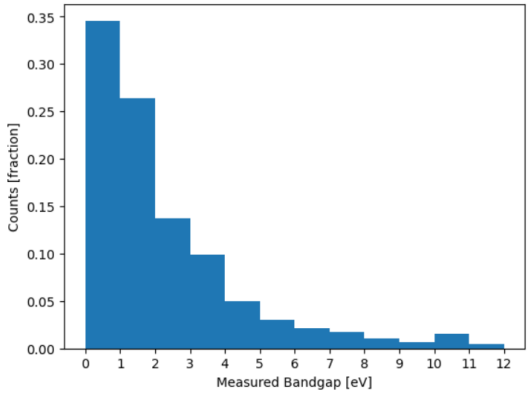
Assessment Figures

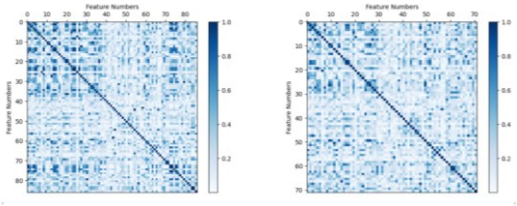



```
# features_generated
features_df.head(7)
```

[41] ✓ 0.0s Python

	AtomicNumber_composition_average	AtomicRadii_composition_average	AtomicVolume_composition_average	AtomicWeight_composition_average	BCCefflatcnt_co
0	6.0	1.1350	9311.576313	12.969702	
1	10.0	1.2700	9169.525548	21.197000	
2	19.0	1.3450	32.035942	43.422500	
3	15.0	1.5600	23.705899	35.645750	
4	28.0	1.4400	32.101458	66.922735	
5	23.0	1.5075	25.028908	57.450650	
6	6.0	0.9250	9300.147671	12.505791	

7 rows × 87 columns

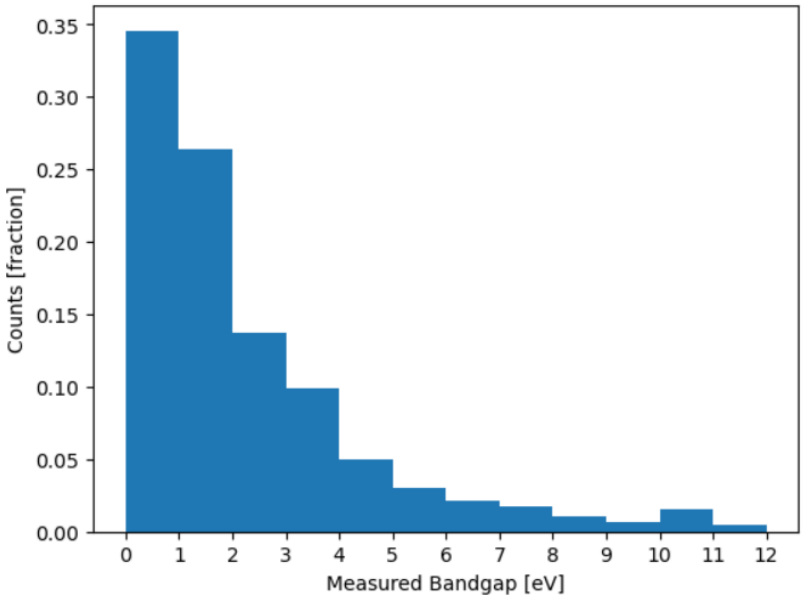


ML4ER Assignment 2

Jiahui Yang
Informatics Skunkworks
MSE 401, 3 Credits
Jul 21 2024

Progress

- Histogram plot before exercise 1.4



According to the histogram plot, we have more data in 0–4 than 5–12. It appears to be right-skewed or positively skewed, meaning that there are more data points at lower band gap values, and the frequency of data points decreases as band gap values increase. Thus, it is not uniformly distributed across its range and our band gap data is not balanced.

Progress

- Screenshot of the “features_df” from section 2

```python
# features generated
features_df.head(7)
```

[41] ✓ 0.0s Python

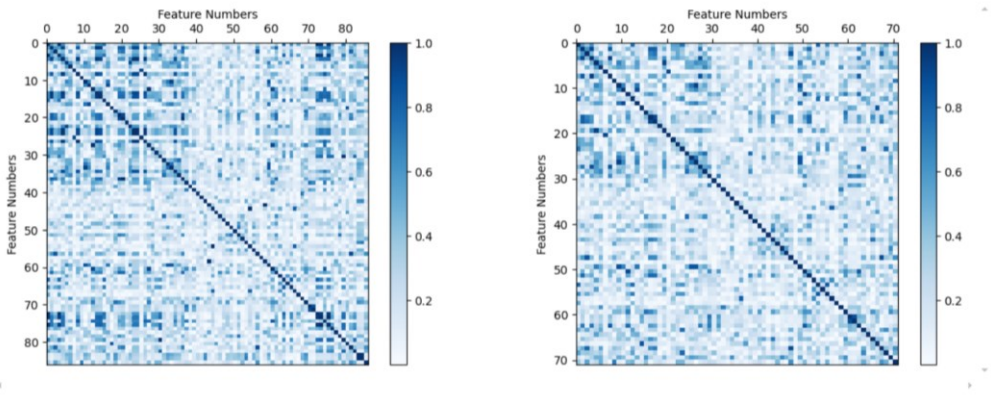
	AtomicNumber_composition_average	AtomicRadii_composition_average	AtomicVolume_composition_average	AtomicWeight_composition_average	BCCefflatcnt_cor
0	6.0	1.1350	9311.576313	12.969702	
1	10.0	1.2700	9169.525548	21.197000	
2	19.0	1.3450	32.035942	43.422500	
3	15.0	1.5600	23.705899	35.645750	
4	28.0	1.4400	32.101458	66.922735	
5	23.0	1.5875	25.028908	57.450850	
6	6.0	0.9250	9300.147671	12.505791	

7 rows × 87 columns

There are 87 features that we generated.

Progress

- Correlation matrix after removing highly correlated features above exercise 3.2




```
len(features_df_lowcorr.columns)
```

[48] ✓ 0.0s Python

71

There are 71 features left.

Problems

- On my local PC, I need to use “import pymatgen.core.composition” for the following code to work, even though we have already used “import pymatgen” at the beginning.

```python
import pymatgen.core.composition

# parse out individual elements for each formula using pymatgen's composition parser
element_list = list()
for idx in mastml_df.index:
    element_list.extend(pymatgen.core.composition.Composition(mastml_df["chemicalFormula Clean"][idx]).elements)
```

[29] ✓ 14.9s Python

Questions

- I am having difficulty understanding correlated features and determining how to choose the appropriate normalization or rescaling method for our features.
- Answered in the Friday discussion

Hours Summary

Date	Hours	Description of Work
07/18/2024	2.5 hours	Complete through section 3 of module 1