IT3385 MACHINE LEARNING OPERATIONS

Assignment (45%)

AY2025 Semester 2

Table of Contents

Introduction

Machine Learning Operations (MLOps) streamlines the process of deploying Machine Learning models to production and then maintaining and monitoring them. MLOps is a collaborative function, often consisting of data scientists, ML engineers, and DevOps engineers. The word MLOps is a compound of two different fields i.e. machine learning and DevOps from software engineering.

MLOps can encompass everything from the data pipeline to machine learning model production. In some places, you will see MLOps implementation is only for the deployment of the machine learning model but you will also find enterprises with the implementation of MLOps across many different areas of ML Lifecycle development, including Exploratory Data Analysis (EDA), Data Preprocessing, Model Training etc.

In this assignment, you will be working with **one/two teammate(s)** to build machine learning models and deploy them using a web application. Please appoint one team member as the leader. The implementation by your team should consider the various aspects of the MLOps lifecycle, i.e. integration with model generation (software development lifecycle and continuous integration/continuous delivery - CI/CD), orchestration, deployment, to deployment metrics etc.

Tasks Involved in Assignment

Below is the list of tasks to be performed in this assignment

S/N	Task	Individual / Team
1.	Perform an Exploratory Data Analysis on a Given Dataset	Individual
2.	Train, validate and develop a Machine Learning pipeline using PyCaret.	Individual
3.	Build and deploy a front-end web application with real-time prediction	Individual
4.	Set up development and deployment environment according to MLOps Lifecycle	Team

Please note that for individual components, each member is required to work on **ONE** dataset given in Annex A. You should not use the same dataset as your teammate(s). Each of you will create an ML Pipeline for your modelling workflow.

The base marks of this assignment are **45 marks,** and it constitutes **45%** of your total ICA marks for this competency unit.

Copy work from other people or the internet is strictly prohibited. If found, it will be considered a case of plagiarism and is subject to disciplinary actions

Late Submission Penalty

Late submission of assignment may subject to penalty as shown below.

No. of Calendar Days	Penalty
If the submission is <= 5 working days	Cap at 50% (of the base marks)
If the submission is > 5 working days	0 marks will be awarded

Submission

Below are the required deliverables for this assignment, only the appointed team leader is to submit these via Brightspace:

1. The source codes from your source code repository in a zip file format for the whole team.

   Note: do not include (installed) libraries or packages which will make your zip file very big and as a result you might be unable to upload it to Brightspace.

2. A demonstration video presented by the team, with a maximum duration of 15 minutes and in MP4 format. All team members should play a part in the presentation. You can use Powerpoint slides to guide your presentation. Please submit the Powerpoint slides if used.

3. A README file for the deployment guide, user guide, team members information & folder structure, along with the URLs to the web application deployed and the source code repository.

Please be reminded to submit all the deliverables via Brightspace **by 22 Feb 2026 (Sunday) at 2359hrs.** Refer to **Annex B** for detailed assessment rubrics.

Task 1: Perform an Exploratory Data Analysis on a Given Dataset (5 marks)

Exploratory Data Analysis (EDA) is an approach that is used to analyse the data and discover trends, and patterns, or check assumptions in data with the help of statistical summaries and graphical representations.

You may make use of Python libraries and visualisation libraries to assess the relationship (correlation, distribution etc) among the variables.

Task 2: Train, validate and develop a Machine Learning pipeline using PyCaret. (10 marks)

You may train and validate your models in Integrated Development Environment (IDE) or Jupiter Notebook either on your local machine or on the cloud. You are required to use PyCaret to develop a machine-learning pipeline and train your models.

