Assignment Write-Up: Applied Data Science with ML and AI — Week 2

Task 1: Descriptive Statistics

1.1 Manual Calculations

Given dataset: `[55, 92, 78, 60, 85, 78, 90, 66, 73, 88]`

- Mean: Sum = 765, Count = 10, Mean = 765 / 10 = 76.5
- Median: Sorted data = `[55, 60, 66, 73, 78, 78, 85, 88, 90, 92]`; median = (78 + 78)/2 = 78
- Mode: Frequency dictionary shows 78 appears twice; all others appear once. Mode = 78
- Range: Max = 92, Min = 55, Range = 92 - 55 = 37
- Variance (Population):
    a. Squared differences from mean: 462.25, 240.25, 2.25, etc.
    b. Sum = 1122.5
    c. Variance = 1122.5 / 10 = 112.25
- Standard Deviation: √112.25 ≈ 10.59

1.2 Using NumPy

Used the following NumPy functions:

- `np.mean(), np.median(), np.var(), np.std(), np.max() - np.min()`

Used `scipy.stats.mode` to find the mode. All results from NumPy matched the manually computed values.

Task 2: Supervised Learning — Linear Regression

2.1 Manual Guess and MSE

Guessed line: `Y_pred = 7X + 40`
Predicted values were calculated using this equation.
Mean Squared Error (MSE) was computed as the average of squared differences between actual and predicted Y values.
MSE = 4.375

2.2 Using Scikit-learn

Used `LinearRegression()` to fit the model.

- **Coefficient (slope) = 6.964**
- **Intercept = 39.286**
- **MSE = 0.4464**
- **R^2 score = 0.9948**

The slope indicates the increase in salary per year of experience. The intercept represents the starting salary. A low MSE and high R^2 score indicate a good model fit.

Task 3: Supervised Learning — Logistic Regression

3.1 Manual Sigmoid Calculation

Used parameters m = 2 and c = -5, so z = 2X - 5.
Applied the sigmoid function: `1 / (1 + exp(-z))` to compute predicted probabilities.
Applied a threshold of 0.5 to classify each value as 0 or 1.
Accuracy = 9 correct predictions out of 10 = 0.9

3.2 Using Scikit-learn

Used `LogisticRegression(solver='liblinear')`

- **Coefficient = 2.197**

- **Intercept = -5.099**
- **Accuracy = 0.9**
- **Confusion Matrix:**

  lua
  CopyEdit

  ```
  [[4 1]

   [0 5]]
  ```

True Positives = 5, True Negatives = 4, False Positives = 1, False Negatives = 0.
The model performs well with only one misclassification.

Task 4: Unsupervised Learning — K-Means Clustering

4.1 Manual Iteration

Initial centroids:

- **Centroid 1 = [2, 10]**
- **Centroid 2 = [2, 5]**

Calculated Euclidean distances from each point to the two centroids and assigned them to the nearest one.
Computed new centroids by averaging the coordinates of the assigned points.

4.2 Using Scikit-learn

Used `KMeans(n_clusters=2, random_state=42, n_init=10)`
Printed final centroids and labels.
Compared to manual results. Assignments were mostly similar, with minor differences due to multiple iterations and better convergence in Scikit-learn's implementation.

Task 1 Output Questions (Descriptive Statistics)

Q: Are the results from manual and NumPy calculations the same?
A: Yes, all values obtained manually matched the results from NumPy functions.

Task 2 Output Questions (Linear Regression)

Q: What do the learned m and c tell you about the relationship between experience and salary in this model?
A: The slope (~6.96) indicates that salary increases by approximately 6.96 units (in thousands) for each additional year of experience. The intercept (~39.29) represents the estimated salary for zero years of experience.

Q: What does your calculated MSE tell you about the model's predictions for this dataset?
A: A low MSE (0.4464) suggests that the model's predicted salaries are very close to the actual values, indicating high accuracy.

Q: What does your calculated R^2 score tell you about how well the model fits this dataset?
A: An R^2 score of 0.9948 means that 99.48% of the variance in salary is explained by years of experience, implying an excellent model fit.

Task 3 Output Questions (Logistic Regression)

Q: What does the accuracy score tell you about this model on this dataset?
A: The accuracy of 0.9 indicates that 90% of the predictions are correct, showing strong model performance on this dataset.

Q: Explain what each part of your confusion matrix represents.
A:

- **True Positives (TP): 5 → predicted pass, actually pass**
- **True Negatives (TN): 4 → predicted fail, actually fail**
- **False Positives (FP): 1 → predicted pass, actually fail**
- **False Negatives (FN): 0 → predicted fail, actually pass**

Task 4 Output Questions (K-Means Clustering)

Q: How do the final centroids and labels from Scikit-learn compare to your manual results? Are they similar? Why might they differ?
A: The final cluster assignments were mostly similar to manual results. Differences may arise because Scikit-learn continues iterating until convergence, whereas the manual process was limited to the first or second iteration.