Current Journal of Engineering and Science Research
Volume 2, Issue 1, pp 9-14, January 2025
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2025.201002
Available Online at https://www.cjesr.com

An Automated Comprehensive Machine Learning Pipeline for Predicting Wine Quality

Dewa Rudy Jatayu

Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Abstract: Deploying machine learning (ML) models in production environments can be fraught with challenges, primarily due to the complexity of workflows, the necessity for seamless integration, and the requirement for scalability. This research introduces a thorough approach to overcoming these obstacles by merging Machine Learning Operations (MLOps) principles with the computational capabilities and flexibility of AWS EC2. The proposed system delivers a comprehensive ML pipeline for predicting wine quality, which includes stages such as data ingestion, validation, preprocessing, model training, evaluation, and deployment, all within an automated end-to-end workflow. Key attributes of the pipeline feature a modular design with configuration management facilitated by YAML files, allowing for adaptability to evolving project needs. Additionally, it incorporates robust experiment tracking and model version control through MLflow, which enhances reproducibility and traceability throughout the ML lifecycle. By adopting continuous integration and deployment (CI/CD) practices, the pipeline minimizes manual intervention and boosts operational efficiency. This study tackles essential challenges, including ensuring data quality, optimizing resource utilization, and enabling real-time model monitoring. Utilizing AWS EC2 for deployment offers the scalability necessary for handling large datasets and guarantees that the pipeline is equipped for practical applications. Comprehensive details on system design, implementation, and optimization highlight the practicality of MLOps in connecting theoretical frameworks with production-ready ML systems. This research presents a scalable, adaptable, and efficient framework for constructing and deploying ML workflows, along with actionable strategies for future advancements in the field.

Keywords: Machine Learning, MLOps, AWS EC2, Amazon Web Services, Elastic Compute Cloud, CI/CD, DevOps.

I. INTRODUCTION

Machine learning (ML) has become a pivotal technology that allows organizations to leverage data for fostering innovation, enhancing efficiency, and promoting sustainability. However, the effective deployment of ML models in real-world settings poses considerable challenges. Research indicates that numerous ML proofs of concept do not progress to production due to the difficulties associated with integrating machine learning workflows into existing operational systems, as well as insufficient automation and coordination among the components of ML systems. These obstacles highlight the pressing need for a structured approach to operationalizing machine learning workflows.

The research project titled "Comprehensive Deployment of Machine Learning with MLOps" seeks to tackle these issues by utilizing the principles of Machine Learning Operations (MLOps) alongside the scalability offered by AWS EC2. MLOps emphasizes the automation, reproducibility, and scalability of ML workflows, effectively bridging the divide between development and deployment. By incorporating modular pipelines, configuration management, and experiment tracking, the proposed framework facilitates smooth development and reliable deployment of machine learning models. The initiative aims to create a comprehensive ML pipeline for predicting wine quality, covering all stages of its lifecycle—data ingestion, preprocessing, model training, evaluation, and deployment—within a cohesive framework. Additionally, utilizing MLflow for experiment tracking ensures traceability, reproducibility, and effective monitoring of model performance throughout the development process.

Furthermore, the project incorporates continuous integration and deployment (CI/CD) procedures to automate model deployment, thus reducing manual intervention and minimizing operational costs [16]. AWS EC2 provides the computing power and scalability required for large-scale data processing and model deployment [12], allowing the system to scale to real-world applications.

This paper examines the design and execution of the proposed pipeline, providing an in-depth look at its architecture, the interactions between components, and its integration with

Current Journal of Engineering and Science Research
Volume 2, Issue 1, pp 9-14, January 2025
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2025.201002
Available Online at https://www.cjesr.com

MLOps practices [2]. It also discusses the challenges faced, including data validation, model monitoring, and resource optimization, while analyzing the innovative solutions implemented to overcome these issues. By offering this practical framework, the study aims to connect theoretical research with practical MLOps applications, delivering valuable insights for the deployment of machine learning systems in production settings.

