Project ARI3205
Interpretable AI for Deep Learning Models

Konstantinos Makantasis

October 24, 2025

This document contains the details for the ARI3205 project, which is marked out of 100%; however, it is equivalent to 80% of the total mark for this unit. Individuals are encouraged to work in teams of not more than **TWO** members per team. Questions related to the project should ideally be discussed with the lecturer. While discussion between students is considered healthy, the final deliverable must be produced by you and not plagiarised.

The deadline for this project is 23:59 Monday, January 19th, 2026. Deliverables and the plagiarism form must be uploaded to the VLE. Projects submitted late will be penalised or may not be accepted. In your deliverable you should clearly state how the work was divided among the team members.

Introduction

This assignment assesses the student's capability in building and interpreting machine learning models following a methodologically robust approach. The focus is on students' ability to identify appropriate techniques for interpreting the behaviour of different machine learning models and justify their choices.

Dataset Description

You will use the following datasets:

1. **California Housing dataset** (regression): It is a public, robust tabular regression dataset with 20,640 samples and 8 predictive features, loadable in one line with scikit-learn (*sklearn.datasets.fetch_california_housing)*.

2. **Adult Census Income dataset from UCI** (binary classification): It is tabular dataset for binary classification with rich categorical and numerical features. It can be downloaded from here.

3. **Fashion-MNIST dataset** (image classification): It is s a dataset of 70,000 28×28 grayscale images of clothing items across 10 categories (60,000 train and 10,000 test), for benchmarking image classification models Instructions for downloading and loading it in Python can be found here.

Specifications

The project consists of the following four (4) parts. You should use the **Python** programming language, explain how you worked on each task and justify your decisions.

Part 1: Feature-Level Interpretability (30 marks)

You will use the California Housing and the Adult Census Income datasets in this part. You should train one feed-forward neural network for each dataset and apply the following interpretability techniques:

1. Partial Dependence Plots (PDP) and Individual Conditional Expectation (ICE) plots **(7 marks)**
    a. Use PDP to examine the average effect of at least two features.
    b. Use ICE plots to explore individual predictions for at least two features.
    c. Explain what insights PDP and ICE give about the model's behaviour.

2. Permutation Feature Importance (PFI) **(7 marks)**
    a. Use PFI to identify the most important features in the model.
    b. Explain what the term "important" means when using the PFI method.

3. Accumulated Local Effects (ALE) **(9 marks)**
    a. Implement ALE plots to investigate the local effects of feature changes.
    b. Compare ALE with PDP and discuss any differences in the interpretability of these techniques.

4. Global Surrogates **(7 marks)**
    a. Build an interpretable model to approximate the predictions of the feed-forward neural network model.
    b. Analyse the surrogate model's effectiveness and discuss when such approximations are helpful.

Part 2: Local Interpretability Techniques (30 marks)

You will use the Adult Census Income dataset in this part. You should use the feed-forward neural network you trained in Part 1 and apply the following interpretability techniques:

1. Local Interpretable Model-agnostic Explanations (LIME) **(10 marks)**
    a. Apply LIME to explain individual predictions from the feed-forward neural network model.
    b. Explain how LIME approximates the local decision boundary and what the interpretation suggests.

2. Shapley Additive Explanations (SHAP) **(10 marks)**
    a. Use SHAP values to explain model predictions.
    b. Compare SHAP explanations with LIME and discuss any differences between the two.

3. Anchors **(10 marks)**
    a. Implement Anchors to interpret model predictions for specific cases.
    b. Compare Anchors with LIME and SHAP, and discuss any differences between them.

Part 3: Example-Based Explanations (20 marks)

You will use the Adult Census Income dataset in this part. You should use the feed-forward neural network you trained in Part 1 and apply the following interpretability techniques:

1. Counterfactual Explanations **(10 marks)**
    a. Generate counterfactuals for at least two incorrect predictions.
    b. Discuss the counterfactual's importance for debugging models and decision-making.

2. Prototypes and Criticisms **(10 marks)**
    a. Identify prototypes and criticisms from the dataset.
    b. Discuss how prototypes and criticisms can be used within the context of interpretable AI.

Part 4: Interpretable AI for Computer Vision (20 marks)

You will use the Fashion-MNIST dataset in this part. You should use a convolutional neural network (you can use a pre-trained network or train one from scratch) and apply the following interpretability techniques:

1. Integrated Gradients **(10 marks)**
    a. Use integrated gradients to interpret model predictions.
    b. Discuss the results obtained.

2. Grad-CAM **(10 marks)**
    a. Implement Grad-CAM to visualise where the CNN is focusing during its predictions.
    b. Discuss the results obtained.

Deliverables

1. **Python Jupyter Notebook**: submit your work via a Python Jupyter Notebook with the filename *"yourFirstname_yourSurname.ipynb"*. You may import and use any publicly available Python packages you deem necessary.
2. **PDF:** The pdf file generated by the Python Jupyter Notebook.
3. Filled in and signed **plagiarism form.**

Final Remarks

If you have difficulties, feel free to contact the lecturer. Any issues, including technical problems, should be identified and highlighted as early as possible to ensure timely resolution.

Good luck!!!