Project final report

Nguyen Xuan Binh
ID: 887799

14/04/2024

—

Multivariate Statistical Analysis
MS-E2112

1. Introduction

New material discoveries are a major driver of technological development. The discovery of steel and bronze in antiquity and the development of synthetic polymers in the 20th century were two examples of how new materials have drastically altered human society. These days, advances in materials science are also essential for addressing some of the most important social issues, like climate change and the future of our energy supply.

Nonetheless, there is still a lot of trial and error involved in materials discovery today. Finding a material that is appropriate for technological uses might take decades of study, and optimizing that material for commercialization can take even much longer time.

In this report, I introduce the Materials Project (www.materialsproject.org) dataset [1], which is a component of the Materials Genome Initiative. One of the Materials Project's key purposes is to compute the properties of compounds for which experimental data may be incomplete. This comprehensive dataset consists of **83989** atoms/molecules. The dataset is available at this URL

https://figshare.com/articles/dataset/Materials_Project_Data/7227749

In this dataset, there are in total 10 columns. The first column contains the chemical formula for the corresponding atom or molecule based on IUPAC nomenclature. Since the formula simply acts as an ID, it is not a feature. Then, the next 6 columns are physical properties of atoms or molecules.

The six columns (properties) are:

- energy_above_hull: This represents the energy above the convex hull for a given material. If a material has an energy above hull of zero, it is on the convex hull and is considered thermodynamically stable. A positive value indicates how much energy would need to be removed from the material to make it as stable as the stable phases on the hull. This is crucial for understanding phase stability in materials science.

- band_gap: This is an energy range in a solid where no electronic states exist. Substances having large band gaps (also called "wide" band gaps) are insulators, those with small band gaps (also called "narrow" band gaps) are semiconductor, and conductors either have very small band gaps or none, because the valence and conduction bands overlap to form a continuous band.

- total_magnetization: This is the density of permanent or induced magnetic dipole moments within a magnetic material.

- total_energy: This includes both kinetic energy (associated with molecular motion) and potential energy (related to chemical bonds) for an atom or the whole molecule

- energy_per_atom: This is the total energy roughly divided by the number of atoms in the unit cell. It tries to compare the energy contents of materials with different sizes and compositions.

- formation_energy_per_atom: This is the energy amount evolved or absorbed when a substance is formed from its constituent elements in a certain temperature and pressure

Finally, the rest 3 columns are elastic_anisotropy, K_VRH (Voigt-Reuss-Hill average of the bulk modulus), G_VRH (Voigt-Reuss-Hill average of the shear modulus) which has more empty entries than nonempty entries. Furthermore, when a material lacks elastic properties, it simultaneously lacks all 3 elastic properties. As a result, we can assign a label of 0 to materials without elastic properties and 1 to materials with elastic properties.

Below are some of the first datapoints. The absence of elastic properties for certain atoms or molecules, such as Xenon (Xe) can be due to its gaseous and liquid states. Elastic properties are defined for solids which have crystal structures, because they relate to the material's ability to resist deformation under stress. However, fluids do not have a crystalline structure, so they do not have elastic properties.

formula	energy_above_hull	band_gap	total_magnetization	total_energy	energy_per_atom	formation_energy_per_atom	elastic_anisotropy	K_VRH	G_VRH
Hf	0.071216	0.0000	-2.050000e-05	-9.883049	-9.883049	0.071216	0.881277	101.242732	44.836516
P	3.509988	2.0113	3.000042e+00	-1.895193	-1.895193	3.509988	10.884643	0.327165	-0.064038
Xe	0.005612	6.1701	0.000000e+00	-0.030139	-0.030139	0.005612	NaN	NaN	NaN
Hg	0.020462	0.0000	-2.800000e-06	-0.283229	-0.283229	0.020462	NaN	NaN	NaN
Br	0.615956	0.0000	-1.807400e-03	-1.013059	-1.013059	0.615956	-60.573886	21.044759	-18.850184

The research question that I am going to propose is as follows

What insights can be derived from the relationships and distributions of electronic, magnetic, and mechanical properties within various materials, and how do these properties influence the likelihood of materials possessing elastic characteristics?

