BUSI 651 – Individual Assignment 1
Machine Learning Algorithms

Note: *Do not include the questions as well as dataset in your submission (to avoid similarity with other submissions)*

Question 1:

In this problem, we aim to develop a predictive model to estimate the energy consumption of buildings equipped with HVAC (Heating, Ventilation, and Air Conditioning) systems. The Dataset.csv contains information on various features that influence energy consumption in buildings. The goal is to leverage these features to accurately predict the energy usage (in kilowatt-hours) for different buildings. The table contains the following columns:

Room Area (sq. ft.)	The area of the rooms in the building, which exhibits very little variation and almost similar numbers across different buildings.
Number of Appliances	The count of appliances present in each building, which shows a moderate correlation of approximately 0.4 with the energy consumption.
Outside Temperature (°C)	The outside temperature recorded at each building's location, which demonstrates a strong negative correlation of approximately -0.7 with the energy consumption. As the temperature increases, the energy required for cooling decreases and vice versa.
Insulation Thickness (inches)	The thickness of insulation in the building's walls, which exhibits a high positive correlation of nearly 0.8 with the energy consumption. Better insulation leads to more efficient temperature control and reduces energy consumption.
Building Type	A categorical feature representing the type of building, such as residential or commercial.
HVAC System	Another categorical feature representing the type of HVAC system installed in the building, such as Central AC, Split AC, or Window AC.
Average Temperature in last 24 hours (°C)	An additional numerical column highly correlated with the Outside Temperature, reflecting a similar trend in terms of impact on energy consumption.
Energy Consumption (kWh)	The output variable, representing the actual energy consumption (in kilowatt-hours) of each building based on the given features.

The objective of this problem is to build a predictive model that accurately estimates the energy consumption of buildings with HVAC systems. By leveraging the provided features, the model will predict the energy usage for new, unseen buildings, assisting homeowners, businesses, and energy providers in optimizing energy consumption and reducing costs.

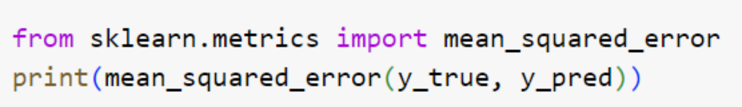
Required Analysis:

a) Determine the primary feature influencing energy consumption prediction? What about the secondary feature? Explain the reasons behind their significance.

b) Identify any feature that may not contribute significantly to prediction accuracy. What is your mitigation strategy? If no such feature found, provide justification for your claim.

c) Apply multiple linear regression to build a prediction model for energy consumption (y) based on the features. Feel free to modify the dataset to enhance prediction accuracy. Use the model to predict energy consumption for point1, point2, and point3:
<u>*point 1:*</u> Room Area (sq. ft.): 279, Number of Appliances: 16, Outside Temperature (°C): 20, Insulation Thickness (inches): 1.7, Building Type: Residential, HVAC System: Central AC, Average Temperature in last 24 hours (°C): 19, Energy Consumption (kWh): 385
<u>*point 2:*</u> Room Area (sq. ft.): 277, Number of Appliances: 22, Outside Temperature (°C): 15, Insulation Thickness (inches): 1.5, Building Type: Commercial, HVAC System: Split AC, Average Temperature in last 24 hours (°C): 14, Energy Consumption (kWh): 425
<u>*point 3:*</u> Room Area (sq. ft.): 276, Number of Appliances: 14, Outside Temperature (°C): 25, Insulation Thickness (inches): 2.2, Building Type: Residential, HVAC System: Window AC, Average Temperature in last 24 hours (°C): 26, Energy Consumption (kWh): 350

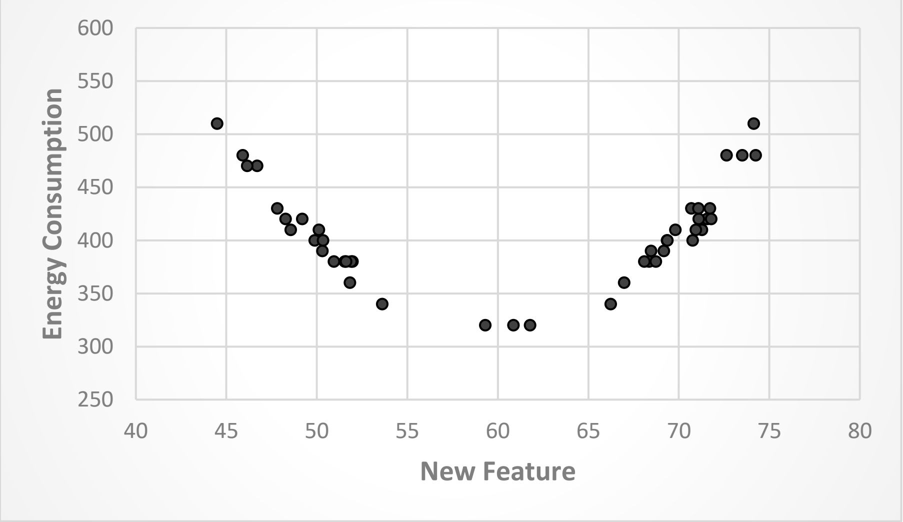
d) Compute the Mean Squared Error (MSE) regression loss using the "mean_squared_error" function from "sklearn.metrics" library, for these points:



```
from sklearn.metrics import mean_squared_error
print(mean_squared_error(y_true, y_pred))
```

where *y_true* is *ground truth (correct) target values* and *y_pred* is *estimated target values* based on the linear regression model. (Note that y_true and y_pred are <u>vectors</u> each containing three values corresponding to point1, point2, point3.)

e) Given the following scatter plot of a new feature X with respect to Energy Consumption, recommend whether to include it in the features for prediction. Provide reason.

