

## АВТОМАТИЗАЦИЯ ПРОЦЕССОВ ВНЕДРЕНИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ В ПРОИЗВОДСТВЕННЫЕ СИСТЕМЫ НА ОСНОВЕ KUBERNETES

Трапезин А. (Университет ИТМО)

Научный руководитель – инженер факультета прикладной информатики

Филянин И.В. (Университет ИТМО)

**Введение.** В современных промышленных предприятиях широко применяются методы машинного обучения (ML) для повышения эффективности производства и принятия управленческих решений [1]. Однако успешная интеграция моделей ML в рабочие процессы требует значительных усилий по автоматизации развертывания, мониторинга и масштабирования [2]. Традиционные подходы, включающие ручное администрирование и настройку инфраструктуры, сопряжены с высокими временными затратами и риском возникновения ошибок. В ответ на эти вызовы активно развивается концепция MLOps, обеспечивающая единый жизненный цикл моделей машинного обучения [3]. В данной работе рассматривается разработка и внедрение Kubernetes-оператора, который интегрирует MLflow для управления версиями моделей и Seldon Core для их развертывания, что позволяет автоматизировать процессы внедрения ML в производственную среду.

**Основная часть.** Цель исследования – разработка механизма автоматизированного развертывания и поставки моделей машинного обучения, обеспечивающего интеграцию MLflow и Seldon Core в Kubernetes для упрощения процесса доставки моделей в промышленную среду. В рамках работы реализован Kubernetes-оператор, который выполняет следующие функции:

- 1) Автоматизированное отслеживание изменений в репозитории моделей MLflow и инициирование развертывания новых версий.
- 2) Генерация SeldonDeployment CRD для обеспечения интеграции моделей в кластер Kubernetes без необходимости ручной настройки.
- 3) Реализация стратегий управления трафиком, включая A/B-тестирование и постепенное увеличение нагрузки на новые версии моделей.
- 4) Интеграция с системами мониторинга (Prometheus, Grafana) для автоматического сбора метрик и оценки производительности моделей.

Результаты экспериментов подтвердили эффективность предложенного решения: внедрение оператора позволило сократить среднее время развертывания моделей (TTD) на 45%, а также уменьшить количество ошибок конфигурации с 15% до 3%. Это свидетельствует о значительном повышении надежности и скорости развертывания ML-моделей в промышленной среде.

**Выводы.** В ходе работы разработан Kubernetes-оператор для автоматизации развертывания и управления моделями машинного обучения в среде Kubernetes. Реализованы механизмы автоматического отслеживания изменений в MLflow, генерации SeldonDeployment CRD и управления нагрузкой посредством A/B-тестирования. Дальнейшие исследования будут направлены на расширение функционала оператора, включая поддержку динамической адаптации моделей и интеграцию с дополнительными инструментами управления инфраструктурой.

### Список использованных источников:

1. Kreuzberger D., Kühl N., Hirschl S. Machine Learning Operations (MLOps): Overview, Definition, and Architecture // IEEE Access. – 2023. – Т. 11. – С. 31866-31879. – DOI: 10.1109/ACCESS.2023.3262138.
2. Tamburri D. A. Sustainable MLOps: Trends and Challenges // 22nd International

Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC). – Timisoara, Romania, 2020. – C. 17-23. – DOI: 10.1109/SYNASC51798.2020.00015.

3. Wazir S., Kashyap G. S., Saxena P. MLOps: A Review // arXiv. – 2023. – DOI: 10.48550/ARXIV.2308.10908.