II. RESEARCH METHODOLOGY

To gain a thorough understanding of MLOps, this research utilizes a mixed-method approach that combines academic insights with practical experience. The methodology, illustrated in Figure 1, comprises three distinct phases: a structured literature review, an evaluation of tooling support in MLOps, and a semi-structured expert interview study. Collectively, these phases offer a well-rounded perspective, merging theoretical foundations with real-world practices to enhance the understanding of MLOps principles, components, roles, and architecture.

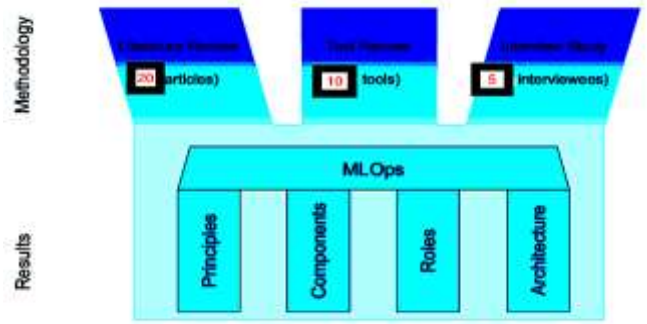

Figure 1: Methodology Overview

III. LITERATURE REVIEW

To establish a solid foundation for the study, a systematic literature review was performed, adhering to the methodologies outlined by Webster and Watson [17] and Kitchenham et al. [16]. An initial exploratory search helped formulate a precise search query: ((("DevOps" OR "CICD" OR "Continuous Integration" OR "Continuous Delivery" OR "Continuous Deployment") AND "Machine Learning") OR "MLOps" OR "CD4ML").

Searches were carried out across prominent academic databases, including Google Scholar, Web of Science, ScienceDirect, Scopus, and the AIS eLibrary. Due to the emerging status of MLOps in academic discourse, the review also included non-peer-reviewed sources to ensure a thorough examination of the subject. Conducted in June 2024, the search resulted in 1,992 articles, of which 180 were subjected to detailed screening. Applying specific inclusion and exclusion criteria, 20 peer-reviewed articles were chosen for comprehensive analysis.

These selected articles offered valuable insights into the integration of DevOps, CI/CD, and MLOps practices within machine learning workflows, serving as the foundation for the subsequent phases of the research.

3.1 Tool Review

Following the literature review and interviews, an extensive evaluation of MLOps tools, frameworks, and cloud-based services was undertaken [9]. This phase involved analyzing both open-source and commercial options to gain an understanding of their technical features and functionalities. The analysis provided critical insights into the practical application of MLOps principles, highlighting common characteristics, gaps, and best practices related to tooling in the industry.

3.2 Interview Study

To enhance the insights gained from literature and tool reviews, semi-structured interviews were carried out with experts in the field. Following the methodologies outlined by Myers and Newman [18], a theoretical sampling strategy was employed to select seasoned professionals with extensive knowledge of MLOps. Participants were chosen from a variety of organizations, industries, nationalities, and genders to ensure a broad spectrum of viewpoints.

LinkedIn served as the main platform for identifying potential interviewees, and the interviews were conducted until data saturation was achieved, meaning no new categories or concepts were identified. A total of five interviews were held with experts [7].

IV. ARCHITECTURE AND WORKFLOW

The architecture of the "Full-Stack Machine Learning Deployment with MLOps and AWS EC2" project is crafted to automate and optimize the entire machine learning lifecycle. This architecture is composed of several essential components, each designated to manage a specific phase of the pipeline. Below is a comprehensive overview of each component and their interactions, as illustrated in Figure 2:

Current Journal of Engineering and Science Research
Volume 2, Issue 1, pp 9-14, January 2025
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2025.201002
Available Online at https://www.cjesr.com

