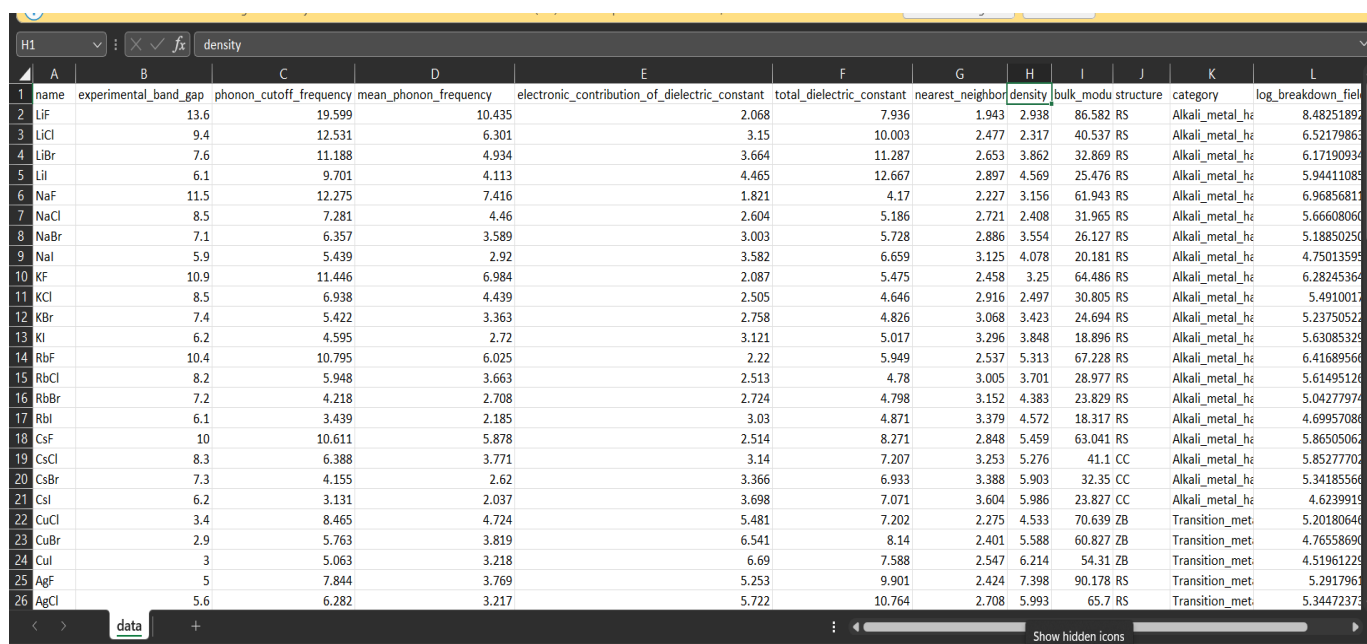


Dielectric Breakdown Prediction By Linear Regression

About Dataset : The challenge of formulating a predictive framework for the electric breakdown behavior of materials has persisted due to the intricate nature of dielectric degradation and breakdown in materials. This complexity arises from the dynamic interactions among various factors, including the intensity of the electric field, the duration of its application, ambient temperature, and the material's condition, particularly its defects and structural characteristics.

Empirically tracking dielectric degradation, which ultimately leads to breakdown, proves challenging due to the gradual formation and accumulation of defects at both the atomic and nanoscale levels. Therefore, the emphasis of this study is on the intrinsic dielectric breakdown field, a quantity calculated from first-principle numerical calculation. This parameter holds particular importance as it represents the maximum electric field that an ideal, defect-free material can withstand, thereby establishing the theoretical threshold for dielectric breakdown.

This dataset contains 82 rows of semiconductor/insulators, 2 columns of materials categories, and 8 columns of experimental/calculated (DFT) features. The prediction target being the intrinsic dielectric breakdown field, which is simplified as "log_breakdown_field" in the data file. The logarithm is taken in accordance with the treatments in the original paper. This dataset is intended to serve as an example for data-driven material discovery and for deriving phenomenological laws for electric breakdown field strength prediction.



