Sustainable Energy 可持续能源, 2023, 13(3), 33-43
Published Online June 2023 in Hans. https://www.hanspub.org/journal/se
https://doi.org/10.12677/se.2023.133004

Hans汉斯

基于机器学习的煤炭资源利用优化策略

——以陕北地区为例

何依婷[1]，谢殷豪[2]，张彤言[3]

[1]延安大学数学与计算机科学学院，陕西 延安
[2]榆林学院化学与化工学院，陕西 榆林
[3]咸阳师范学院化学与化工学院，陕西 咸阳

收稿日期：2023年9月2日；录用日期：2023年10月2日；发布日期：2023年10月11日

摘　要

陕北地区煤炭资源丰富，是我国重要的能源化工基地。为了寻找合理的煤炭资源利用策略，实现资源的可持续发展，本文根据大量陕北地区的煤炭资源数据，包括陕北地区煤炭使用量和浪费量，并使用机器学习算法进行回归分析。首先，对数据进行预处理，去除无用信息。然后，使用五类回归模型分析煤炭使用量数据，选出最佳模型对煤炭浪费量进行准确预测。最后，制定出符合可持续发展原则的策略，减少不必要的资源浪费和环境污染，提高煤炭资源的综合利用效率。

关键词

可持续发展，机器学习，煤炭资源，回归分析

Optimization Strategy for Coal Resource Utilization Based on Machine Learning

—Taking the Northern Shaanxi Region as an Example

Yiting He[1], Yinhao Xie[2], Tongyan Zhang[3]

[1]School of Mathematics and Computer Science, Yan'an University, Yan'an Shaanxi
[2]School of Chemistry and Chemical Engineering, Yulin University, Yulin Shaanxi
[3]School of Chemistry and Chemical Engineering, Xianyang Normal University, Xianyang Shaanxi

Received: Sep. 2nd, 2023; accepted: Oct. 2nd, 2023; published: Oct. 11th, 2023

Abstract

The northern Shaanxi region is rich in coal resources and is an important energy and chemical

文章引用: 何依婷, 谢殷豪, 张彤言. 基于机器学习的煤炭资源利用优化策略[J]. 可持续能源, 2023, 13(3): 33-43.
DOI: 10.12677/se.2023.133004

industry base in China. In order to find reasonable coal resource utilization strategies and achieve sustainable development of resources, this article is based on a large amount of coal resource data in the northern Shaanxi region, including coal usage and waste, and uses machine learning algorithms for regression analysis. Firstly, preprocess the data to remove useless information. Then, use five types of regression models to analyze coal usage data and select the best model to accurately predict coal waste. Finally, develop strategies that comply with the principles of sustainable development, reduce unnecessary resource waste and environmental pollution, and improve the comprehensive utilization efficiency of coal resources.

Keywords

Sustainable Development, Machine Learning, Coal Resources, Regressive Analysis

1. 引言

1.1. 背景与研究动机

当今社会，煤炭资源是世界能源产业的重要组成部分。然而，随着全球可持续发展议程的不断推进，煤炭资源的开采和利用方式面临着日益严峻的挑战。煤炭资源的不合理开采和过度利用已经导致了环境污染、生态破坏、资源浪费以及社会问题。因此，迫切需要采取措施来改善煤炭资源的利用策略，以实现可持续发展目标。本文的研究动机有以下四个方面。

1) 可持续发展需求。随着全球对环境保护和可持续发展的关注度不断提高，各国政府都在积极推动低碳经济的发展，以实现经济、社会和环境的协调发展。中国政府也明确了实现低碳经济和环境友好型发展的目标，并为此采取了一系列措施，以减少对环境的破坏和资源的浪费。因此需要针对这一问题开展深入研究，为实现可持续发展提供有力支撑。

2) 机器学习技术的崛起。近年来，随着大数据和人工智能技术的不断发展，机器学习技术在各个领域的应用越来越广泛。机器学习技术在数据分析和预测领域表现出了显著的优势，可以帮助人们更好地把握市场变化和资源需求，进而为企业和政府决策提供更加准确的数据支持。针对煤炭资源的利用问题，机器学习技术的应用可以帮助我们更好地预测煤炭市场波动、资源需求以及环境影响等方面的情况，为优化煤炭资源的利用策略提供了新的可能性。

3) 陕北地区煤炭资源的特殊性。陕北地区是我国煤炭资源最为丰富的地区之一，也是我国能源的重要基地。然而，该地区面临着许多特殊的问题，如环境脆弱、资源枯竭等。在煤炭资源利用方面，需要采取科学合理的方式，既要充分发挥煤炭资源的优势，促进地区经济发展，又要积极探索新的技术和方法，减少对环境的破坏和资源的浪费。因此，针对陕北地区煤炭资源的科学管理对于该地区的可持续发展至关重要，同时也为其他煤炭资源富集地区提供了有价值的经验。

