ML_Molecule_Structure_Predictor

Public

main 1 Branch 0 Tags Go to file t Go to file Add file Code

cellister added R... 64f5429 · 1 minute ago 15 Commits

Data	exploring dataset	2 months ago
Photos	added RBF SVM visual	1 minute ago
Project_PDFs	added RBF SVM visual	1 minute ago
.gitignore	Initial Commit	2 months ago
ML_Molecular_...	added RBF SVM visual	1 minute ago
README.md	added RBF SVM visual	1 minute ago

README

Machine Learning Molecule Structure Predictor: Perovskite Classification

This project was completed as the final Phase 3 assessment in the Flatiron School's Data Science Bootcamp.

Analysis by Erin Wasserman, April 2024

Overview

This project explores different machine learning algorithms to predict the lowest distortion in perovskite structures, which is a critical aspect in ceramic science, materials physics, and solid-state inorganic chemistry. Researchers have identified a total of 73 elements in the A and B cation sites of ABO3 structures, leading to numerous oxides of the perovskite type.

About

No description, website, or topics provided.

Readme
Activity
0 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Jupyter Notebook 100.0%

The objective is to develop models capable of accurately classifying perovskite structures based on features such as electronegativity, ionic radius, valence, and bond lengths of A-O and B-O pairs, thereby aiding in the prediction and optimization of material properties. Model performance is assessed using the Accuracy Score metric.

Business Problem

A Research Laboratory has asked for an analysis of a dataset to help understand the relationships between various physical and chemical properties of perovskite materials and their crystal structure types.

Based on this analysis, the following recommendations are made:

- Utilize the Random Forest model.
- Enhance dataset with additional empirical data.
- Address missing values and discrepancies using insights from feature importance analysis to ensure data quality.
- Strategically target discrete values: Leverage RBF SVM's proficiency in capturing nonlinear relationships to target molecules with discrete values for improved predictive accuracy.

Data Understanding

Dataset Description

[Kaggle Crystal Structure Dataset] (https://www.kaggle.com/datasets/meetnagadia/crystal-structure-dataset/data). Each record represents a different Perovskite structures based on their characteristics. The data consists of 4,165 ABO3 perovskite-type oxides. Each observation is described by 13 feature columns and 1 class column which identifies it to be either a cubic, tetragonal, orthorhombic, and rhombohedral structure.

Data Features

List of Feature (or attributes, either work) Descriptions

- vA : Valence of A
- vB : Valence of B
- r_A6 : ionic radius of A cation (A6)
- r_A12 : ionic radius of A cation (a second one, A12)
- r_B6 : ionic radius of B cation
- EN_A : Average electronegativity value of A cation
- EN_B : Average electronegativity value of B cation
- bond_len_AO : Bond length of A-O pair
- bond_len_BO : Bond length of B-O pair
- ENR_diff : Electronegativity difference with radius

- tG : Goldschmidt tolerance factor
- tau : New tolerance factor
- mu : Octahedral factor

Data Splits

- The dataset was randomly split into training (80%) and test (20%) sets to evaluate model performance.
- Kfold Stratified crossvalidation was use to ensure that each class of the target variable is represented proportionally in each split.

Data Quality Issues

- No significant data quality issues were identified in the dataset.
- Nonphysical data was identified and removed using domain knowledge.

Data Licensing and Usage

- Database: Open Database, Contents: Copyright Original Authors

Methodology

Data Preparation

Data Cleaning and Preprocessing

- Missing values and placeholders were identified and removed from the dataset.
- Histograms, Correlation Plots, and Pair plots were used to inform data cleaning.

Modeling

- Decision Tree
- Random Forest
- Support Vector Machine

Evaluation

- Kfold Cross-validation
- Accuracy
- Comparison Table

Limitations

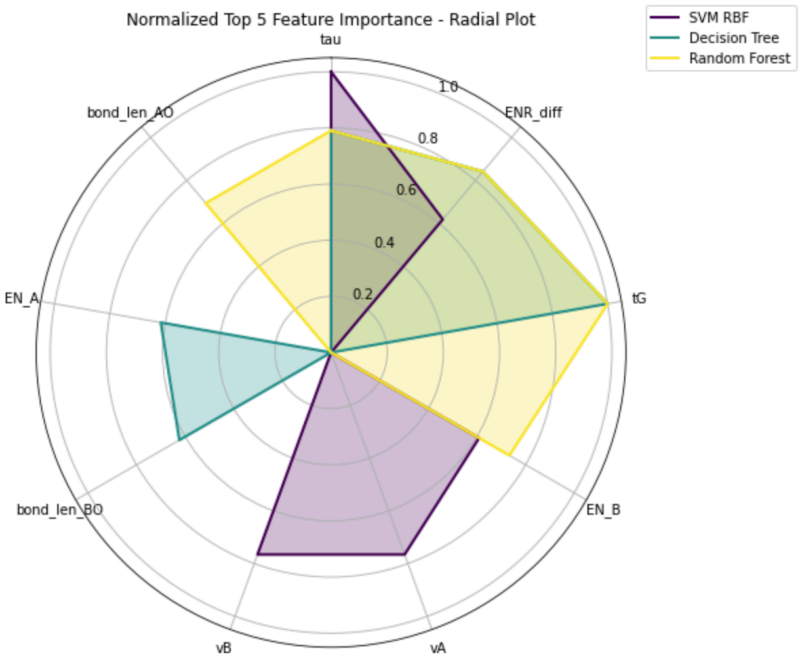
- More detailed explanation of the feature variables would have maximized the data cleaning.