	A	B	C	D	E	F	G	H	I	J	K	L
1	name	experimental_band_gap	phonon_cutoff_frequency	mean_phonon_frequency	electronic_contribution_of_dielectric_constant	total_dielectric_constant	nearest_neighbor	density	bulk_modulus	structure	category	log_breakdown_field
2	LiF	13.6	19.599	10.435	2.068	7.936	1.943	2.938	86.582	RS	Alkali_metal_he	8.48251892
3	LiCl	9.4	12.531	6.301	3.15	10.003	2.477	2.317	40.537	RS	Alkali_metal_he	6.52179863
4	LiBr	7.6	11.188	4.934	3.664	11.287	2.653	3.862	32.869	RS	Alkali_metal_he	6.17190934
5	LiI	6.1	9.701	4.113	4.465	12.667	2.897	4.569	25.476	RS	Alkali_metal_he	5.94411085
6	NaF	11.5	12.275	7.416	1.821	4.17	2.227	3.156	61.943	RS	Alkali_metal_he	6.96856811
7	NaCl	8.5	7.281	4.46	2.604	5.186	2.721	2.408	31.965	RS	Alkali_metal_he	5.66608060
8	NaBr	7.1	6.357	3.589	3.003	5.728	2.886	3.554	26.127	RS	Alkali_metal_he	5.18850250
9	NaI	5.9	5.439	2.92	3.582	6.659	3.125	4.078	20.181	RS	Alkali_metal_he	4.75013595
10	KF	10.9	11.446	6.984	2.087	5.475	2.458	3.25	64.486	RS	Alkali_metal_he	6.28245364
11	KCl	8.5	6.938	4.439	2.505	4.646	2.916	2.497	30.805	RS	Alkali_metal_he	5.49100171
12	KBr	7.4	5.422	3.363	2.758	4.826	3.068	3.423	24.694	RS	Alkali_metal_he	5.23750522
13	KI	6.2	4.595	2.72	3.121	5.017	3.296	3.848	18.896	RS	Alkali_metal_he	5.63085329
14	RbF	10.4	10.795	6.025	2.22	5.949	2.537	5.313	67.228	RS	Alkali_metal_he	6.41689560
15	RbCl	8.2	5.948	3.663	2.513	4.78	3.005	3.701	28.977	RS	Alkali_metal_he	5.61495120
16	RbBr	7.2	4.218	2.708	2.724	4.798	3.152	4.383	23.829	RS	Alkali_metal_he	5.04277974
17	RbI	6.1	3.439	2.185	3.03	4.871	3.379	4.572	18.317	RS	Alkali_metal_he	4.69957086
18	CsF	10	10.611	5.878	2.514	8.271	2.848	5.459	63.041	RS	Alkali_metal_he	5.86505062
19	CsCl	8.3	6.388	3.771	3.14	7.207	3.253	5.276	41.1	CC	Alkali_metal_he	5.85277702
20	CsBr	7.3	4.155	2.62	3.366	6.933	3.388	5.903	32.35	CC	Alkali_metal_he	5.34185566
21	CsI	6.2	3.131	2.037	3.698	7.071	3.604	5.986	23.827	CC	Alkali_metal_he	4.62399191
22	CuCl	3.4	8.465	4.724	5.481	7.202	2.275	4.533	70.639	ZB	Transition_metal	5.20180640
23	CuBr	2.9	5.763	3.819	6.541	8.14	2.401	5.588	60.827	ZB	Transition_metal	4.76558690
24	CuI	3	5.063	3.218	6.69	7.588	2.547	6.214	54.31	ZB	Transition_metal	4.51961229
25	AgF	5	7.844	3.769	5.253	9.901	2.424	7.398	90.178	RS	Transition_metal	5.29179611
26	AgCl	5.6	6.282	3.217	5.722	10.764	2.708	5.993	65.7	RS	Transition_metal	5.34472373

Figure 1 : Outlook of The Dataset

Selection of Machine Learning Algorithm : Linear Regression was chosen for this dataset because it is a simple and effective approach to establish a baseline model for predicting the log_breakdown_field. The dataset consists of numerical and categorical features, some of which, like experimental_band_gap, may have a linear relationship with the target variable. Linear Regression is computationally efficient and interpretable, making it ideal for smaller datasets like this one, where overfitting is less of a concern.