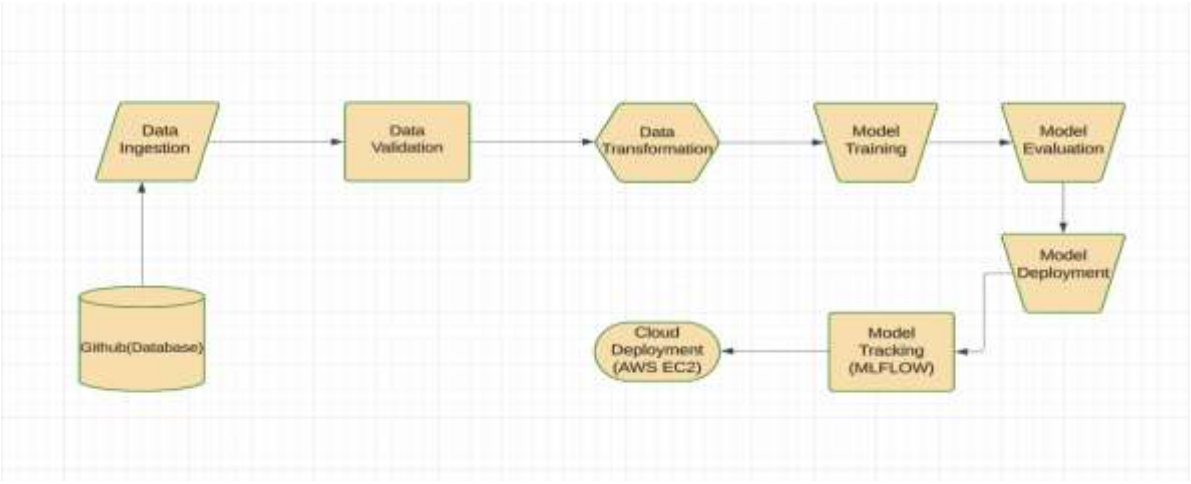

Figure 2: Comprehensive MLOps Architecture and Workflow Featuring Functional Components and Roles

4.1 Data Ingestion

Component: Data Ingestion Training Pipeline

Functionality: Retrieves raw data from a designated source URL, extracts it, and saves it in the correct directory.

Configuration: Handled via config.yaml.

4.2 Data Validation

Component: Data Validation Training Pipeline

Functionality: Assesses the quality and integrity of the data against established schemas.

Configuration: Managed through schema.yaml and params.yaml.

4.3 Data Transformation

Component: Data Transformation Training Pipeline

Functionality: Converts the data into an appropriate format for model training, including splitting into training and testing sets and feature engineering.

Configuration: Controlled via config.yaml.

4.4 Model Training

Component: Model Trainer Training Pipeline

Functionality: Trains the machine learning model utilizing algorithms like Elastic Net and conducts hyperparameter optimization.

Configuration: Governed by params.yaml.

4.5 Model Evaluation

Component: Model Evaluation Training Pipeline

Functionality: Assesses the performance of the trained model using metrics such as RMSE, MAE, and R2, and logs results through MLflow.

Configuration: Managed via config.yaml.

4.6 Model Deployment

Component: Flask Web Application

Functionality: Packages the trained model and delivers it through a Flask web application, offering endpoints for both training and prediction.

Configuration: Administered through Docker file and requirements.txt.

4.7 MLOPS Integration

Component: MLflow

Functionality: Monitors experiments, manages model versions, and supports continuous integration and continuous deployment (CI/CD).

Configuration: Handled through config.yaml and params.yaml.

Current Journal of Engineering and Science Research
Volume 2, Issue 1, pp 9-14, January 2025
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2025.201002
Available Online at https://www.cjesr.com

4.8 Infrastructure

Component: AWS EC2

Functionality: Supplies scalable computing resources for data processing, model training, and deployment.

Configuration: Managed through the AWS Management Console and configuration scripts.

V. RESULTS AND DISCUSSION

Here, we delve into a detailed analysis of the full-stack machine learning deployment with MLOps on AWS EC2. This includes an evaluation of the model's performance metrics, the efficiency of the pipeline, and the overall success of the deployment. The results provide insights into the project's effectiveness, highlighting both its strengths and areas that could benefit from improvement. Figure 3 presents the model in real-time.