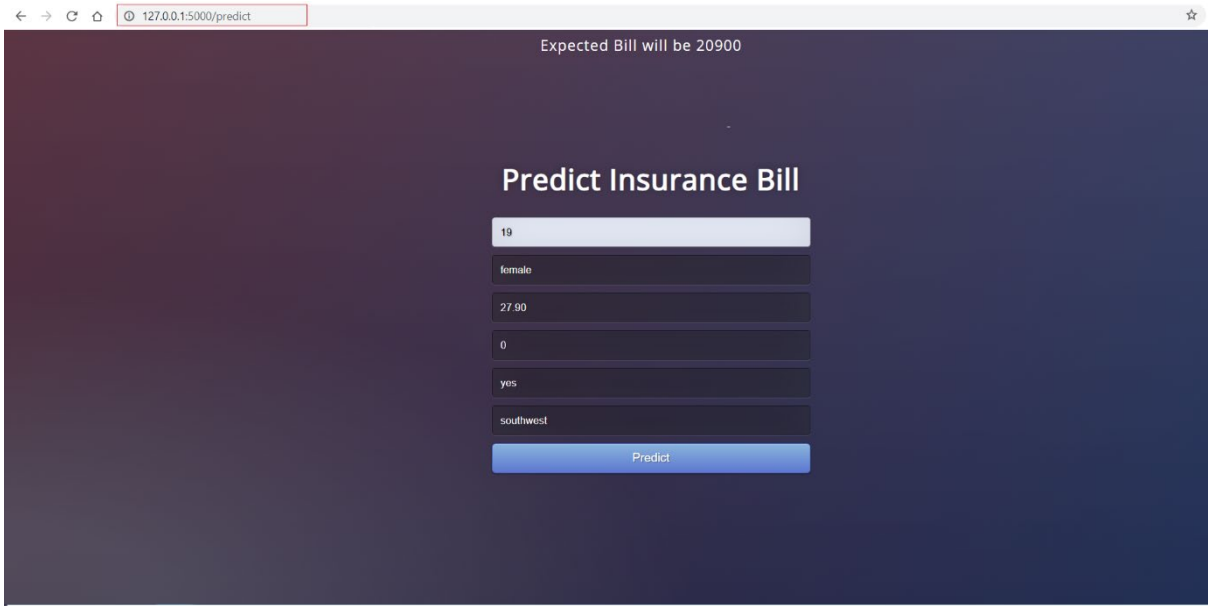
1. The training environment should be initialised with some additional preprocessing tasks such as scaling and normalisation, automatic feature engineering and binning continuous data into intervals etc based on your modelling and dataset which subsequently creates a transformation pipeline.
2. Models are trained and evaluated according to multiple performance metrics using k-fold cross-validation.
3. Analyse the performance of a trained model on the test set using the plot_model or evaluate_model function.
4. Use the final model to generate predictions on unseen data with the predict_model function.
5. Save the entire pipeline
6. Enable experiment logging during the training of models using MLFlow
7. Proper documentation is required for each step
8. Register the model with MLFlow

Task 3: Build and deploy a front-end web application with real-time prediction (15 marks)

Once your machine learning pipeline and model are ready, you can start building a web application that can connect to them and generate predictions on new data in real time. There are two parts to this application:

- Front-end (web-based) to get inputs from users
- Back-end (developed using Flask, Django or Streamlit with Python) – connect the model and return predictions based on a given input.

Below is a simple example of a real-time prediction of insurance bills based on inputs entered by users. You are advised to develop an attractive, modern and user-friendly user interface.



The web application should be tested on a local machine and deployed to the cloud or any platform as a service (PaaS) that enables the deployment of web apps based on a managed container system, with integrated data services and a powerful ecosystem e.g. Heroku, Streamlit or Render.

As you are working in a team, you would need to discuss with your team on the common approach to develop the web application. Your web application should appear as integrated.

Task 4: Set up development and deployment environment according to MLOps Lifecycle (15 marks)

The development and deployment environment for your team should be created according to the best practices for ML Model productization. Below are some of the recommendations for ML Model productization

- Follow a standard ML project folder structure
- Use Poetry for the dependency management
- Use Hydra for managing configuration files and minimizing hard coding
- Use DVC for data version control
- Use Branching for source code version control
- Use CI/CD feature in Github/Gitlab

The environment should support collaborative work, easy check-in & out of codes and quick deployment. The team is required to do a 15-minute presentation for the demonstration of the entire MLOps lifecycle using their environment and ML web application.