Code :

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt

data = pd.read_csv('data.csv')
print(data)
X = data.drop(columns=['log_breakdown_field'])
y = data['log_breakdown_field']
X = pd.get_dummies(X, drop_first=True)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
print(f"R-squared: {r2}")
plt.scatter(y_test, y_pred, color='blue')
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], color='red', linestyle='--')
plt.title("Actual vs Predicted")
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.grid(True)
plt.show()
```

Figure 2 : Implemented Code

Implementation Steps :

- **Load the Dataset:** Load the dataset into a pandas DataFrame from a CSV file.
- **Select Feature and Target:** Choose one feature as the independent variable (X) and the target variable (y).
- **Split the Data:** Divide the data into training and testing sets using an 80-20 split.
- **Initialize the Model:** Initialize the Linear Regression model using Scikit-Learn.
- **Train the Model:** Train the model on the training dataset to learn the relationship between the feature and target.
- **Make Predictions:** Use the trained model to predict the target variable on the test dataset.
- **Evaluate the Model:** Assess the model's performance using Mean Squared Error (MSE) and R-squared (R^2).
- **Visualize the Results:** Create a scatter plot comparing actual vs predicted values to understand model performance.

Result :

Mean Squared Error: 0.2670159421581349

R-squared: 0.8530525711326221

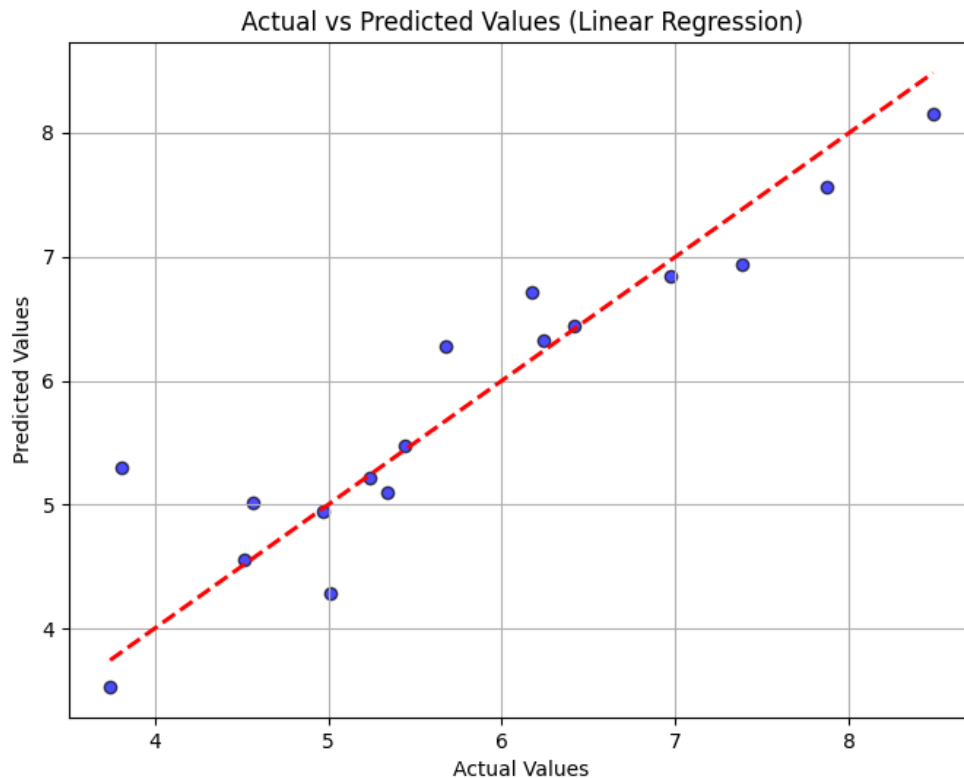


Figure 3 : Visualization of Result

Conclusion : The Simple Linear Regression model performed well, achieving a Mean Squared Error (MSE) of 0.2409 and an R-squared (R^2) value of 0.8605, indicating that approximately 86.05% of the variance in the target variable (log_breakdown_field) is explained by the feature (experimental_band_gap). However, the 24% error suggests that the model does not fully capture the variability in the data. This is likely due to the simplistic assumption of a purely linear relationship and the exclusion of other potentially influential features, such as phonon_cutoff_frequency or total_dielectric_constant, which may contribute to the target variable. Additionally, some of the variability may arise from noise or unmeasured factors. Despite these limitations, the model demonstrates a strong overall fit, and further improvements could be achieved by including more features or exploring non-linear models to better capture complex relationships.

Reference :

<https://www.kaggle.com/datasets/chaozhuang/dielectric-breakdown-prediction-dataset>