Programming Question

Question 1 :The Mystery Polynomial [Programming Question]:

You have been given a dataset in regression\_data.csv. This data was generated from a noisy polynomial function with a degree of 6 or less. Your task is to find this function. Instructions: • Load the data and split it into an 80% training set and a 20% testing set. • Build and evaluate six polynomial regression models with degrees n = 1 through n = 6. • For each model, calculate and record error metrics of your choice (e.g., Mean Squared Error, R-squared) on both the training and testing sets. • Report the most likely polynomial function that generated the data, along with your justification based on the error metrics. Note: Increasing the degree of the polynomial will obviously reduce training error, so do not rely solely on training errors for model selection.

Question 2 :

KNN with Cross-Validation [Programming Question]:

You are given a dataset point KNN.csv. Each row corresponds to a 2D point with

coordinates (x1, x2) and a class label (0 or 1).

Your tasks are:

(a) Load the dataset from point KNN.csv, and split it into 80% training and 20%

testing sets.

(b) Implement the K-Nearest Neighbors (KNN) algorithm for classification.

(c) Use 5-fold cross-validation on the training set to select the optimal number

of neighbors k. The aim is to avoid both high bias (underfitting) and high

variance (overfitting).

(d) Report the following:

• The value of k selected by cross-validation.

• Training accuracy, validation accuracy, and test accuracy.

Question 3:

Using Python (scikit-learn), generate a synthetic regression dataset with n = 200 observations

and p = 5 predictors.

(a) Fit a linear regression model (parametric, inflexible).

(b) Fit a decision tree regressor (non-parametric, flexible).

(c) Compare their test mean squared errors (MSE) on a held-out test set.

(d) Plot the training error and test error of the decision tree model as a function of tree depth.

(e) Using the plot, explain how the bias–variance trade-off is illustrated. At what depth

do you see signs of overfitting?

(f) Discuss under what conditions the flexible method outperforms the inflexible one, and

vice versa.