# COMP 1100 - Spring 2025 [Evan Hodges]
# Assignment 3 – Data Science Fundamentals

Note: This assignment mentions “Energy_Consumption_kWh”, which isn’t included in the dataset. I copy-pasted the “Energy_Consumption_kWh” column from last assignment into this dataset to compensate.

Problem 1

**2a**. Numerical: Temperature_C | Occupancy | Energy_Consumption_kWh | Humidity_% | CO2_Level_ppm
Categorical: Motion_Sensor | Building_ID | Energy_Usage_Class

**2b**. Target for…
Logistic Regression and Decision Tree: Energy_Usage_Class (“High” or “Low”)
Linear Regression: Energy_Consumption_kWh

**3a**.

| ---------- | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|
| Temperature_C | 21.943 | 3.294 | 11.9 | 31.6 |
| Occupancy | 243.145 | 136.577 | 0 | 499 |
| Energy_Consumption_kWh | 132.371 | 28.671 | 78.26 | 197.42 |
| Humidity_% | 50.169 | 10.025 | 14.3 | 77.6 |
| CO2_Level_ppm | 522.129 | 70.639 | 364.3 | 660.4 |

Observations: Temperature ranges fairly widely, occupancy varies dramatically, energy consumption has a mean of ~132 kWh (moderate to high energy usage), humidity averages ~50% but can drop to as low as ~14% or rise to ~78%, CO2 levels range from ~364ppm to over 660ppm (some rooms/buildings experience quite elevated CO2 concentrations).

Problem 2

**1.**

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 1 -R 1.0E-8 -num-decimal-places 4
Relation:     Smart_Building_Energy_Classification_Dataset-weka.filters.unsupervised.attribute.NumericToNominal-R1,7-weka.filters.unsupervised.attribute.Remove-R1,5,7-8
Instances:    200
Attributes:   4
              Temperature_C
              Occupancy
              Energy_Consumption_kWh
              CO2_Level_ppm
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


Linear Regression Model

Energy_Consumption_kWh =

      1.5014 * Temperature_C +
      0.2118 * Occupancy +
     -0.0146 * CO2_Level_ppm +
     55.5519

Time taken to build model: 0 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.9788
Mean absolute error                      4.5737
Root mean squared error                  5.8605
Relative absolute error                 19.088  %
Root relative squared error             20.3801 %
Total Number of Instances              200
```

**2a**. Temperature_C has the strongest immediate impact.

**2b**. The correlation coefficient is very close to 1 and the error indicators are all relatively low; the model did a good job of predicting Energy_Consumption_kWh.

Problem 3

**1.**

```
=== Run information ===

Scheme:       weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
Relation:     Smart_Building_Energy_Classification_Dataset-weka.filters.unsupervised.attribute.Remove-R1,4-5,7
Instances:    200
Attributes:   4
              Temperature_C
              Occupancy
              CO2_Level_ppm
              Energy_Usage_Class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                   Class
Variable             Low
=======================
Temperature_C    -0.5079
Occupancy        -0.0704
CO2_Level_ppm     0.0324
Intercept        10.5975


Odds Ratios...
                   Class
Variable             Low
=======================
Temperature_C     0.6018
Occupancy          0.932
CO2_Level_ppm      1.033


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         185               92.5    %
Incorrectly Classified Instances        15                7.5    %
Kappa statistic                          0.85
Mean absolute error                      0.105
Root mean squared error                  0.2375
Relative absolute error                 21.0007 %
Root relative squared error             47.4994 %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.920    0.070    0.929      0.920    0.925      0.850    0.980     0.981     Low
                 0.930    0.080    0.921      0.930    0.925      0.850    0.980     0.982     High
Weighted Avg.    0.925    0.075    0.925      0.925    0.925      0.850    0.980     0.981