Annex A: Datasets

1. Measurements of geometrical properties of kernels belonging to three different varieties of wheat (Kama, Rosa and Canadian) - https://www.kaggle.com/datasets/dongeorge/seed-from-uci

2. Air Pollution, Alcohol, Smoking & Risk of Lung Cancer - https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

3. Comprehensive Health Information for Alzheimer's Disease - https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset

4. Heart Failure Prediction Dataset - 11 clinical features for predicting heart disease events - https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Annex B: Assessment Rubrics

Individual Component

Criteria	Advanced	Proficient	Functional	Developing	Not Competent
Task 1 - Perform an Exploratory Data Analysis on a Given Dataset **(5 marks)**	Intense data exploration by looking at data in many different ways. All the visualisations address the required analysis concisely and clearly. *(4.1 – 5 marks)*	Plenty of data exploration and investigation on the relationship among the variables. Most of the visualisations address the required analysis concisely and clearly. *(3.1 – 4 marks)*	Some data exploration and investigation on the relationship among the variables. At least half of the visualisations address the required analysis concisely and clearly. *(2.1 – 3 marks)*	Limited data exploration and investigation on the relationship among the variables. Only a few visualisations address the required analysis concisely and clearly. *(1.1 – 2 marks)*	No evidence of data exploration and investigation on the relationship among the variables. The visualisation did not address the required analysis concisely and clearly. *(0 – 1 mark)*
Task 2 - Train, validate and develop a Machine Learning	Performed thorough and comprehensive data cleansing	Performed most of the required data cleansing and transformation	Performed some of the required data cleansing and transformation	Performed little data cleansing and transformation	No evidence of data cleansing and transformation

pipeline using PyCaret. **(10 marks)**	and transformation Able to choose the best performing model by performing thorough tuning for the best hyper-parameter values Uses the comprehensive performance metric for the model evaluation *(8.1 – 10 marks)*	Able to choose the best performing model by performing tuning for the good hyper-parameter values. Uses the appropriate performance metric for the model evaluation *(6.1 – 8 marks)*	Able to choose the best performing model using PyCaret function Uses some performance metric for the model evaluation *(4.1 – 6 marks)*	Able to choose the best performing model using PyCaret function Uses inappropriate performance metric for the model evaluation *(2.1 – 4 marks)*	Did not choose the best performing model using PyCaret function Did not use performance metric for the model evaluation *(0 – 2 marks)*
Task 3 - Build and deploy a front-end web application with real-time prediction **(15 marks)**	Source codes are working flawlessly, very organised and with detailed comments providing explanation for complex steps. Codes are well structured and follows good programming practices. ML application was deployed and generated real-time predictions for single and batch request. *(12.1 – 15 marks)*	Source codes are working fine, organised and properly commented. Codes are clear. Follows good programming practices. ML application was deployed and generated real-time predictions for single request. *(9.1 – 12 marks)*	Source codes are working with few bugs, organised and minimal comments. Codes are unclear and difficult to follow. Follows some good programming practices. ML application was deployed and generated real-time predictions for single request with some bugs. *(6.1 – 9 marks)*	Source codes are working with some bugs, disorganised and minimal comments. Codes are unclear and difficult to follow. Follows some good programming practices. ML application was deployed and no prediction generated. *(3.1 – 6 marks)*	Source codes are buggy, not working, disorganised and without comments. Hardly follows any good programming practices. ML application was not deployed *(0 – 3 marks)*

Group Component

Task 4 - Set up development and deployment environment according to MLOps Lifecycle					
Criteria	**Advanced**	**Proficient**	**Functional**	**Developing**	**Not Competent**
Set up development and deployment environment according to MLOps Lifecycle **(10 marks)**	All the tools and libraries are properly integrated and functional for supporting CI/CD operations in ML model deployment. *(8.1 – 10 marks)*	Majority of the tools and libraries are properly integrated and functional for supporting CI/CD operations in ML model deployment. *(6.1 – 8 marks)*	Some of the tools and libraries are properly integrated and functional for supporting CI/CD operations in ML model deployment. *(4.1 – 6 marks)*	Some of the tools and libraries are properly integrated but not functional for supporting CI/CD operations in ML model deployment. *(2.1 – 4 marks)*	None of the tools and libraries are used. *(0 – 2 marks)*
Presentation **(5 marks)**	Presentation is very clear, well organised, and compelling. *(4.1 – 5 marks)*	Presentation are mostly clear, reasonably well crafted and organised. *(3.1 – 4 marks)*	Presentation are mostly clear and reasonably organised. *(2.1 – 3 marks)*	Presentation are unclear and not organised. *(1.1 – 2 marks)*	Presentation are confusing, clearly unpolished and disorganised *(0 – 1 mark)*