To tackle this research question, I will try to uncover data patterns from the dataset with three stages

Stage 1: Univariate analysis, using general location and scatter statistics and histograms

Stage 2: Bivariate analysis, using pairplots and pairwise Pearson correlation coefficients

Stage 3: Multivariate analysis, using traditional principal component analysis (PCA), Fisher linear discriminant analysis (Fisher LDA) for classification and K-means for clustering analysis.

2. Univariate analysis

First and foremost, I derived the basic statistical values to measure the distribution for the 6 physical properties. They are mean, standard deviation, skewness, kurtosis, min, median and max.

stats	energy_above_hull	band_gap	total_magnetization	total_energy	energy_per_atom	formation_energy_per_atom
mean	0.142806	1.18132	5.315005	-171.8224	-5.815823	-1.475265
std	0.423493	1.56429	12.624827	185.5540	1.823426	1.247408
skewness	5.920774	1.34649	4.908275	-2.227180	0.318152	0.530711
kurtosis	40.817304	1.49737	40.911982	7.051213	0.226503	0.507995
min	0.000000	0.00000	-84.003218	-1834.877	-14.331771	-4.522664
median	0.027675	0.27240	0.001509	-107.7252	-6.036196	-1.613350
max	5.892481	17.8914	279.988888	-0.016100	-0.016100	4.828697

Based on the table, we can propose some notable features for each physical property

1. Energy above hull: The mean energy above hull is relatively low at 0.142806, meaning that most materials are close to being thermodynamically stable. However, the distribution of energy above hull is highly skewed (5.920774) and has a very high kurtosis (40.817304), suggesting a heavy concentration of values near zero with only few extreme outliers.

2. Band gap: The average is 1.181326, with a wide range up to 17.8914, indicating a diverse set of materials from conductors to wide-band-gap insulators. The band gap shows a positive skewness (1.346495) and a moderate kurtosis (1.497373), showing that many materials have low band gaps.

