi: 10.1016/j.dsm.2021.12.002.

[9] K. E. Schackart, H. J. Imker, and C. E. Cook, "Detailed Implementation of a Reproducible Machine Learning-Enabled Workflow," Data Sci. J., vol. 23, no. 1, pp. 1–14, 2024, doi: 10.5334/dsj-2024-023.

[10] F. Le Piane, M. Vozza, M. Baldoni, and F. Mercuri, "Integrating high-performance computing, machine learning, data management workflows, and infrastructures for multiscale simulations and nanomaterials technologies," Beilstein J. Nanotechnol., vol. 15, 2024, doi: 10.3762/BJNANO.15.119.

[11] S. Raschka, J. Patterson, and C. Nolet, "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence," Inf., vol. 11, 2020, doi: 10.3390/info11040193.

[12] Z. N. Jawad and V. Balázs, "Machine learning-driven optimization of enterprise resource planning (ERP) systems: a comprehensive review," Beni-Suef Univ. J. Basic Appl. Sci., vol. 13, no. 4, 2024, doi: 10.1186/s43088-023-00460-y.

[13] L. K. Andersen and B. J. Reading, "A supervised machine learning workflow for the reduction of highly dimensional biological data," Artif. Intell. Life Sci., vol. 5, 2024, doi: 10.1016/j.ailsci.2023.100090.

[14] G. Gouthami, N. Siddartha, and D. P. Borugadda, "Data Science: the Impact of Machine Learning," Futur. Trends Artif. Intell. Vol. 3 B. 8, vol. 3, 2024, doi: 10.58532/v3bgai8p2ch7.

[15] N. B. Kilaru, "AUTOMATE DATA SCIENCE WORKFLOWS USING DATA ENGINEERING TECHNIQUES," Int. J. Res. Publ. Semin., vol. 2, no. 2, 2024.

[16] M. Ogrizović, D. Drašković, and D. Bojić, "Quality assurance strategies for machine learning applications in big data analytics: an overview," J. Big Data, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-01028-y.

[17] H. Hassani and E. S. Silva, "The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field," Big Data Cogn. Comput., vol. 7, no. 2, 2023, doi: 10.3390/bdcc7020062.

[18] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," SN Comput. Sci., 2021, doi: 10.1007/s42979-021-00765-8.