Assignment: Polynomial Regression for Manufacturing Quality Prediction

Scenario:
You are working with a dataset from a manufacturing process. The goal is to understand how key process parameters, specifically Temperature (°C) and Pressure (kPa), relate to the final Quality Rating of the product. You suspect these relationships might be non-linear and want to explore the use of Polynomial Regression to potentially model them more accurately than simple linear models.

Dataset:
The dataset contains observations from a manufacturing process, including process parameters and the resulting product quality rating.

- Dataset Source: Manufacturing Data for Polynomial Regression on Kaggle
- Key Columns for this Assignment:
    - Temperature (°C): Process temperature (Input Feature).
    - Pressure (kPa): Process pressure (Input Feature).
    - Quality Rating: Product quality rating (Target Variable).
    - *Note: The dataset also contains interaction and derived features (Temperature x Pressure, Material Fusion Metric, Material Transformation Metric), but for this assignment, we will primarily focus on modeling Quality Rating based on Temperature (°C) and Pressure (kPa) individually using polynomial features.*
- Loading the Data: You will need to download the dataset from the Kaggle link and load it into a pandas DataFrame (e.g., manufacturing_df). Ensure you handle the file path correctly after downloading.

```
# Example loading (replace 'path/to/your/downloaded/file.csv' with the actual path)

# import pandas as pd

# file_path = 'path/to/your/downloaded/file.csv'

# manufacturing_df = pd.read_csv(file_path)

# print(manufacturing_df.head())

# print(manufacturing_df.info())
```

Assignment Questions:

Question 1: Data Loading and Initial Visualization

- (a) Load the manufacturing dataset from the downloaded CSV file into a pandas DataFrame. Display the first few rows, check the column names, and view the data types using .info().
- (b) Create two separate scatter plots:
    - Plot 1: Temperature (°C) (x-axis) vs. Quality Rating (y-axis).
    - Plot 2: Pressure (kPa) (x-axis) vs. Quality Rating (y-axis).
      Label your axes appropriately and give each plot a title.
- (c) Based *only* on the visual inspection of the scatter plots, does a simple linear relationship seem adequate to model Quality Rating based on Temperature (°C)? What about based on Pressure (kPa)? Explain your reasoning for each plot. Do either suggest a polynomial relationship might be more suitable?

Question 2: Linear vs. Quadratic Models for Temperature

- (a) Select Temperature (°C) as your feature (X) and Quality Rating as your target (y). Fit a simple Linear Regression model (degree 1) using the *entire* dataset. Report the R-squared (R^2) score and Mean Squared Error (MSE).
- (b) Using sklearn.preprocessing.PolynomialFeatures, create degree-2 polynomial features for Temperature (°C). Fit a Linear Regression model using these features on the *entire* dataset. Report the R^2 score and MSE.
- (c) Create a scatter plot of Temperature (°C) vs. Quality Rating. On this same plot, overlay the prediction lines from both the simple linear model (degree 1) and the quadratic model (degree 2) trained in parts (a) and (b). Add a legend.
- (d) Compare the metrics (R^2, MSE) and the visual fit from part (c). Does the quadratic model appear to provide a better fit for the relationship between Temperature (°C) and Quality Rating compared to the linear model? Justify your answer.

Question 3: Exploring Higher-Degree Polynomials for Temperature & Overfitting

- (a) Split the data into a training set (80%) and a testing set (20%) using Temperature (°C) as the feature and Quality Rating as the target. Use random_state=42 for reproducibility.
- (b) Fit polynomial regression models for degrees 1, 2, 3, 4, 5, and 8, using *only* Temperature (°C) as the input feature. For each degree:

    - Create polynomial features (fit on training data, transform train and test).
    - Train a linear regression model on the transformed training data.
    - Calculate and record the R^2 score on *both* the training set and the testing set.
- (c) Create a plot with the polynomial degree (1, 2, 3, 4, 5, 8) on the x-axis and the R^2 scores on the y-axis. Plot both training R^2 and testing R^2 scores on the same graph. Include a legend.
- (d) Analyze the plot from part (c). How do the training and testing R^2 scores change as the polynomial degree increases? Is there evidence of overfitting? If so, around which degree does overfitting seem to become significant for the Temperature feature? Explain.

Question 4: Optimal Model Selection for Temperature using Cross-Validation

- (a) Using the *entire* dataset (Temperature (°C) as X, Quality Rating as y), perform k-fold cross-validation (use k=5 or k=10) to evaluate polynomial regression models for degrees 1 through 6.
    - For each degree, create a Pipeline containing PolynomialFeatures and LinearRegression.
    - Use cross_val_score with the pipeline, the full dataset, your chosen cv, and scoring='neg_mean_squared_error'.
    - Calculate and record the average *positive* MSE for each degree (by negating the mean of the scores).
- (b) Plot the polynomial degree (1 through 6) on the x-axis and the average cross-validated MSE on the y-axis.
- (c) Based on the cross-validation MSE plot, which polynomial degree appears to be the best choice for modeling the relationship between Temperature (°C) and Quality Rating? Justify your choice, considering the trade-off between model performance (low MSE) and complexity.

Question 5: Final Model, Prediction, and Limitations

- (a) Based on the optimal degree identified for Temperature (°C) in Question 4, create the corresponding polynomial features using the *entire* dataset. Train a final LinearRegression model using these features.

- (b) Use this final model to predict the Quality Rating for a hypothetical process run with a Temperature (°C) of 215°C.
- (c) Briefly interpret the relationship between Temperature (°C) and Quality Rating as captured by your final polynomial model (consider the shape of the curve).
- (d) Discuss *two* potential limitations of using this final model (which is based *only* on Temperature (°C)) for predicting Quality Rating in a real manufacturing setting. Consider factors like other variables (e.g., Pressure (kPa), interaction terms), the range of data, and potential extrapolation issues.