研究生课程项目报告

（适用于课程论文、提交报告）

分数： 评语： 评阅人（签字）：

科　目：大数据分析与挖掘 **教　师：**向朝参、欧阳德强

姓　名：王颖 **学　号：**20213917

专　业：计算机科学与技术 **类　别：**硕士

上课时间：2024 年 11 月 至 2025 年 1 月

重庆大学研究生院制

基于全球能源数据的可再生能源应用影响因素探究

摘要: 在全球能源危机的背景下，可再生能源的应用和推广在如今的世界日趋重要。可再生能源的应用受到复杂的因素影响，包括技术、政策、经济、地域、地理位置等各方面因素。分析其不同的影响因素可以为可再生能源的应用与推广做出更加有效的、有针对性应用方式与策略，推动能源结构转型。本文基于 Kaggle 网站的公开数据集“Global Renewable Energy Usage（2020-2024）”，尝试探究数据集中影响月可再生能源使用量“Monthly_Usage_kWh”的关键因素。本文通过数据分析研究相关变量的相关性，进一步依据相关性分析的结果建立随机森林回归模型，分析影响可再生能源使用的关键因素。

关键词: *可再生能源, 大数据分析, 能源应用, 随机森林*

1. 引言

在全球气候危机的背景下，能源问题成为世界各国亟待解决的重要议题。化石燃料的广泛使用不仅造成了温室气体的大量排放，还加剧了能源安全问题。因此，发展和推广可再生能源被认为是实现可持续发展的关键路径之一。太阳能、风能、水能等清洁能源因其环保、可再生的特性，正在逐步取代传统能源，成为全球能源结构的重要组成部分。

然而，可再生能源的推广与使用仍面临诸多挑战。社会经济因素（如收入水平、补贴政策）、技术条件（如能源转换效率、储能技术）以及地理环境（如区域气候、基础设施建设）均对其使用量产生重要影响。理解这些影响因素是推动可再生能源应用的关键。

2. 基于随机森林回归模型的多因素分析算法

2.1 算法概述

随机森林（Random Forest）是一种基于集成学习的机器学习算法，被广泛用于回归问题。它通过使用多个决策树对数据进行建模，并将它们的预测结果进行集成，从而提高了模型的性能和稳定性。

2.2 算法原理

随机森林回归的基本原理和流程如下：

(1) 随机选择样本：从原始训练集中随机选择一部分样本，构成一个子样本集。这样可以使得每棵决策树都在不同的样本集上进行训练，从而增加模型的多样性。

(2) 随机选择特征：对于每个决策树的每个节点，在选择最佳划分特征时，只考虑随机选择的一部分特征。这样可以防止某些特征对整个模型的影响过大，从而提高模型的鲁棒性。

(3) 构建决策树：在每个子样本集上使用某种决策树算法（如 CART 算法）构建一棵决策树。决策树的生长过程中，通常采用递归地选择最佳划分特征，将数据集划分为不纯度最小的子集。

(4) 集成预测：对于新的输入样本，通过将多棵决策树的预测结果进行平均或加权平均，从而得到最终的回归结果。

2.3 算法的使用

借助 Python 中 Scikit-learn 库的 RandomForestRegressor 类可以较为快速、方便地构建随机森林回归模型，并且对其进行训练和结果评估，本文即采用该模块来进行主要的模型构建与训练等工作。

3. 数据集

本文采用数据集为 Kaggle 网站上的公开数据集“Global Renewable Energy Usage（2020-2024）”，该数据集提供了 2020 年至 2024 年全球家庭可再生能源使用情况的信息。它包括可再生能源采用情况、月度能源使用量、家庭规模、收入水平和城乡差异的数据。该数据集还包含政府补贴以及家庭采用太阳能、风能、水能和地热能等可再生能源所节省的成本详情。这是一个了解可再生能源采用趋势、分析影响能源使用的因素以及探索社会经济特征与能源消耗之间关系的有用资源。

数据集的维度为 1000×12，数据集的样例如表 3-1 所示：

Household_ID	Region	Country	Energy_Source	Monthly_Usage_kWh	Year	Household_Size	Income_Level	Urban_Rural	Adoption_Year	Subsidy_Received	Cost_Savings_USD
H00706	Africa	Egypt	Hydro	274.46	2022	7	Middle	Rural	2010	No	65.98
H00107	South America	Brazil	Biomass	375.99	2022	2	High	Urban	2011	No	20.93

表 3-1 数据集样例表

使用该数据集需要在 Python 代码中导入 Pandas 库，使用其中的函数 read_csv(file_path) 读入 csv 文件即可。

4. 实验

在实验阶段主要进行两部分工作，一是进行相关性分析，大致分析各个变量与可再生能源使用量之间的相关性，探究主要分析的方向；二是建立随机森林回归模型进行量化分析，研究影响可再生能源使用量的主要影响因素。

4.1 相关性分析

将特征分类，分为数值型特征与离散型特征。对数值型特征，将其与可再生能源使用量分别进行相关性分析并可视化结果；对离散型特征，对其进行特征转换，通过量化编码使其便于进行相关性分析。同时，增加一项新的特征“使用年数”，由“统计年”-“采用年”得到。特征进行一定的转换之后再分别与可再生