4) 政策制定的需求。政府和决策者在进行煤炭资源管理和政策制定时需要科学、准确的数据支持。通过对历史数据的分析，以及未来市场和环境变化的预测，可以为政府部门提供更加准确、实时的决策支持。机器学习技术的应用可以帮助我们更好地分析数据，把握市场和环境的变化趋势，进而为政府制定更加科学合理的政策和措施提供有力支撑。

1.2. 煤炭资源利用现状分析

煤炭作为世界三大工业能源之一，其开采量与使用量逐年递增。我国虽然地大物博，资源储量大，但在长期不合理开采下，煤炭所带来的高污染、高浪费问题日益突显。现有的煤炭资源开采技术有“长壁开采 121 工法”，这种开采体系不仅造成严重资源浪费，开采率不足 50%，煤炭资源浪费量达到 20%至 25%，煤炭资源每年的浪费量高达 34.5 亿多吨，价值 2 万多亿元，这种方法还带来安全隐患，破坏原有的生态环境，造成严重的环境污染。为解决此类问题，需要合理设置开发条件，提高煤炭资源利用率。我国西部地区煤炭资源丰富，开发时间早，开采条件优越。陕西省煤炭资源发布见图 1。

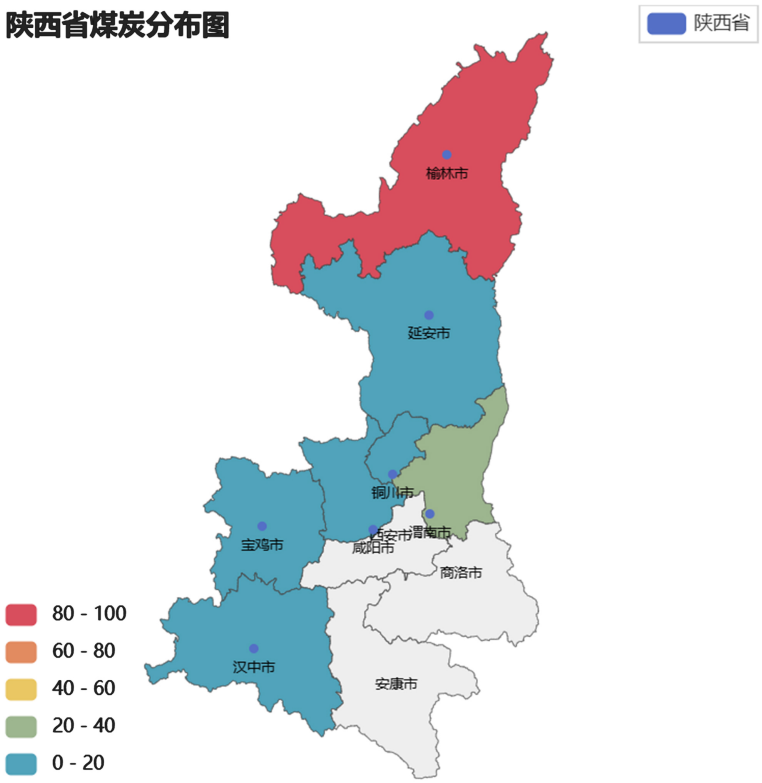

Figure 1. Coal distribution map of Shaanxi Province
图 1. 陕西省煤炭分布地图

伴随着经济的高速发展，煤炭开采技术的不断进步，陕北地区煤炭产业链面临巨大挑战。在煤炭资源合理利用方面，马洁琼[1]提出了煤炭经济可持续发展的相关对策，希望能为推动我国煤炭经济的转型发展提供有益参考；黄晨玮[2]通过 ARIMA 模型对陕煤集团未来 3 年的可持续发展能力作预测，对拓宽评价思路、创新评价方法具有重要价值；张晶[3]围绕可持续发展理念下的煤炭经济发展问题进行探讨，提出煤炭经济发展中存在的问题，并对基于问题明确具体的可持续发展路径；宋思远[4]基于改进不平衡数据的 Stacking 模型进行信用风险的预测研究，并将其应用到实际煤炭企业数据中；周相团等[5]分析了我国煤炭经济发展现状及基于可持续发展视角下所存在的问题，并提出了我国煤炭经济可持续发展的相关解决措施。

本文通过分析各城市煤炭数量和总产能(见图 2)和陕西省原煤分月产量及增速(见图 3)，针对陕西省的煤炭资源开采与利用提出了更为精准的优化预测方案，并对陕北地区的资源进行精准开采与合理利用，使得资源利用率大大提高，为能源安全上了一份保险。通过考虑各种因素引起的资源浪费，进一步细化

