**基于天气自动传感器大数据(Beach Water Quality)数据集对海洋能开发进行分析和预测**

**李浩铭 郑嘉文 尹壹 梁陈俊**

1. **数据说明**

这个数据集统计了在芝加哥的若干个海滩中传感器每日测试的数据，这些数据包括日期、水温、浑浊度、换能器深度(m)、波浪高度(m)、波浪周期(s)、传感器电池剩余情况这些属性。一共34924条数据。

任务1：多分类。每个海滩的属性有差异，可以根据这些差异训练分类器完成对不同海滩的分类操作。

任务2：序列预测。对于某个沙滩，其水温、波浪高度、周期等属性可能会发生周期性变化。可以根据过去几年观测到的数据预测接下来一段时间的情况。

任务3：聚类。给定测试数据，将其聚类，看看每一个类是否对应于一个海滩。

任务4：数据预处理。

1. **项目目标**

根据所给数据集中，分别计算分析得到不同海洋能所对应的最佳开发海滩位置。

1. **项目步骤**
2. 对数据集进行数据预处理，根据海滩名对各海滩进行分类，并按时间顺序排序。同时，将海滩的时间数据计算转换。
3. 输出处理后的数据集“processed\_data.csv”
4. 加载数据集，检查数据。
5. 定义不同海洋能类型及相关特征（所需数据集中的数据列）。
6. 标准化特征。
7. 假设海洋能潜力与相关的特征数据呈现正相关，利用简单回归模型分析。
8. 输出结果
9. **项目内容**

“data\_preprocessing.py”对数据集进行预处理

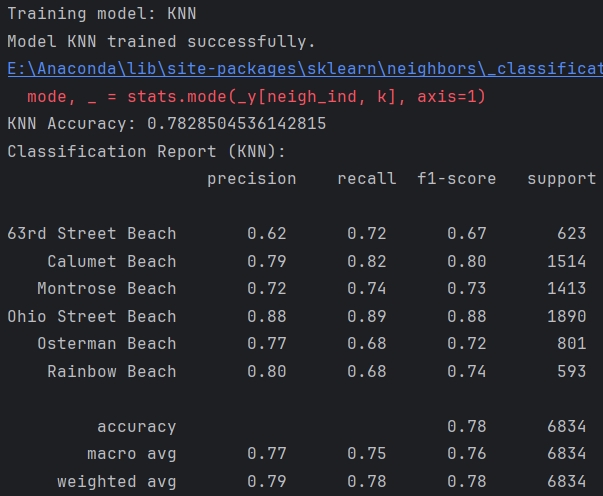
“data\_clustering.py”对数据集进行多分类任务

“data\_multiple\_classification”对数据集进行聚类任务

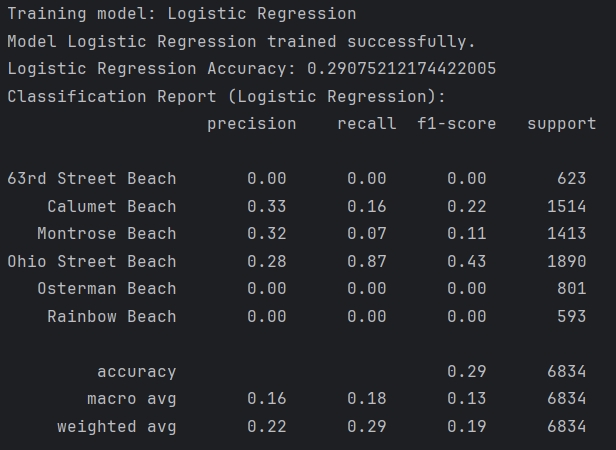
“data\_series\_prediction.py”对数据集进行序列预测任务

