**MLOps Assignment 1**

**Task M2 Summary**

**Description of the Work Completed**

This project focused on implementing MLOps practices using **MLflow** for experiment tracking and **DVC** for data versioning, applied to a machine learning task involving wine quality prediction.

The combination of MLflow and DVC in this project exemplifies how MLOps tools can enhance the robustness and reproducibility of machine learning projects, aligning with industry best practices for efficient model development and deployment.

The main components of the project were:

|  |  |
| --- | --- |
| Experiment Tracking with MLflow: | * We implemented a machine learning pipeline that involved training three different models:   + Decision Tree,   + Logistic Regression, and   + Support Vector Machine (SVM). * Each model was trained on a standardized dataset split into training and testing sets. * The pipeline was designed to log model parameters, evaluation metrics (accuracy, precision, recall, and F1-score), and model artifacts using MLflow. * This enabled us to track and compare the performance of each model across different runs. |
| Data Versioning with DVC: | * We employed DVC to manage the versioning of the winedata.csv dataset, ensuring reproducibility and efficient data management. * DVC was used to track the dataset and facilitate easy switching between different versions, thus providing a robust mechanism to manage changes in the data over time. |

**Justification for the Choices Made**

|  |  |
| --- | --- |
| MLflow for Experiment Tracking: | * **Ease of Use:** MLflow offers a user-friendly interface and seamless integration with Python, making it easy to implement and track experiments. * **Comprehensive Logging:** MLflow's capability to log various metrics and parameters provides valuable insights into model performance, facilitating better decision-making. * **Reproducibility:** By tracking experiments, we ensure that results can be reproduced and validated, which is crucial for maintaining model performance over time. |
| DVC for Data Versioning: | * **Data Management:** DVC's ability to handle large datasets efficiently and track data versions allows for consistent and reliable data management. * **Reproducibility:** Versioning datasets ensure that experiments can be replicated with the exact data used during initial model training. * **Integration with Git:** DVC seamlessly integrates with Git, providing a familiar workflow for version control and collaboration. |