了能源使用方案，为煤炭工业的发展带来了一个光明的前景。

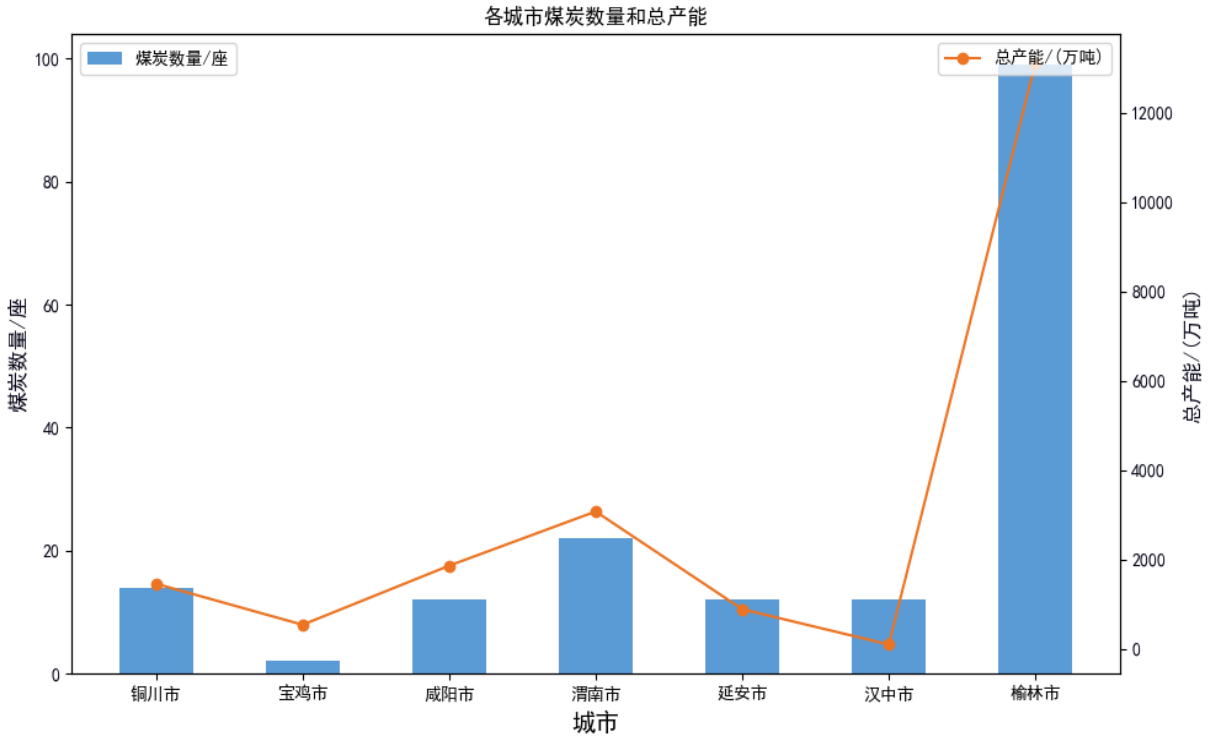

Figure 2. Quantity and total production capacity of coal in each city
图 2. 各城市煤炭数量和总产能

