PASS TASK (WEEK 6)

About this task

Step-1

At the completion of week 5 and 6 modules, you are required to complete a lesson review to tell us what you learnt and how you learnt it by submitting evidence requested at the end of this file.

Step-2

Your tutor will then review your submission and will give you feedback. If your submission is incomplete the tutor will ask you to include missing parts. Tutor can also ask follow-up questions, either to clarify something that you have submitted or to assess your understanding of certain topics.

Feedback and submission deadlines

Feedback deadline: 15th May (No submission before this date means no feedback!)

Submission deadline: Before creating and submitting portfolio.

Evidence of Learning

1. Submit a summary report (pdf format) in Ontrack (https://ontrack.deakin.edu.au)
    1.1. Summarise the main points that is covered in the week 5 and 6.
    1.2. Provide summary of your reading list – external resources, websites, book chapters, code libraries, etc.
    1.3. Reflect on the knowledge that you have gained by reading contents of the week 5 and 6 with respect to machine learning.
    1.4. Attempt the quiz given in weekly content (5.18 and 6.15) and add screenshot of your score (>85% is considered completion of this task) in this report.
2. Complete the following problem-solving task given in weekly content, and submit your code file (.ipynb) separately to OnTrack (https://ontrack.deakin.edu.au).

Problem Solving Task

Instructions:
Ensure you provide well-structured code for each question. Clearly explain your reasoning where required and submit all necessary files, including the modified dataset after performing the required transformations.

Dataset Description: Energy Efficiency Data (Dataset.csv)
This dataset contains information about **768 simulated residential buildings**, each described by architectural design parameters. The goal of this dataset is to predict **how much energy is needed to heat and cool** the buildings under different design configurations.

Column	Name	Description
X1	Relative Compactness	A measure of how compact the building shape is. Higher compactness often improves energy efficiency.
X2	Surface Area (m^2)	Total external surface area of the building.
X3	Wall Area (m^2)	Total area of the external walls.
X4	Roof Area (m^2)	Total roof area.
X5	Overall Height (m)	Height of the building (either single-storey or two-storey).
X6	Orientation	Building orientation: 2 = East, 3 = South, 4 = West, 5 = North.
X7	Glazing Area (%)	Total area of windows as a percentage of the building's floor area. Values used are 0%, 10%, 25%, and 40%, representing the extent of window coverage.
X8	Glazing Area Distribution	Describes how the glazing area is distributed across the sides of the building. Values range from 0 to 5, representing: 0 = No glazing, 1 = 55% North-facing + 15% others, 2 = 55% East-facing + 15% others, 3 = 55% South-facing + 15% others, 4 = 55% West-facing + 15% others, 5 = Uniform (25% on each side). This feature influences how sunlight enters the building and can be treated as categorical.
Y1	Heating Load (kWh/m^2)	Amount of energy required to heat the building.
Y2	Cooling Load (kWh/m^2)	Amount of energy required to cool the building.

Q1: Splitting the Dataset

- Load the Energy Efficiency dataset (Dataset.csv).
- Create training and test datasets using two different strategies:
  - Random split: 70% for training, 30% for testing
  - Group-based split: Hold out all samples with two unique values of “X8” (Glazing Area Distribution) for testing
- Print the number of training and test samples for each method.

- Explain how the two splitting methods might affect how well your model performs on new building designs.

Q2: Orientation Distribution in Train/Test Splits (X6)

- For both the training and test datasets from Q1, create a bar chart showing how many buildings face each orientation using feature X6: 2 = East, 3 = South, 4 = West, 5 = North.
- Plot the number of buildings for each orientation in both training and test sets.
- Check whether certain orientations are over- or under-represented.
- Please explain the following in your answers
  - Are both sets balanced in terms of orientation?
  - If not, how could this affect your model's predictions?
  - Why is it important for your model to learn from all orientations?

Q3: Predicting Heating Load with Linear Regression

- Train a linear regression model to predict Heating Load (Y1) using the original input features (X1–X8).
- Use the training set from Q1 (random split).
- Evaluate your model using at least two metrics, such as:
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - R^2 Score
- In your answers please Explain:
  - Why did you choose these metrics?
  - How well does your model perform based on these results?

Q4. Using PCA to Reduce Dimensionality

- Apply Principal Component Analysis (PCA) to reduce the input features (X1–X8) to just the first three principal components.
- Use the same train-test split from Q3.
- Train a linear regression model using the PCA-transformed data to predict Heating Load (Y1).
- Compare the performance of this model with your model from Q3.
- In your answers please Explain:
  - Did PCA help or hurt performance?
  - What are the pros and cons of using only a few PCA components (like 3) instead of using all the original features (X1–X8) in your model?

Q5. Ridge Regression with PCA Features

- Apply **Ridge Regression (L2 regularisation)** using the same PCA-transformed data and split from Q4.
- Train the Ridge model to predict Heating Load (Y1).
- Evaluate performance using the same two metrics.
- Compare this model's results with your model from Q4.
- In your answers please Explain:
  - Did regularisation improve performance or stability?
  - Why might/ if regularisation be useful when working with reduced or correlated features?