**Q0.**

1. The main difference between regression and classification is that regression predicts continuous values, whereas classification predicts categorical labels.
2. A confusion matrix is a table that is used to define the performance of a classification algorithm (sciencedirect.com). It shows true positives, false positives, true negatives, and false negatives. It helps us understand:

* How well the model distinguishes between classes.
* Whether the model is biased toward certain predictions.
* Where misclassifications occur.

1. SSE measures the total squared difference between actual and predicted values in regression. A lower SSE indicates a better-fitting model.
2. **Overfitting**: The model learns too much noise, performing well on training data but poorly on unseen data.

**Underfitting**: The model is too simple and fails to capture patterns, leading to poor performance on both training and test data.

1. Splitting the data into training and testing sets helps ensure the model learns patterns from the training data but is evaluated on unseen data. This prevents overfitting (where the model memorizes training data but performs poorly on new data).

Choosing **k** based on test accuracy or SSE helps find the best balance:

* **Too small k** → Model is too sensitive (overfitting).
* **Too large k** → Model is too simple (underfitting).

1. **Class Label Prediction:**

Strengths: simple and easy to interpret/ useful when only one clear decision is needed

Weaknesses: does not show confidence in the prediction/ can be misleading if the model is uncertain

**Probability Distribution Prediction**

Strengths: provides more information on uncertainty/ useful in high-risk decisions/ allows threshold-based decision-making

Weaknesses: more complex to interpret/ can be misleading if probabilities are not well-calibrated