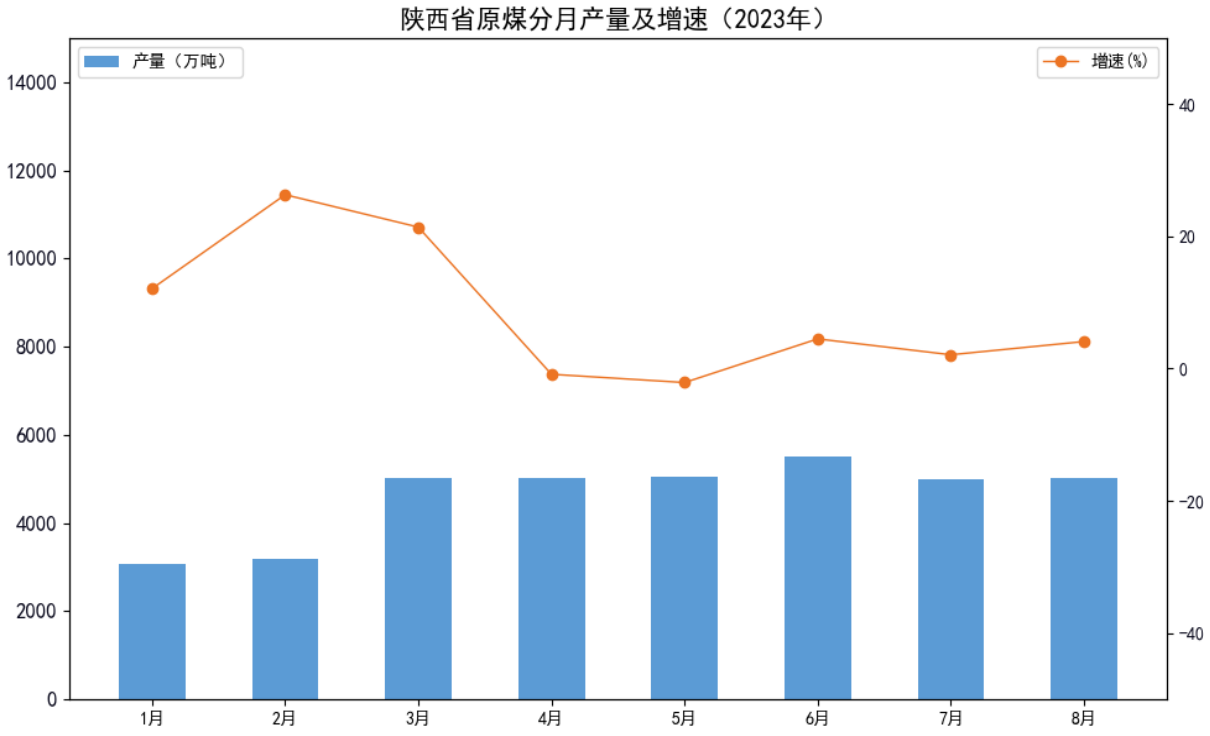

Figure 3. Monthly production and growth rate of raw coal in Shaanxi Province
图 3. 陕西省原煤分月产量及增速

2. 实验设计

2.1. 数据收集与预处理

2.1.1. 数据获取

陕北地区煤炭使用量和浪费量的数据通过当地的煤炭企业和政府部门进行获取，同时查阅了中国煤炭市场等权威数据发布机构的公开信息。这些资源可以帮助我们了解煤炭行业的整体趋势和发展动态，进而对陕北地区的煤炭使用和浪费情况进行更为全面和深入的了解。

2.1.2. 数据归一化

数据归一化是一种重要的数据预处理技术，它将具有不同尺度和范围的数据转化为统一的尺度，以便更好地进行比较和分析。有助于消除数据之间的量纲差异，提高了数据的可比性。通过数据归一化，能够更清晰地理解陕北地区煤炭数据之间的关系，从而提高模型稳定性。这种处理方法使不同特征具有相似的尺度，有效避免了尺度差异可能引发的偏差问题。

2.1.3. 数据清洗

煤炭数据清洗是数据预处理的关键环节，旨在识别和纠正数据集中的错误、缺失、重复或异常数据，以确保数据的准确性、一致性和完整性。包括去除重复数据、填补缺失值、处理异常值、数据类型转换和标准化数据。数据清洗有助于提高数据质量，使其适用于后续的聚类分析、建模和决策制定。具体描述见表 1。

Table 1. Coal data cleaning
表 1. 煤炭数据清洗

数据清洗操作	描述	处理数量
去除重复数据	检查并去除数据集中的重复数据记录	108
填补缺失值	检查缺失值情况，对于缺失的字段使用平均值/中位数/众数进行填充	20
处理异常值	检查异常值情况，遇到异常值直接删除	51
数据类型转换	确保数据字段的类型正确，将日期字段转换为日期格式	0
标准化数据	对数值型字段进行标准化，确保数据在相同的尺度和范围内	17

2.2. 机器学习模型选择与建立

线性回归[6] (Linear Regression)是一种基本的监督学习模型，用于建立输入特征与连续目标变量之间的线性关系。该模型通过拟合一条直线(在简单线性回归中)或一个超平面(在多元线性回归中)来最小化实际观测值与模型预测值之间的差异。线性回归的目标是找到最佳拟合系数，使平方误差最小化。模型结构图见图 4。

决策树[7] (Decision Tree)是一种非参数的监督学习模型，它广泛应用于分类和回归任务。与其他机器学习模型相比，决策树具有直观易懂、易于解释等优点。决策树的基本原理是构建一棵树状结构，以树根为起点，通过各个分支将数据集逐步划分为不同的子集，最终到达各个叶节点。每个叶节点代表一个类别(对于分类任务)或一个数值(对于回归任务)。在构建决策树的过程中，每个分支都基于一个特征进行划分。选择哪个特征进行划分通常由信息增益、基尼不纯度等指标来衡量。通过选择最优特征，可以将数据集划分成更纯的子集，从而使得每个子集更加容易进行分类或回归。模型结构图见图 5。

