BigDataCourse

Лабораторная работа #1 (Классический жизненный цикл разработки моделей машинного обучения)

В рамках данной работы были разработаны CI/CD pipeline для ML модели с достижением метрик моделей и качества.

Задача - классификация пингвинов по исходным признакам. Использованная модель - KNN (experiments/knn.sav), dataset (penguins) (https://www.kaggle.com/parulpandey/palmer-archipelago-antarctica-penguin-data)

Конфигурации: Dockefile (Dockerfile), Docker-Compose (docker-compose.yml), CI/CD (.github/workflows/docker-image.yml)

Также были реализованы тесты (src/unit_tests)

Лабораторная работа #2 (Взаимодействие с источниками данных)

В рамках данной работы была произведена выгрузка исходных данных и отправка результатов модели с использованием опередлнного источника данных.

В качестве источника данных был выборан PostgreSQL.

Реализация базы данных находится тут (src/database.py)

Все креды для доступа к базе данных лежат в Github secrets

Лабораторная работа #3 (Размещение секретов в хранилище)

В рамках данной работы все секреты(логин/пароль/хост и т.д.) были размещены в хранилище, с помощью которого с ними и было произведено взаимордействие.

В качестве хранилища был выбран Hashicorp Vault (vault-cli)

Вся работа с секретами находится тут (.github/workflows/docker-image.yml)

Vault находится в stateless состоянии, так как все секреты хранятся в github secrets.

Лабораторная работа #4 (Интеграция Apache Kafka сервиса)

В рамках данной работы были реализованы Kafka Producer и Kafka Consumer, а также была произведена интеграйия их в проект.

Реализация Kafka Producer тут (src/kafka_producer.py)

Реализация Kafka Consumer тут (src/kafka_consumer.py)