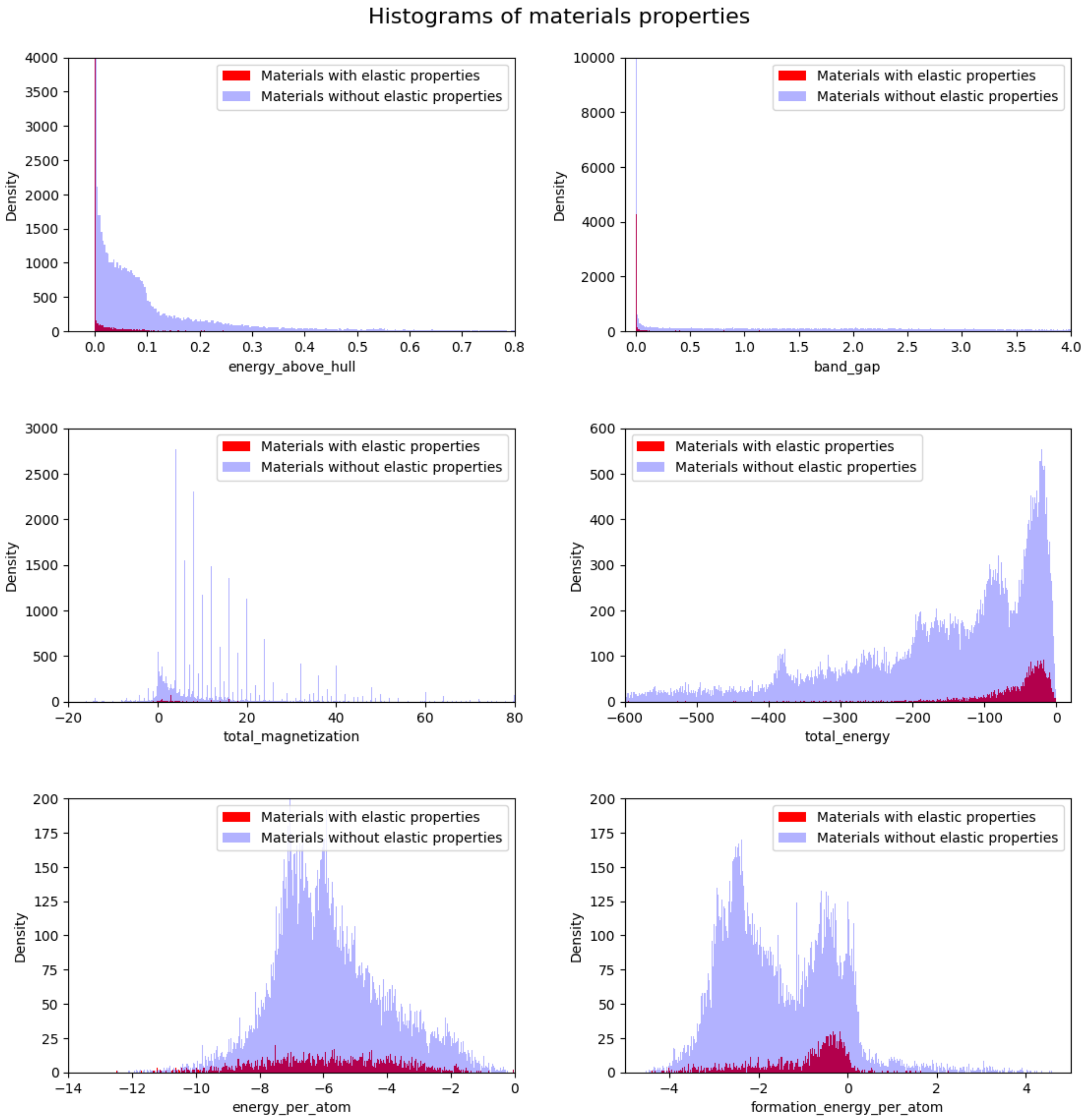
3. Total magnetization: it varies widely, from -84.003218 to 279.988888, with a mean close to 5.315005, showing diverse magnetization among the materials. The data is positively skewed (4.908275) and has high kurtosis (40.911982), showing many materials with near-zero magnetization.

4. Total energy: it has a mean of -171.822498 and ranges significantly from -1834.877179 to just -0.016100, showing a wide variation in the stability and energy states of the materials. The distribution of total energy is negatively skewed (-2.227180), showing that most materials cluster around a lower energy state, but there are some with much higher energy levels.

5. Energy per atom: average is -5.815823, with values ranging from -14.331771 to -0.0161, indicating that it varies significantly across different materials. This feature shows positive skewness (0.318152), which means that most materials have energy values slightly lower than the mean.

6. Formation energy per atom: The mean formation energy per atom is -1.475265, showing general trends in the energy released or absorbed during the formation of these materials. It also shows slight positive skewness (0.530711) and a low kurtosis (0.507995), suggesting a near normal distribution.