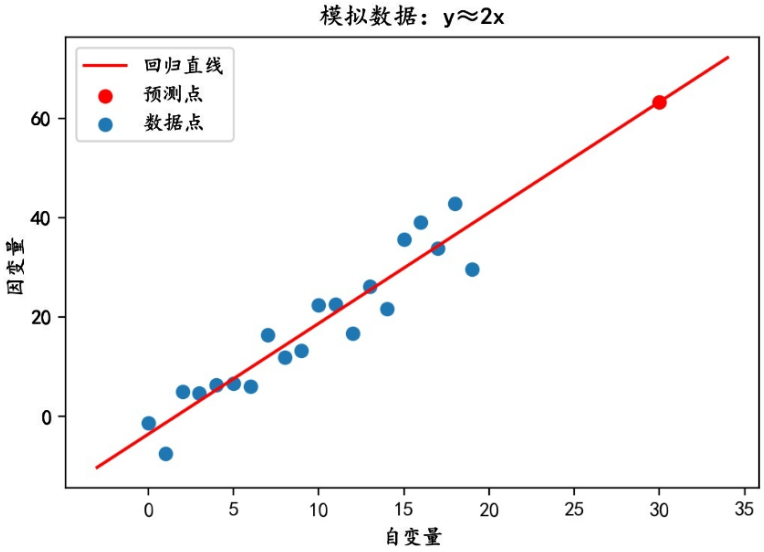

Figure 4. Linear regression model
图 4. 线性回归模型

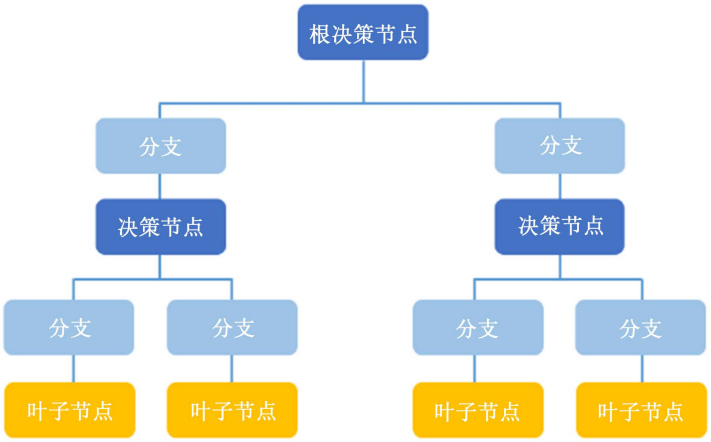

Figure 5. Decision tree model
图 5. 决策树模型

随机森林[8] (Random Forest)是一种集成学习算法，它建立在多个决策树的基础上，通过投票(对于分类问题)或平均(对于回归问题)来提高模型性能和泛化能力。每个决策树都在随机子集上训练，以减少过拟合风险。随机森林对数据中的噪声具有鲁棒性，广泛应用于各种分类和回归任务。模型结构图见图 6。

K 最近邻[9] (KNN)是一种基于实例的监督学习算法，它根据特征空间中数据点的距离度量来进行分类或回归。KNN 算法的核心是距离度量，通常使用欧几里得距离或曼哈顿距离来计算数据点之间的距离。K 值的选择对结果也有很大影响，较小的 K 值可能导致过拟合，而较大的 K 值可能导致欠拟合。在实际应用中，通常通过交叉验证来选择最优的 K 值。对于分类问题，KNN 查找最接近的 K 个邻居，并使用它们的多数类别来做出决策。对于回归问题，KNN 使用 K 个邻居的平均值作为预测结果。模型结构图见图 7。

支持向量回归[10] (SVR)是一种回归方法，它使用支持向量机(SVM)的思想来拟合数据并预测连续目标变量。与传统的线性回归不同，SVR 的目标是最大化在一定容忍度内的预测误差，以找到最佳拟合曲线或超平面。SVR 在处理高维数据时表现出色。模型结构图见图 8。