- An accruacy of 100% was not achieved.

Key Findings

Comparison of Model Evaluation Scores

Model	Precision	Recall	Accuracy	F1-Score
Decision Tree	81%	81%	81%	81%
Random Forest	85%	86%	86%	85%
RBF Support Vector Machine	83%	83%	83%	82%



Rank	Decision Tree	Random Forest	RBF SVM
1	tG	tG	tau
2	tau	ENR_diff	vA
3	ENR_diff	tau	vB
4	bond_len_BO	EN_B	EN_B
5	EN_A	bond_len_AO	ENR_diff

*Radial Basis Function Kernel Support Vector Machine (RBF SVM)

Actionable Insights

1. Utilize the Random Forest model.
2. Enhance dataset with additional empirical data.
3. Address missing values and discrepancies using insights from feature importance analysis to ensure data quality.

4. Strategically target discrete values: Leverage RBF SVM's proficiency in capturing nonlinear relationships to target molecules with discrete values for improved predictive accuracy.

Next Steps

There are several directions that could be explored in the next phase of this project.

1. Gather Additional Empirical Data:
    - Focus on collecting empirical data, especially experimental data related to molecular structures. Bond lengths, bond angles, valences are critical empirical parameters.
    - Consider expanding the dataset to include a wide range of molecular structures and variations to capture diverse structural characteristics.
2. Improve Electronegativity and Ionic Radius Estimates:
    - Investigate methods to improve the accuracy of electronegativity and ionic radius estimates using data-driven approaches. Machine learning models can be trained to refine and optimize these parameters based on observed structural characteristics.
    - Assess the impact of adjusted electronegativity and ionic radius values on predictive model performance and molecular structure determination.
3. Feature Segmentation:
    - Utilize unsupervised machine learning techniques like clustering to identify patterns and group similar molecular structures. This can help in segmenting the dataset into distinct clusters without relying on labeled data.
    - Explore the use of clustering algorithms such as K-means or hierarchical clustering to identify structural similarities and differences.
4. Utilize Structural Parameters for Prediction:
    - Leverage structural parameters such as tG, tau, and mu to predict molecular properties and behaviors. These parameters can serve as valuable predictors in machine learning models.
    - Explore existing relationships between structural parameters and molecular properties, and consider incorporating them into the feature set for predictive modeling.
5. Deep Learning:
    - Investigate the application of deep learning models, such as neural networks, for molecular structure determination. Deep learning techniques can effectively capture complex relationships and patterns in the data.
    - Explore the use of deep learning architectures like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) for feature extraction and classification tasks.
6. Analyze Halides for Big Data Processing:

- Explore the inclusion of halides in the dataset for big data processing and analysis. Investigate how variations in halide compositions contribute to structural diversity and affect predictive model outcomes.
- Consider incorporating halide-specific features or parameters into the dataset to enhance the predictive capabilities of machine learning models.

These next steps aim to address identified inconsistencies, enhance the robustness of the analysis, and provide more actionable insights for the researcher's decision-making processes.

Author

Name: Erin Wasserman

GitHub: Cellister

Email address: cellister at gmail .com

Repository Structure

- **Jupyter Notebook**

The Jupyter Notebook is the key deliverable and contains the details of my data strategy, methodology, data cleaning, visualizations, and actionable insights.

- **Presentation**

This 5-7 minute, non-technical presentation was made in Canva and gives an impactful and brief overview of the key insights and recommendations.

- **Data**

The data used in this analysis can be found in the 'Data' folder. Some data can be found on the Kaggle Dataset.

```
├── Photos
│   ├── Other
│   ├── EDA
│   ├── SVM
│   ├── Random_Forest
│   └── Decision_Tree
├── Data
│   ├── Perovskite_train.csv
│   ├── Perovskite_test.csv
├── Project_PDFs
│   ├── ML_Molecule_Structure_Predictor.pdf
│   ├── ML_Molecule_Structure_Predictor.pdf
```