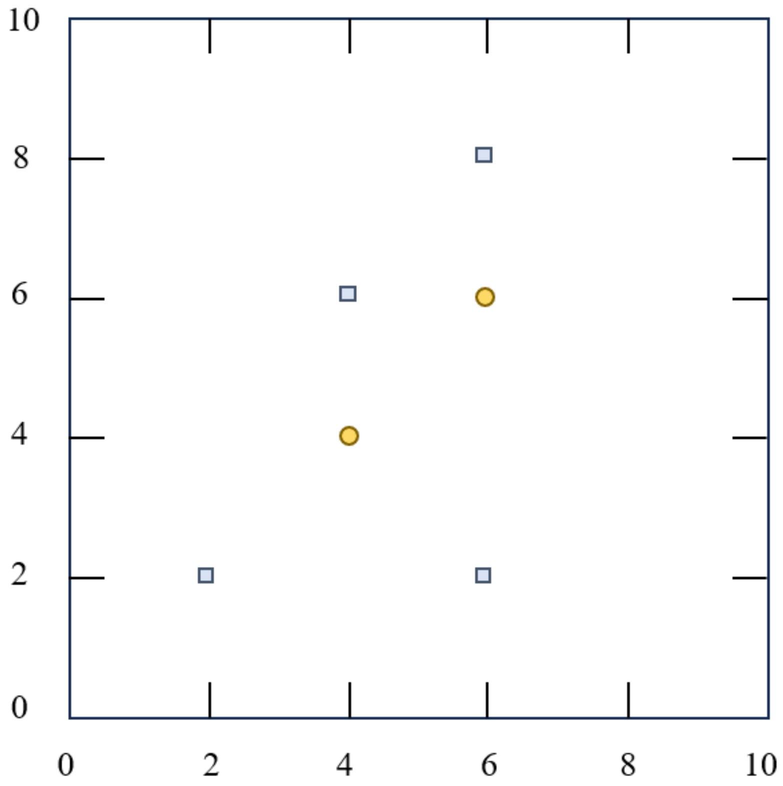


Question 2:

Consider the figure below where data points are divided into two classes, yellow circles and blue squares. Please answer the following questions:

a) Draw the decision boundary for KNN-algorithm when:
   - ○ K=1
   - ○ K=3

b) How will the following points be classified by K=1 and K=3 classifiers:
   - ○ point (8, 6)
   - ○ point (8, 4)



Question 3:

Given the following table where data points belong to two classes - and +. Using the KNN algorithm to predict the label (+,-) for the following points:

- point1: (7.81, 5.33)
- point2: (9.43, 5.29)

a) What is the predicted label based on K=1? What about k=3?
b) What is the index of the closest neighbors in each case?

Index	X1	X2	Y
1	8.27	5.59	+
2	1.58	5.87	-
3	5.92	5.87	-
4	9.44	5.83	+
5	2.11	5.57	-
6	4.71	5.94	+
7	3.82	5.84	+
8	6.98	5.91	-
9	3.15	5.42	-
10	8.9	5.94	-
11	7.65	5.77	+
12	9.83	5.29	-
13	1.94	5.36	+
14	7.13	5.28	-
15	5.77	5.47	-
16	4.36	5.31	+
17	5.09	5.65	-
18	3.42	5.24	+
19	2.76	5.71	+
20	9.6	5.52	-

Delivery:

For the final submission, create tables like this and fill them with your results/answers:

Question 1:

a) primary feature: secondary feature: **reason:**
b) feature not contributing: **mitigation strategy:** **If no such feature exists, justify your claim:**
c) Apply multiple linear regression: **energy consumption for point 1:** **energy consumption for point 2:** **energy consumption for point 3:**
d) Mean Squared Error (MSE) regression loss:
e) recommend to include the new feature (with reason):

Question 2:

Decision Boundary for K=1:

Decision Boundary for K=3:

Classify:

point (8, 6):
point (8, 4):

Question 3:

Predicted label based on K=1:
Point 1:
Point 2:
Predicted label based on K=3:
Point 1:
Point 2:

Index of closest neighbor for K=1:
For Point 1:
For Point 2:

Index of closest neighbors for K=3:
For Point 1:
For Point 2:

1.0 Grading Rubric

Key Points	**Grade Allocation (%)**
Format (font type, size, table, formulas), overall content, including references if required (APA Style)	10
Results, analysis and assumptions	80
Novelty and creativity in solution	10

N.B. Failure to comply with the above would result in low grades.

2.0 Format and Deadline

Submission: a single word/pdf document

Deadline: As posted in Course Shell