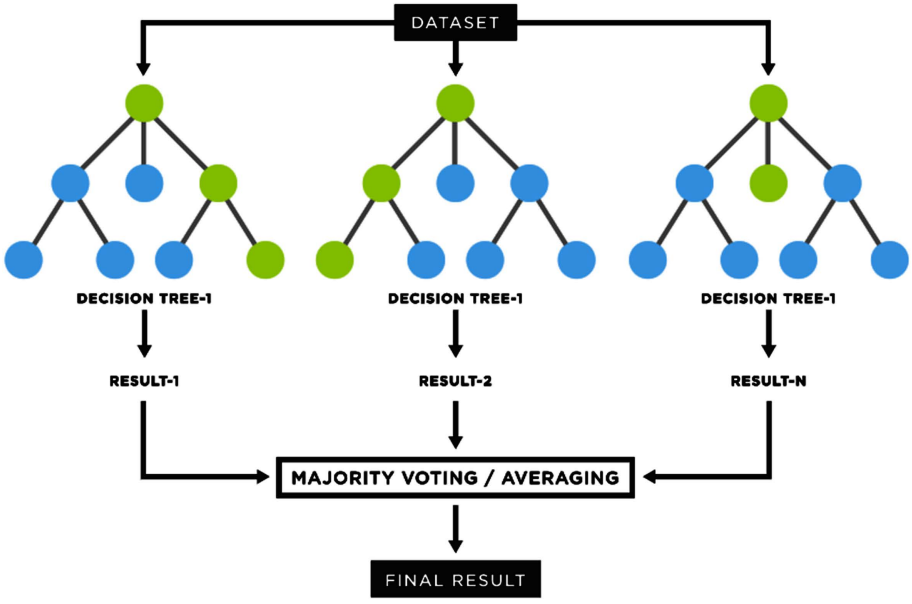

Figure 6. Random forestmodel
图 6. 随机森林模型

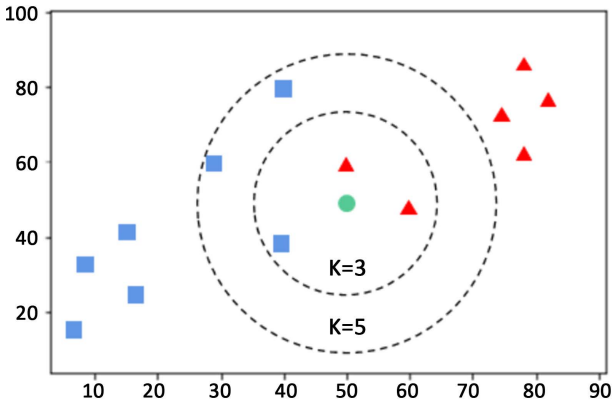

Figure 7. K-nearest neighbor model
图 7. K 最近邻模型

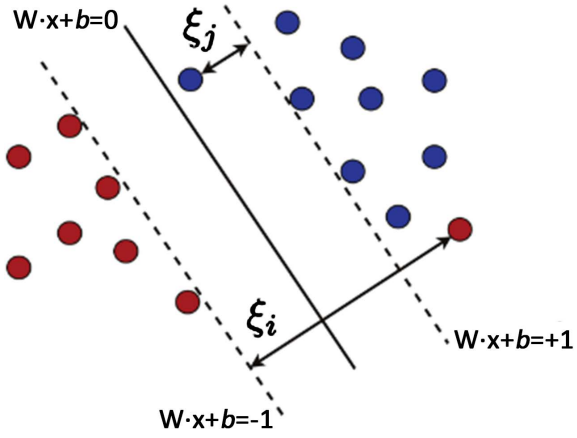

Figure 8. Support vector regression Model
图 8. 支持向量回归模型

2.3. 实验分析

2.3.1. 实验流程

实验的整体流程如下所示。

1) 煤数据收集和预处理。从相关数据源收集所需数据，并进行预处理。包括数据清洗、缺失值填充、异常值处理等，确保数据的质量和适用性。

2) 模型选择与训练。根据问题的特性和数据的类型，选择合适的回归模型进行训练。包括线性回归、决策树回归、随机森林回归、K 最近邻回归和支持向量回归五个类别。

3) 模型评估。使用 MSE (均方误差)、MAE (平均绝对误差)、RMSE (均方根误差)、MAPE (平均绝对百分比误差)和 NMSE (归一化均方误差)等指标对模型的性能进行评估。

4) 模型优化与调整。根据评估结果，对模型进行优化和调整，包括参数调整、模型融合等。

5) 结果分析。对比和分析不同模型的性能，找出各模型的优势和局限，以及探讨各模型在不同指标上的表现。

2.3.2. 影响因素

在具体实验中，“影响因素”可能在模型的训练、优化、评估等阶段起到影响作用。实验过程中具体的影响因素如下。

1) 数据质量。数据质量对模型的性能有很大影响。高质量的数据可以更准确地反映真实世界的情况，从而提高模型的预测精度。

2) 模型参数调整。模型的参数对其性能有很大影响。例如，对于随机森林和决策树，如果设置过大的树深度，可能会导致过拟合；而对于支持向量回归，如果设置过小的惩罚参数 C，可能会导致模型过于复杂，出现过拟合。

3) 评估指标的选择。不同的评估指标可能会得出不同的评价结果。例如，MAE 和 RMSE 对噪声大的数据更为敏感，而 MSE 则对所有误差同等对待。根据实际问题选择合适的评估指标很重要。

3. 结果分析

3.1. 模型评估

通过 MSE、MAE、RMSE、MAPE、NMSE [11]五类指标评估各回归模型性能，这些指标主要用于评估回归模型的预测性能，帮助评估模型的准确性，并在不同的业务场景中选择最合适的评价指标。

