#### 1. 问题描述

为研究形成性评价成绩与终结性考核成绩之间的关系，本文采用随机森林回归模型（Random Forest Regression）。假设数据集中的每一对数据（形成性评价成绩、终结性考核成绩）对应着一名学生的评分情况。目标是通过随机森林回归模型对形成性评价成绩（作为自变量）与终结性考核成绩（作为因变量）之间的非线性关系进行拟合与预测。

#### 2. 随机森林回归模型简介

随机森林回归是一种集成学习方法，它通过构建多个决策树并将它们的预测结果加权平均来进行回归预测。每棵树在训练过程中会选择数据的不同子集与特征，从而提高模型的泛化能力和鲁棒性。随机森林回归模型的主要优点是能够处理复杂的非线性关系并有效避免过拟合。

#### 3. 模型建立

设定随机森林回归模型的输入为学生的形成性评价成绩 XXX，输出为学生的终结性考核成绩 YYY。模型通过以下步骤建立：

**数据预处理**：将原始数据（包括形成性评价成绩和终结性考核成绩）整理为两列数据，分别表示自变量和因变量。形成性评价成绩作为自变量 XXX，终结性考核成绩作为因变量

**模型构建**：采用 sklearn.ensemble.RandomForestRegressor 模块建立随机森林回归模型，设定决策树的数量为 100（即 n\_estimators=100），随机种子为 42，以保证结果的可重复性。

**模型训练**：通过训练数据集训练模型，拟合形成性评价成绩与终结性考核成绩之间的非线性关系。

**预测与拟合**：使用训练好的模型对形成性评价成绩的不同值进行预测，得到对应的终结性考核成绩预测值，构建回归曲线。

#### 4. 模型求解

模型训练完成后，通过计算皮尔逊相关系数（rrr）与均方误差（MSE）评估模型的拟合效果：

**皮尔逊相关系数**：皮尔逊相关系数衡量的是两个变量之间的线性相关性，其计算公式如下：

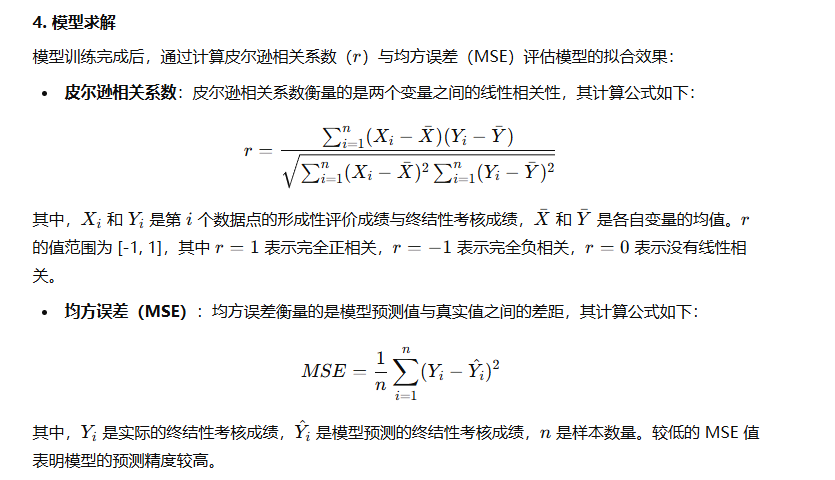
r=∑i=1n(Xi−Xˉ)(Yi−Yˉ)∑i=1n(Xi−Xˉ)2∑i=1n(Yi−Yˉ)2r = \frac{\sum\_{i=1}^{n} (X\_i - \bar{X})(Y\_i - \bar{Y})}{\sqrt{\sum\_{i=1}^{n} (X\_i - \bar{X})^2 \sum\_{i=1}^{n} (Y\_i - \bar{Y})^2}}r=∑i=1n​(Xi​−Xˉ)2∑i=1n​(Yi​−Yˉ)2​∑i=1n​(Xi​−Xˉ)(Yi​−Yˉ)​

其中，XiX\_iXi​ 和 YiY\_iYi​ 是第 iii 个数据点的形成性评价成绩与终结性考核成绩，Xˉ\bar{X}Xˉ 和 Yˉ\bar{Y}Yˉ 是各自变量的均值。rrr 的值范围为 [-1, 1]，其中 r=1r = 1r=1 表示完全正相关，r=−1r = -1r=−1 表示完全负相关，r=0r = 0r=0 表示没有线性相关。

**均方误差（MSE）**：均方误差衡量的是模型预测值与真实值之间的差距，其计算公式如下：

MSE=1n∑i=1n(Yi−Yi^)2MSE = \frac{1}{n} \sum\_{i=1}^{n} (Y\_i - \hat{Y\_i})^2MSE=n1​i=1∑n​(Yi​−Yi​^​)2

其中，YiY\_iYi​ 是实际的终结性考核成绩，Yi^\hat{Y\_i}Yi​^​ 是模型预测的终结性考核成绩，nnn 是样本数量。较低的 MSE 值表明模型的预测精度较高。



#### 6. 结果与分析

模型训练后，得到皮尔逊相关系数 r=0.764，表明形成性评价成绩与终结性考核成绩之间有较强的正相关关系；同时，均方误差为 MSE=76.23，说明模型的预测误差较小，拟合效果较好。

通过图1中的散点图与拟合曲线，可以清晰看到数据的非线性关系，随机森林模型较好地捕捉到了这一趋势。在低分段，形成性评价成绩与终结性考核成绩之间的关系较为平稳；而在高分段，随着形成性评价成绩的提高，终结性考核成绩的增速趋于平缓。