## **P02a: Optimizing Regression Models for Disease Progression Prediction**

This lab exercise emphasizes parameter optimization, regularization, and hyperparameter tuning techniques to enhance the performance of regression models for disease progression prediction. Applying these techniques and understanding their impact on model generalization and robustness.

**Scenario:**

You are a data scientist working for a medical research institution. You are tasked with predicting the progression of a chronic disease based on various patient characteristics and medical history. The dataset you are working with is large and complex, with potentially non-linear relationships between variables. You want to build a model that can accurately predict the disease progression to assist in early diagnosis and personalized treatment planning. You will explore linear and polynomial regression models, focusing on optimizing their parameters to achieve the best possible performance.

**Dataset:**

For this exercise, you will use the synthetic dataset generated using NumPy. This dataset includes features such as age, blood pressure, cholesterol levels, and other relevant medical indicators, along with a target variable representing disease progression.

**Task 1: Data Preparation and Exploration**

1. **Import Libraries:** Import the necessary libraries for data manipulation, visualization, model building, and evaluation.
2. **Data Splitting:** Divide the dataset into training, validation, and testing sets. Explain the rationale behind this three-way split and your choice of splitting ratios.
3. **Exploratory Data Analysis:** Perform exploratory data analysis to understand the distribution of features, identify potential outliers, and visualize relationships between variables.

**Task 2: Model Building and Parameter Tuning**

1. **Linear Regression:** Build a linear regression model and explore techniques like feature scaling (e.g., standardization) to potentially improve its performance.
2. **Polynomial Regression:** Experiment with polynomial regression models, systematically adjusting the degree of the polynomial to find the optimal balance between complexity and generalization.
3. **Regularization:** Implement regularization techniques (e.g., Ridge, Lasso) with either linear or polynomial regression to mitigate overfitting and enhance model robustness. Explore different regularization strengths and analyze their impact on model performance.
4. **Hyperparameter Optimization:** Utilize techniques like Grid Search or Randomized Search to fine-tune hyperparameters for both linear and polynomial regression models, including regularization parameters and polynomial degree.

**Task 3: Model Evaluation and Selection**

1. **Comprehensive Evaluation:** Evaluate the performance of all models (linear, polynomial with varying degrees, and regularized models) using appropriate metrics such as Mean Squared Error (MSE), R-squared, and potentially others relevant to medical prediction. Compare performance across training, validation, and test sets.
2. **Learning Curves:** Visualize learning curves for selected models to assess their learning behavior and identify potential areas for improvement.
3. **Model Selection:** Based on the evaluation results, select the best-performing model for predicting disease progression. Justify your choice using evidence from the evaluation metrics and learning curves.

**Task 4: Model Interpretation and Reporting**

1. **Feature Importance:** Analyze the coefficients or feature importance scores of the selected model to understand the contribution of each patient feature to disease progression prediction.
2. **Clinical Implications:** Discuss the clinical implications of the findings, focusing on potential applications for early diagnosis and personalized treatment planning.
3. **Reporting:** Prepare a comprehensive report that includes details of the data preparation, model building, parameter tuning, evaluation, and interpretation steps. Highlight the key findings and their significance in the context of the medical scenario.

**Grading Rubric:**

* Correct implementation of data preparation, model building, parameter tuning, and evaluation steps.
* Thorough exploration of parameter adjustments and their impact on model performance.
* Clear and concise explanation of the rationale behind model selection and interpretation.
* Effective use of visualizations to present results and insights.
* Comprehensive discussion of the findings and their implications for disease prediction and personalized treatment planning.