能源使用量进行相关性分析。

相关性分析的结果通过可视化的相关图表与相关性矩阵来评估其结果好坏与相关性大小。

4.2 建立随机森林回归模型分析

借助相关性分析中已经转换完成的特征，首先划分目标值与特征值，其中特征分为多个维度，包括家庭规模、节省金额、采用年限、是否补贴、收入水平、地区等，目标值为月可再生能源使用量，即数据集中的“ Monthly_Usage_kWh ”这一维度；接下来将数据集按照 8:2 的比例划分为训练集和测试集，通过 Scikit-learn 库的 RandomForestRegressor 类构建随机森林回归模型并指定参数，对其进行训练和结果评估。

模型的量化评估指标主要为平均平方误差 MSE、平均绝对误差 MAE、R^2 得分（决定系数）、交叉验证 R^2 得分与平均交叉验证 R^2 得分五项指标，分别衡量模型的误差大小与模型对数据变化的解释能力。

4.3 实验结果分析

在相关性分析阶段，得到各特征之间的相关性矩阵如图 4-1。由相关性矩阵可以看出，各特征之间的相关性并不明显，相对来说，与（可再生能源）月使用量“Monthly_Usage_kWh”这一特征相关性较大的特征是家庭规模“House_hold”、（家庭）收入水平“Income_Level”与采用年限“Years_Since_Adoption”。其中前两者为负相关，第三者为正相关。

相关性矩阵：

	Monthly_Usage_kWh	Household_Size	Cost_Savings_USD	Years_Since_Adoption	Subsidy_Received	Income_Level	Region
Monthly_Usage_kWh	1.000000	-0.024876	0.000437	0.010724	-0.000032	-0.060829	-0.006430
Household_Size	-0.024876	1.000000	-0.032560	-0.049226	-0.004842	0.048577	-0.033540
Cost_Savings_USD	0.000437	-0.032560	1.000000	0.018023	-0.021612	-0.029430	-0.063868
Years_Since_Adoption	0.010724	-0.049226	0.018023	1.000000	0.059953	-0.005708	-0.018304
Subsidy_Received	-0.000032	-0.004842	-0.021612	0.059953	1.000000	-0.009852	-0.043347
Income_Level	-0.060829	0.048577	-0.029430	-0.005708	-0.009852	1.000000	-0.076069
Region	-0.006430	-0.033540	-0.063868	-0.018304	-0.043347	-0.076069	1.000000

图 4-1 相关性矩阵

在建模分析阶段，可得到五种量化指标值分别如图 4-2。其中与模型误差相关的 MSE、MAE 指标的数值都偏大，而与模型对数据变化的解释能力相关的 R^2、CV R^2 与 MCV R^2 指标值都偏小，模型在此情景下的性能表现较差。

```
Mean Squared Error (MSE): 200279.44250248582
Mean Absolute Error (MAE): 387.156121
R² Score: -0.09028118922283013
Cross-Validation R² Scores: [-0.07599936 -0.13013265 -0.1231749  -0.02965206 -0.03844313]
Mean CV R² Score: -0.0794804189143076
```

图 4-2 五种性能量化指标

5. 总结与讨论

通过对数据的探索性分析，我们发现部分变量与家庭可再生能源使用量之间的相关性较弱。这表明单一因素可能不足以解释能源使用量的差异，而需要更多复杂的交互作用或其他隐藏变量的补充分析。

通过查阅相关资料分析，实验结果不理想的原因可能有以下三方面：

(1) 因素复杂性：低相关性可能反映了整体趋势的复杂性，可能需要通过更细粒度的分析来挖掘隐含关系。

(2) 数据集质量：变量间的相关性较弱可能与数据量、数据采集范围和变量设计有关。例如，未包含更细粒度的经济、政策或技术因素可能是研究结果的一大局限性。这一结果提示我们需要更全面的数据集，涵盖更广泛的影响因素以解释能源使用的复杂性。同时，数据在采集过程中是否出现偏颇性也是需要考量的问题。

(3) 模型适配性：随机森林虽然能捕捉非线性关系，但在特征过少或噪声过大的情况下，仍可能表现不佳。模型可能在训练数据上学习不足，或者模型复杂度未能适应数据的复杂性。

研究过程中发现，部分变量的影响可能较弱，实验结果虽不太理想，但这为后续进一步挖掘潜在机制提供了重要启示。

参考文献

[1] 胡欢武.机器学习基础[M].电子工业出版社:201904.384..

[2] 李英冰,张岩.应急大数据的空间分析与多因素关联挖掘[M].武汉大学出版社:202106.191.

[3] 张松慧,陈丹.机器学习 Python 实战[M].人民邮电出版社:202209.180.

[4] 伍中信,刘思菡.基于随机森林回归分析的上市公司碳排放效应研究[J].湖南财政经济学院学报,2024,40(04):27-38.DOI:10.16546/j.cnki.cn43-1510/f.2024.04.003.

[5] 曹馨,吴琪,王萱,等.基于随机森林算法的中长期天然气需求影响因素分析[J].煤炭经济研究,2024,44(03):6-14.DOI:10.13202/j.cnki.cer.2024.03.022.

[6] 曹军威, 袁仲达, 明阳阳, 等. 能源互联网大数据分析技术综述[J]. 南方电网技术, 2015, 9(11): 1-12.