# 从建模目标、建模算法和评估方面，阐述机器学习建模与传统统计模型有哪些不同？

**1. 建模目标**

**机器学习建模**:

主要关注预测性能，强调模型在未见数据上的泛化能力。目标常常是最大化模型的预测准确性或其他性能指标（如F1分数、AUC等）。可以用于复杂的高维数据问题，适用于需要自动化学习和适应性强的场景。

**传统统计模型**:

主要关注参数估计和推断，强调对数据背后关系的理解。着重于建立一个可解释的模型，通常试图揭示自变量与因变量之间的因果关系。某些情况下，模型更多的用于假设检验，而非纯粹的预测。

**2. 建模算法**

**机器学习建模**:

使用复杂的算法，如决策树、随机森林、支持向量机、神经网络等。侧重于大量数据和特征的处理，能够自动检测特征之间的复杂关系。会利用集成学习、深度学习等新兴技术，适应不同的数据环境。

**传统统计模型**:

往往使用线性回归、逻辑回归、时间序列分析等较为简单的模型。这些模型通常假设数据符合某种分布（如正态分布）并且遵循特定的条件。强调模型的简洁性和可解释性。

**3. 评估方面**

**机器学习建模**:

使用交叉验证、混淆矩阵、ROC曲线等方法评估模型的预测性能。注重于模型在训练集之外的数据表现，常用A/B测试等方法检验模型的实际效果。可能更有关注模型的不确定性和鲁棒性，通过调参、正则化等手段提升模型稳定性。

**传统统计模型**:

评估方法常包括R方、p值、残差分析等，更多关注模型的显著性和适用性。对模型的适合度、假设检验等有明确的统计标准。评估结果需要结合背景知识进行解读，以确保模型的实际应用合理性。

## 对于doubs中的鱼群数据，按照样地，计算各样地鱼类Shannon多样性指数，并新增mpg列。

# 安装并加载必要的包

install.packages(c("ade4", "vegan"))

library(ade4)

library(vegan)

# 加载Doubs数据集

data(doubs)

# 计算Shannon多样性指数

shannon\_diversity <- diversity(doubs$fish, index = "shannon")

# 创建新的数据框

doubs\_diversity <- data.frame(

site = rownames(doubs$fish),

shannon\_index = shannon\_diversity,

mpg = NA

)

## 利用train()，训练随机森林（randomForest）模型

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf \_\_")

## 通过trainControl()，向train()添加重采样10-fold cross-validation，以优化参数

fitControl <- trainControl(method = " repeatedcv ", number = 10, repeats = 5)

model\_rf <- train(mpg ~ ., data = training\_data, method = "rf ", trControl =fitControl)

## 在train()中，增加中心化和标准化等数据预处理，提高模型精度

model\_rf <- train(mpg ~ ., data = training\_data, method = " rf ",

preProcess = c('scale', 'center'),

trControl =fitControl)

## rf有mtry和tree两个参数，可以通过expand.grid()设置调优，并在train()添加

grid <- expand.grid(.mtry=c(1:10))

model\_rf <- train(mpg ~ ., data = training\_data, method = "rf",

preProcess = c('scale', 'center'),

trControl =fitControl,

\_ tuneGrid\_\_ = \_grid\_)

# 什么是递归消除选择？在caret包中，为何选择随机森林等树模型时，没有特征选择这个过程？

递归特征消除（RFE）是一种迭代的特征选择方法，其核心目标是从完整的特征集中逐步筛选出对模型预测性能最重要的特征子集。

而在随机森林等树模型中，看似没有显式的特征选择过程，实际上是因为这些模型天然具有内置的特征选择机制。在决策树的每个节点，算法会基于信息增益或基尼系数选择能最大程度降低不确定性的特征，通过评估特征对模型预测能力的贡献来确定其重要性。随机森林更进一步，通过随机选择特征子集、构建多棵决策树并集成，不仅能有效减少单个特征的过度拟合影响，还能自动捕捉特征间的非线性关系和复杂交互效应。这种隐式的特征选择方法相较于传统的显式特征选择技术，具有处理复杂数据模式、不需要严格线性假设、计算效率高等优势，特别适合生态学等需要处理复杂非线性关系的研究领域。在实际应用中，研究者可以通过varImp()函数直观地评估特征重要性，并结合领域专业知识对模型进行解释和优化。