## 《综合项目实践》任务书

**课程题目：**基于机器学习和极化雷达数据的小时降雨量估测

**课程目标：**

**整体目标**：能够综合利用多种机器学习技术，根据极化雷达数据估测每小时降雨量。

**具体目标**如下：

* 数据分析：了解数据集，进行数据统计分析和探索性数据分析。
* 特征工程：数据清洗（异常值处理、缺失值处理）、特征归一化、特征聚合和特征选择。
* 选择5种机器学习/深度学习算法实现小时降雨量估测。
* 模型训练和评估。
* 比较不同模型的性能，并分析它们的优缺点。
* 研究不同特征工程方法对模型的影响。
* 使用集成学习中的Stacking和贝叶斯模型平均法（BMA）将5种不同模型的估测结果进行组合，从而提高降雨量估测的精度。

**算法技能目标：**能够掌握相关机器学习、深度学习和集成学习算法。

**编程技能目标：**能够熟练使用python编程语言进行数据的处理、分析、建模和评估。

**思政目标：**学好机器学习技术，更好地服务于国家和人民。

**任务流程：**

一、**数据分析：（1次课）**

~~1、 通过给定数据的网址，仔细阅读相关数据说明，理解所用数据集的格式和内容~~。

https://www.kaggle.com/competitions/how-much-did-it-rain-ii/overview#

https://www.kaggle.com/competitions/how-much-did-it-rain-ii/data

2、~~对数据进行统计分析和探索性数据分析，并进行可视化，深入了解数据分布、缺失、异常及相关性等情况~~。

<https://mp.weixin.qq.com/s/V9W0nZzqPdvpp2IGjJRb5w>

https://mp.weixin.qq.com/s/9lxM8U5OqRKVodzE1Q5FCQ

二、**特征工程（2次课）**

1、~~数据清洗：对数据进行缺失值处理和异常值处理，并给出处理前后的可视化对比图。~~

~~总结常用的缺失值处理方法，并实现2种缺失值处理方法~~。

https://www.cnblogs.com/jamchen2021/p/14821584.html

https://mp.weixin.qq.com/s/\_OZmMZMUpo0MNzZ32KR-7w

https://mp.weixin.qq.com/s/eZAweJ9QsiWgKmuO9RbnyA

总结常用的异常值的判别和处理方法，并实现2种异常值判别和2种异常值处理方法。

https://mp.weixin.qq.com/s/YRG\_H8FL4FLBHUSejRafsw

https://mp.weixin.qq.com/s/qsXBHyNdiDX9Q1WP8g5Iag

2、~~数据格式转换：将现有数据整理成为时间序列预测模型所需的时间序列格式。~~

3、~~特征归一化：总结常用的特征归一化方法，并实现最小-最大归一化、Z-score标准化方法。~~

4、特征聚合（特征衍生—分组聚合）：总结常用的特征聚合方法，并实现其中2种方法。

5、特征选择：总结常用的特征选择方法，并实现其中1种方法。

**三、算法选择与实现（2次课）**

1、调研常用的机器学习回归算法，并实现其中4种算法。

2、调研常用的深度学习时间序列预测算法，并实现其中1种算法。

**四、模型训练和评估（4次课）**

1、利用K折交叉验证对数据集进行划分。

2、确定各个模型的参数、超参数、损失函数和优化算法等，并训练模型，模型的训练过程需要进行可视化。

3、考虑并解决可能存在的模型过拟合和欠拟合问题。

4、从模型运行时间和模型精度等方面对模型进行评估。

**五、特征工程对模型效果的影响（2次课）**

1、不同缺失值处理方法对模型效果的影响。

2、不同特征归一化方法对模型效果的影响。

3、不同特征聚合方法对模型效果的影响。

**六、基于集成学习的降雨量估测（2次课）**

1、基于Stacking的降雨量估测

2、基于BMA的降雨量估测（加分项，选做题）