“Ocean\_Energe\_Development.py”为项目实现的最终代码

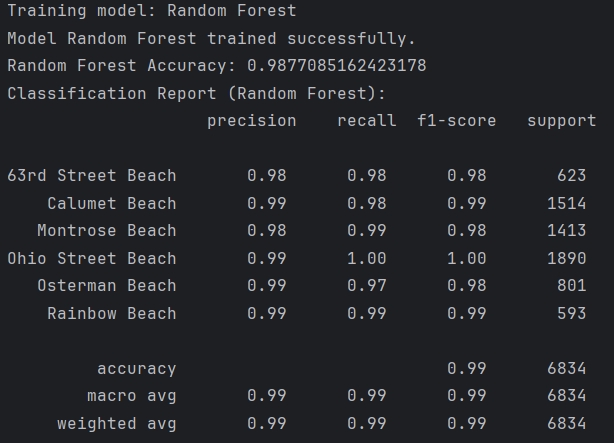
1. **项目结果**
2. 多分类模型结果：

****

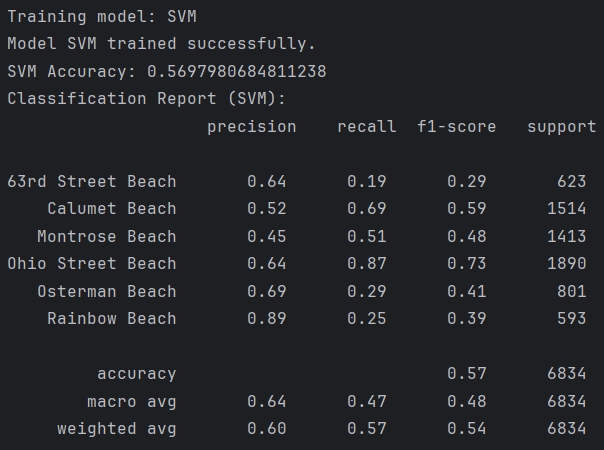
*KNN*

****

*Logistic Regression*

****

*Random Forest*

****

*SVM*

1. 聚类模型结果：

****

*K-Means（左）和DBSCAN（右）*

1. 序列预测模型误差结果：



*Linear Regression*



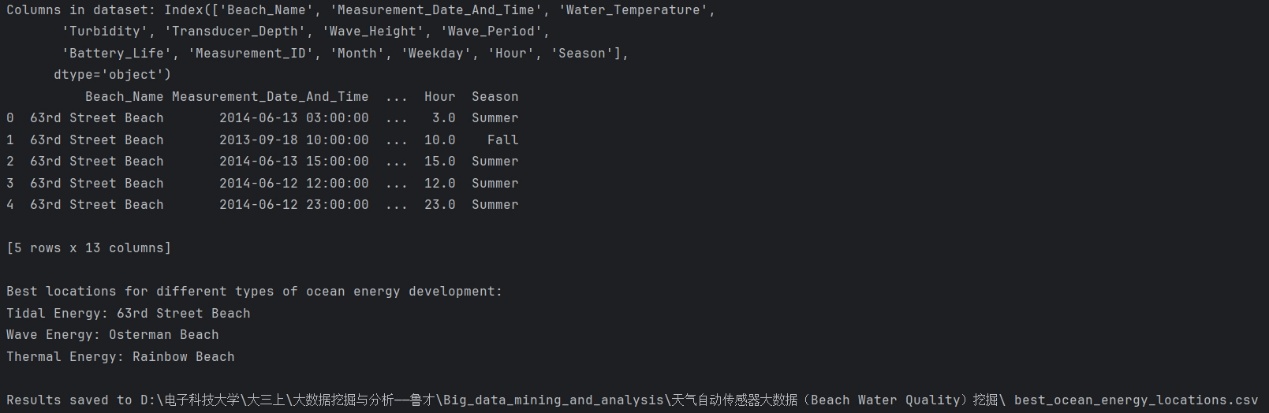


*LSTM*



*Support Vector Regression*

1. 海洋能最佳开发位置：

****

*海洋能最佳开发位置分析输出*

*各海洋能最佳开发位置对应海滩：*

*潮汐能 —— 63rd Street Beach*

*波浪能 —— Osterman Beach*

*温差能 —— Rainbow Beach*

1. **项目开源地址**

[**KapokInHometown/Big\_data\_mining\_and\_analysis: 大数据挖掘与分析**](https://github.com/KapokInHometown/Big_data_mining_and_analysis)

1. **小组成员分工**

李浩铭：项目目标及内容构思、代码的实现与调试、实验最终报告撰写

郑嘉文、尹壹：汇报PPT制作及汇报

梁陈俊：汇报PPT制作