MSE (Mean Squared Error，均方误差)是预测值与实际值之差平方的期望值。衡量模型预测精度的常用指标，取值越小，表示模型预测精度越高。(y_i 表示真实值， $\hat{y}_i$ 表示预测值，n 表示样本数量)

$$\mathrm{MSE} = \frac{\sum\left(y_i - \hat{y}_i\right)^2}{n} \tag{1}$$

MAE (Mean Absolute Error，平均绝对误差)是绝对误差的平均值，能反映预测值误差的实际情况。MAE 取值越小，模型准确度越高。(Q 表示观测值，P 表示预测值，n 表示观测数量)

$$\mathrm{MAE} = \frac{\sum\left(Q - P\right)}{n} \tag{2}$$

RMSE (Root Mean Squared Error，均方根误差)是均方误差的算术平方根，该结果与实际数据的数量级一样。RMSE 取值越小，模型准确度越高。(y_i 表示真实值， $\hat{y}_i$ 表示预测值，n 表示样本数量)

$$\mathrm{RMSE} = sqrt\left(\frac{\sum\left(y_i - \hat{y}_i\right)^2}{n}\right) \tag{3}$$

MAPE (Mean Absolute Percentage Error，平均绝对百分比误差)是实际值与预测值之间的差异与真实值的比例的平均值。MAPE 越小，预测准确性就越高。(Q 表示观测值，P 表示预测值，n 表示观测数量)

$$\mathrm{MAPE} = \frac{1}{n} * \sum\left|\left(\frac{O-P}{O}\right)\right| * 100 \tag{4}$$

NMSE (Normalized Mean Squared Error，归一化均方误差)是一种归一化的 MSE，其计算方式是将 MSE 除以实际值的平方。NMSE 的值越接近于 1，说明预测值与实际值的差距越小，模型的准确性就越高。(MSE 表示均方误差，var(y)表示目标变量的方差)

$$\mathrm{NMSE} = \frac{\mathrm{MSE}}{\mathrm{var}\left(y\right)} \tag{5}$$

各模型评估结果见表 2，由表可知，线性回归模型在拟合度方面通常表现出色，模型的拟合程度较高，预测效果较好。这意味着该模型能够很好地适应数据集，并预测目标变量(因变量)和自变量之间的关系。同时，线性回归模型可以有效地利用历史数据和已知的自变量信息来预测未来的因变量值。与其他复杂的机器学习模型相比，线性回归模型在解释性方面也更加直观易懂。

Table 2. Regression model comparison
表 2. 回归模型对比

	MSE	MAE	RMSE	MAPE	NMSE
Linear Regression	1.683	0.210	1.871	12.541	0.112
Decision Tree	9.319	0.302	3.548	39.870	0.314
Random Forest	6.541	0.357	2.980	45.417	0.177
KNN	109.134	3.548	11.647	16.389	0.767
SVR	212.610	5.644	14.924	26.372	1.018

3.2. 线性回归预测煤炭浪费量

对五种回归模型对比分析可知，线性回归模型优于其他模型。因此对线性回归模型增加正则项(L1 或 L2 正则项)，使得模型复杂度和拟合效果之间取得平衡；减少特征输入个数，即对特征进行筛选，剔除对目标变量影响较小的特征。对数据训练 40 轮后得到五种指标变化数据(见图 9)。根据图 4 可以看出，随着迭代论述不断增加，回归优化模型的效果越来越好，可以用于预测煤炭资源走势，导入实际数据，我国煤炭行业在开采阶段浪费的煤炭资源比例高达 50%。这意味着如果按照年产煤量 35 亿吨计算，我们一年浪费的煤炭资源有 17.5 亿吨。

4. 结论

首先，为了提高陕北地区煤炭资源的利用率，我们通过五种回归模型分析了煤炭数据，发现回归模型对煤炭数据预测效果更优。在此基础上，尝试通过对线性回归模型增加正则项来优化模型，优化后的模型在复杂度和拟合效果之间取得平衡，减少了特征输入个数，剔除了对目标变量影响较小的特征，预测分析效果更优。最后，通过智能开采、资源再利用、洁净煤技术、碳捕获和储存技术、建立循环经济模式、环境监测与评估、政策引导等策略，减少不必要的资源浪费和环境污染，极大的提高煤炭资源的