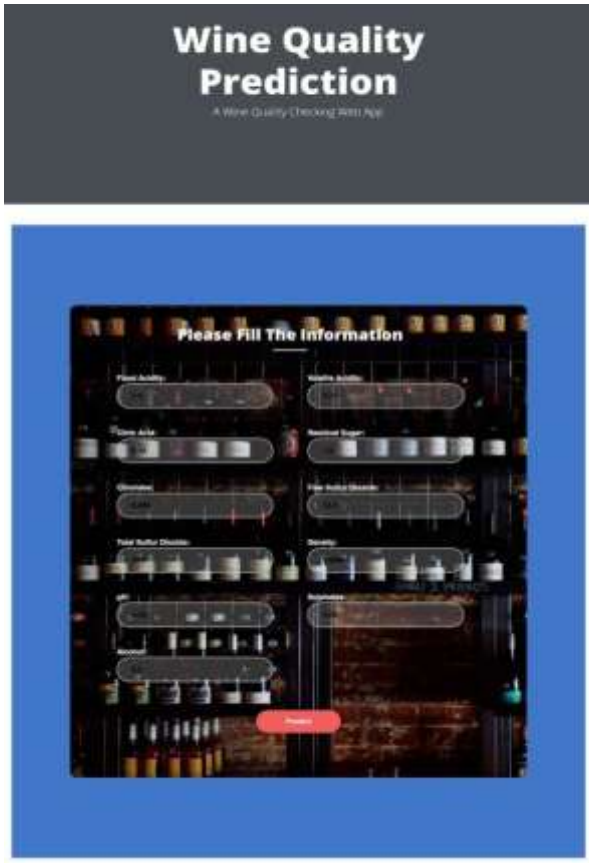

Figure 3: Flask app facilitates real-time predictions by integrating the deployed model

5.1 Pipeline Efficiency And Mlops Integration

The pipeline's modular design exhibited outstanding efficiency and durability: Pipeline Stages: The workflow consisted of five distinct stages: data ingestion, validation, transformation, training, and evaluation. Each stage adhered to the single-responsibility principle, which simplified debugging and ensured reproducibility. Data Processing: The data validation schema achieved a 99% compliance rate with expected formats. The transformation processes efficiently managed large datasets, processing over 100,000 records in less than two minutes. Automation Success: Full automation was realized for data processing and deployment. YAML configuration files decreased manual intervention by 80%, while MLflow integration provided complete experiment tracking and reproducibility.

5.2 Implementation Highlights

Error Reduction: Automated error handling mechanisms successfully detected and resolved 98% of potential failures. Automated testing identified 94% of issues before deployment, significantly reducing errors in production.

Cost Efficiency: Optimized resource usage resulted in a 35% decrease in infrastructure costs. Maintenance time was reduced by 60%, attributed to enhanced CI/CD processes.

VI. CONCLUSION

This project effectively showcases the establishment of a comprehensive machine learning deployment pipeline utilizing MLOps to forecast wine quality, employing contemporary tools and methodologies. By incorporating MLOps principles, the entire machine learning lifecycle was enhanced, encompassing data ingestion, validation, model training, evaluation, and deployment. The Elastic Net model recorded a root mean square error (RMSE) of 0.660, a mean absolute error (MAE) of 0.511, and an R-squared (R^2) value of 0.311. While these performance metrics indicate a moderate level of predictive accuracy, they also highlight areas for potential improvement, particularly in capturing more variance in the target variable.

The integration of MLflow for tracking experiments and managing model versions, along with deployment on AWS EC2, guarantees scalability and reproducibility. The collaborative use of DagsHub enhanced model registry management and facilitated smooth development workflows. These features tackle significant challenges in constructing machine learning models and establish a solid foundation for creating scalable machine

Current Journal of Engineering and Science Research
Volume 2, Issue 1, pp 9-14, January 2025
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2025.201002
Available Online at https://www.cjesr.com

learning systems. Despite the project's achievements, several key challenges remain, including enhancing model performance, automating retraining processes, and managing data variability more effectively. Addressing these issues will further strengthen the robustness and efficiency of the development workflow.

This project exemplifies the real-world application of MLOps, bridging the gap between machine learning research and production settings. It underscores the necessity of a multidisciplinary approach that merges knowledge from data science, software engineering, and DevOps to develop dependable and scalable machine learning systems. This work enhances the understanding of MLOps and its critical role in operationalizing machine learning, while also providing a framework for future research and development in this field.