=== Confusion Matrix ===

  a  b   <-- classified as
 92  8 |  a = Low
  7 93 |  b = High
```

**2a**. Accuracy: 92.5% | Precision: Low [0.925], High [0.921] | Recall: Low [0.920], High [0.930]

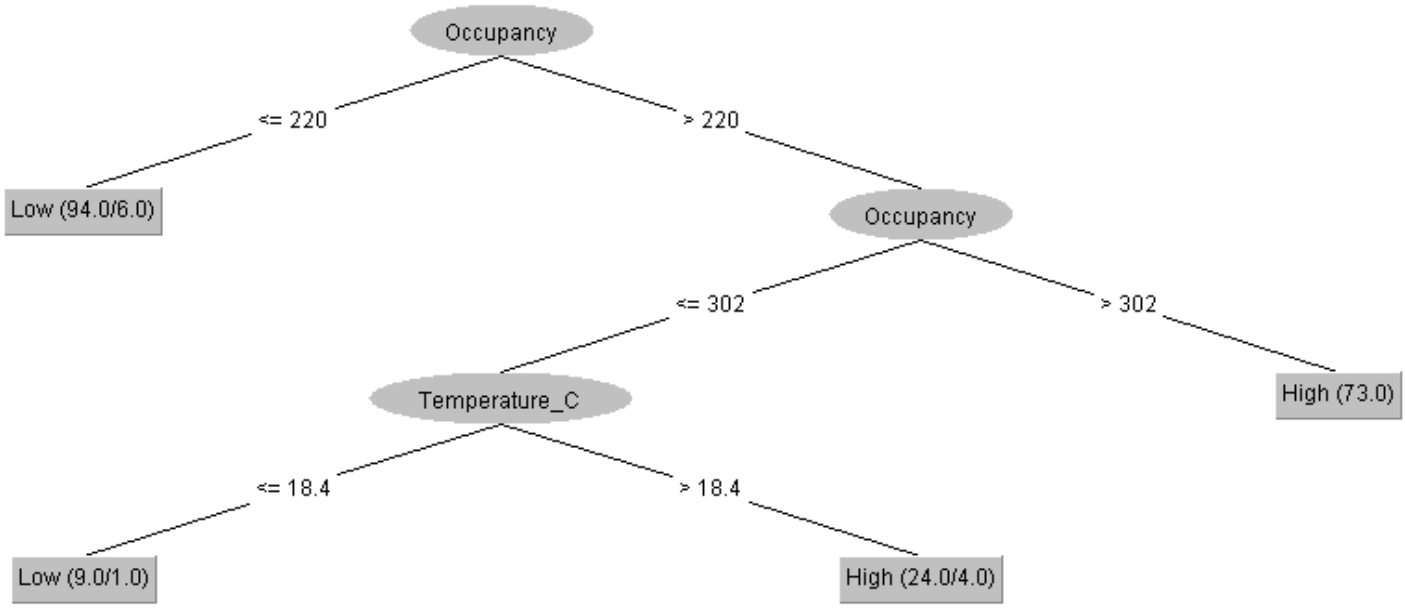
**2b**. Yes, it does. Temperature shows the most pronounced effect- 0.60 for Low.

Problem 4

**1.**

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Smart_Building_Energy_Classification_Dataset-weka.filters.unsupervised.attribute.Remove-R1,4-5,7
Instances:    200
Attributes:   4
              Temperature_C
              Occupancy
              CO2_Level_ppm
              Energy_Usage_Class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Occupancy <= 220: Low (94.0/6.0)
Occupancy > 220
|   Occupancy <= 302
|   |   Temperature_C <= 18.4: Low (9.0/1.0)
|   |   Temperature_C > 18.4: High (24.0/4.0)
|   Occupancy > 302: High (73.0)

Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         183               91.5    %
Incorrectly Classified Instances        17                8.5    %
Kappa statistic                          0.83
Mean absolute error                      0.1277
Root mean squared error                  0.2773
Relative absolute error                 25.5305 %
Root relative squared error             55.4627 %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.920    0.090    0.911      0.920    0.915      0.830    0.924     0.876     Low
                 0.910    0.080    0.919      0.910    0.915      0.830    0.924     0.925     High
Weighted Avg.    0.915    0.085    0.915      0.915    0.915      0.830    0.924     0.900

=== Confusion Matrix ===

  a  b   <-- classified as
 92  8 |  a = Low
  9 91 |  b = High
```

Tree View




**2a**. The first decision split is occupancy.

**2b**. Low usage: Occupancy ≤ 220 = Low | Occupancy > 220 but ≤ 302 and Temperature_C ≤ 18.4 = Low
High usage: Occupancy > 302 = High | 221 < Occupancy < 302 but Temperature_C > 18.4 = High

Problem 5

I applied Linear Regression, Logistic Regression, and a Decision Tree to predict building energy usage. Linear Regression (R = 0.9788, RMSE = 5.86) accurately forecasted kWh consumption, with temperature having the strongest impact. Logistic Regression achieved 92.5% accuracy, also highlighting temperature as key. The Decision Tree provided clear rules, using occupancy as its primary split. Although all models performed well, Linear Regression is best for continuous forecasting, while Logistic Regression and Decision Trees help classify usage patterns. Machine learning can reduce waste by automatically adjusting HVAC and occupancy-based systems to optimize energy efficiency.