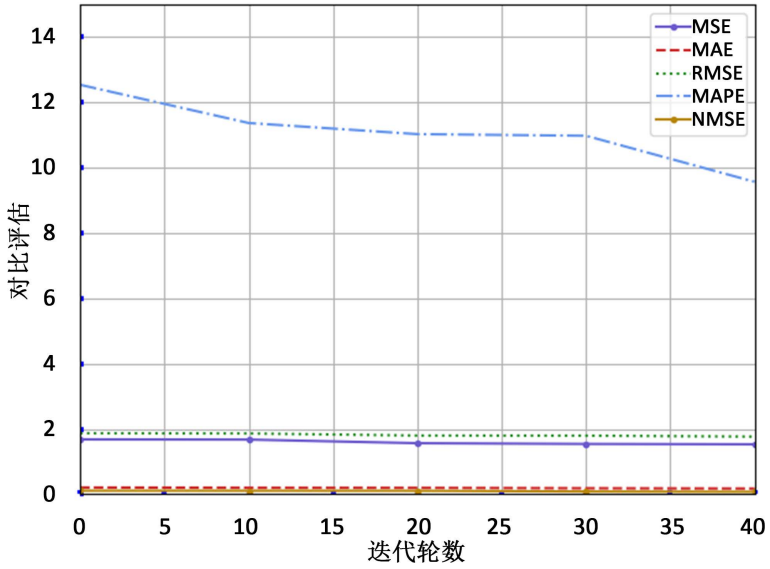

Figure 9. Model prediction results
图 9. 模型预测结果

综合利用效率，推动煤炭行业的可持续发展。

机器学习在煤炭资源利用优化策略中的应用是一个相对较新的研究领域，其潜力主要在于通过数据分析、模式识别和预测模型，以改进煤炭资源的利用和管理。基于机器学习的煤炭资源利用优化策略的优点如下所示。

1) 数据驱动的决策。机器学习技术可以处理大量的数据，通过识别和解读数据中的模式，为决策者提供更精确的指导。与传统的经验或规则驱动的决策方法相比，机器学习可以更好地利用历史数据和实时数据，提高决策的科学性和效率。

2) 预测性维护和管理。通过机器学习模型，可以对煤炭开采设备的性能进行预测性维护和管理，以预防潜在的问题和故障，降低维修成本。同时，这也有助于提高设备的使用寿命和整体的生产效率。

3) 优化煤炭资源配置。通过机器学习算法，可以优化煤炭资源的配置，包括确定最佳的采矿方法和地点。这有助于提高煤炭资源的利用率和企业的盈利能力。

4) 智能化决策支持：机器学习技术可以为煤炭资源的管理和利用提供智能化的决策支持。例如，通过自然语言处理(NLP)技术，可以从大量的文献和报告中提取有关煤炭资源管理和利用的信息，为决策者提供更全面的信息支持。

参考文献

[1] 马洁琼. 可持续发展视角的煤炭经济发展对策分析[J]. 现代工业经济和信息化, 2022, 12(3): 227-228. http://doi.org/10.16525/j.cnki.14-1362/n.2022.03.085

[2] 黄晨玮. 关于煤炭企业可持续发展能力的评估[D]: [硕士学位论文]. 曲阜: 曲阜师范大学, 2021. http://doi.org/10.27267/d.cnki.gqfsu.2021.000076

[3] 张晶. 基于可持续发展的煤炭经济发展探究[J]. 内蒙古煤炭经济, 2021(10): 209-210. http://doi.org/10.13487/j.cnki.imce.020326

[4] 宋思远. 基于机器学习的大中型煤炭企业信用风险预警研究[D]: [硕士学位论文]. 西安: 西安建筑科技大学, 2022. http://doi.org/10.27393/d.cnki.gxazu.2022.000100

[5] 周相团, 赵彬峰, 闫东, 等. 基于可持续发展视角的煤炭经济发展对策[J]. 内蒙古煤炭经济, 2021(6): 154-155. http://doi.org/10.13487/j.cnki.imce.019743

[6] 庄孝准. 核心素养之数学建模能力的培养——以线性回归模型为例[J]. 学周刊, 2023(25): 45-47. http://doi.org/10.16657/j.cnki.issn1673-9132.2023.25.015

[7] 孙泽东. 基于决策树蜂糖李等级分类系统[D]: [硕士学位论文]. 重庆: 重庆三峡学院, 2023.

[8] 张锟滨, 陈玉明, 吴克寿, 等. 粒向量驱动的随机森林分类算法研究[J/OL]. 计算机工程与应用: 1-12. http://kns.cnki.net/kcms/detail/11.2127.TP.20230904.1118.002.html, 2023-09-09.

[9] 刘斌毓. 轨迹数据 k 最近邻查询方法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2023.

[10] 李陈杰, 韩强, 庹先国, 等. 基于多核支持向量回归的浓香型白酒风味成分逐步预测模型研究[J]. 食品安全质量检测学报, 2023, 14(15): 185-194. http://doi.org/10.19812/j.cnki.jfsq11-5956/ts.2023.15.045

[11] Chicco, D., Warrens, M.J. and Jurman, G. (2021) The Coefficient of Determination R-Squared is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Computer Science*, **7**, e623. https://doi.org/10.7717/peerj-cs.623