

Tarefa 2 - Formation Energy Regression

Alexandre Esposte Santos

27 de setembro de 2023

1 Objetivo

O objetivo desta atividade consiste em desenvolver um modelo de regressão para estimar os valores da energia de formação de diferentes materiais.

Os modelos a serem utilizados serão exclusivamente a regressão linear, Lasso e Ridge.

2 Desenvolvimento

Na lista a seguir apresentamos os passos efetuados durante a tarefa.

1. Carregamento e conhecimento da base de dados.
2. Verificação da integridade da base, isto é, verificação de valores ausentes/nulos e instâncias duplicadas.
3. Uma breve análise exploratória dos dados.
4. Divisão da base em treino e teste.
5. Definimos os modelos junto com uma pipeline que provê tratamentos específicos para cada variável do dataset. Os tratamentos consistem em preenchimento de valores faltantes, padronização das variáveis numéricas e a codificação de variáveis categóricas em variáveis numéricas.
6. Aplicou-se validação cruzada no conjunto de treino, dessa maneira formamos o conjunto de treino e validação.
7. Identificação de possíveis outliers e a remoção dos mesmos na base de treino e retreino do modelo através de validação cruzada.
8. Adicionando novas features através de transformações polinomiais e retreinando com validação cruzada.
9. Treinando os modelos com toda a base de treino e validando com a base de teste.
10. Analisando os coeficientes ajustados pelos modelos.
11. Avaliando minuciosamente os erros de cada modelo afim de detectar padrões.

Os passos listados acima são apenas um resumo do que foi feito. No notebook enviado há mais detalhes acerca de cada passo executado.