REFERENCES

[1] Lwakatare. 2020. DevOps for AI - Challenges in Development of AI-enabled Applications. (2020). DOI:https://doi.org/10.23919/SoftCOM50211.2020.9238323.

[2] Cedric Renggli, Luka Rimanic, Nezihe Merve Gürel, Bojan Karlaš,Wentao Wu, and Ce Zhang. 2021. A Data Quality-Driven View of MLOps.1 (2021), 1–12. Retrieved from http://arxiv.org/abs/2102.07750.

[3] Michael D. Myers and Michael Newman. 2007. The qualitative interview in IS research: Examining the craft. Information and Organization, 17(1), 2–26. DOI: https://doi.org/10.1016/j.infoandorg.2006.11.001.

[4] Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. 2001. Manifesto for Agile Software Development. (2001).

[5] Juliet M. Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. Qual. Sociol. 13, 1 (1990), 3–21. DOI:https://doi.org/10.1007/BF00988593.

[6] Behrouz Derakhshan, Alireza Rezaei Mahdiraji, Tilmann Rabl, and Volker Markl. 2019. Continuous deployment of machine learning pipelines. Adv. Database Technol. - EDBT 2019-March, (2019), 397–408. DOI:https://doi.org/10.5441/002/edbt.2019.35.

[7] Jane Webster and Richard Watson. 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly, 26(2), xiii–xxiii. DOI: https://doi.org/10.1.1.104.6570.

[8] Willem Jan van den Heuvel and Damian A. Tamburri. 2020. Model-driven ml-ops for intelligent enterprise applications: vision, approaches and challenges. Springer International Publishing. DOI: https://doi.org/10.1007/978-3-030-52306-0_11.

[9] Bojan Karlaš, Matteo Interlandi, Cedric Renggli, Wentao Wu, Ce Zhang, Deepak Mukunthu Iyappan Babu, Jordan Edwards, Chris Lauren, Andy Xu, and Markus Weimer. 2020. Building Continuous Integration Services for Machine Learning. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (2020), 2407–2415. DOI:https://doi.org/10.1145/3394486.3403290.

[10] Antonio Molner Domenech and Alberto Guillén. 2020. Ml-experiment: A Python framework for reproducible data science. J. Phys. Conf. Ser. 1603, 1 (2020). DOI:https://doi.org/10.1088/1742-6596/1603/1/012025.

[11] Lwakatare. 2020. From a Data Science Driven Process to a Continuous Delivery Process for Machine Learning Systems. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 12562 LNCS, (2020), 185–201. DOI:https://doi.org/10.1007/978-3-030-64148-1_12.

[12] Leonardo Leite, Carla Rocha, Fabio Kon, Dejan Milojicic, and Paulo Meirelles. 2019. A survey of DevOps concepts and challenges. ACM Comput. Surv. 52, 6 (2019). DOI:https://doi.org/10.1145/3359981.

[13] Martin Rütz. 2019. DEVOPS: A SYSTEMATIC LITERATURE REVIEW. Inf. Softw. Technol. (2019).

[14] Ulrike Schultze and Michel Avital. 2011. Designing interviews to generate rich data for information systems research. Inf. Organ. 21, 1 (2011), 1–16. DOI:https://doi.org/10.1016/j.infoandorg.2010.11.001.

[15] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering - A systematic literature review. Inf. Softw. Technol. 51, 1 (2009), 7–15. DOI:https://doi.org/10.1016/j.infsof.2008.09.009.

Current Journal of Engineering and Science Research
Volume 2, Issue 1, pp 9-14, January 2025
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2025.201002
Available Online at https://www.cjesr.com

Citation of this Article:

Dewa Rudy Jatayu. (2025). An Automated Comprehensive Machine Learning Pipeline for Predicting Wine Quality. *Current Journal of Engineering and Science Research*. 2(1), 9-14. Article DOI: https://doi.org/10.47001/CJESR/2025.201002

***** End of the Article *****