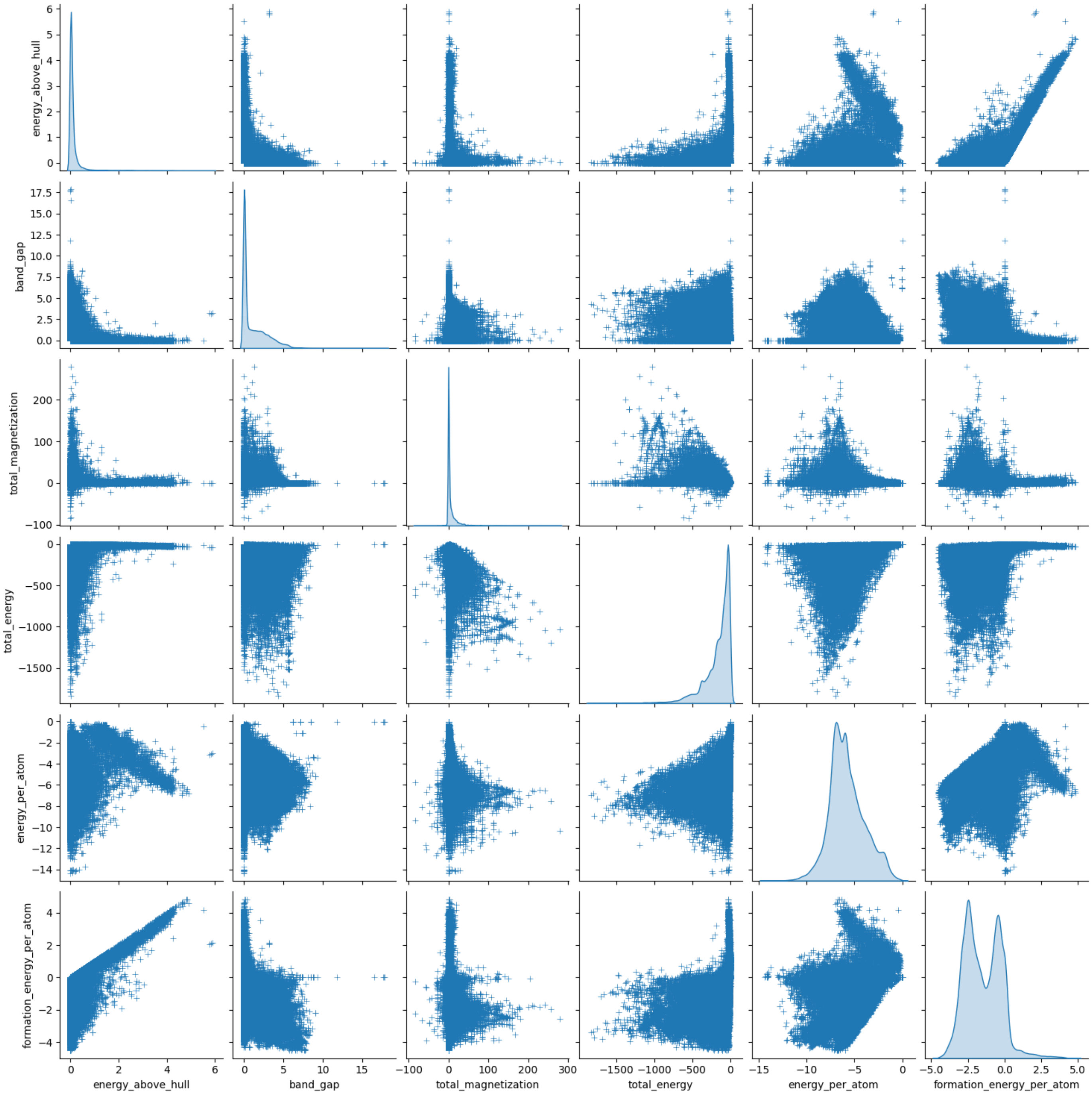
We need to notice that the table above show the statistics calculated for both materials with elastic properties (MWE) and materials without elastic properties (MWOE). However, we are also interested in the relative distribution of each feature for each material type as well. The patterns are clearer when we plot the features using a count histogram (not frequency histograms).



From the histograms, energy above hull and band gap are extremely skewed towards 0.0 value. For the rest features, they are more uniformly distributed, especially energy per atom and formation energy. The histogram shape of MWE and MWOE, while similar in general distribution, differ a lot in various local positions such as total energy, energy per atom and formation energy per atom, where there are several parts that their distributions do not resemble. This can be reasons to justify the motivation for clustering and classification later in multivariate analysis.

3. Bivariate analysis

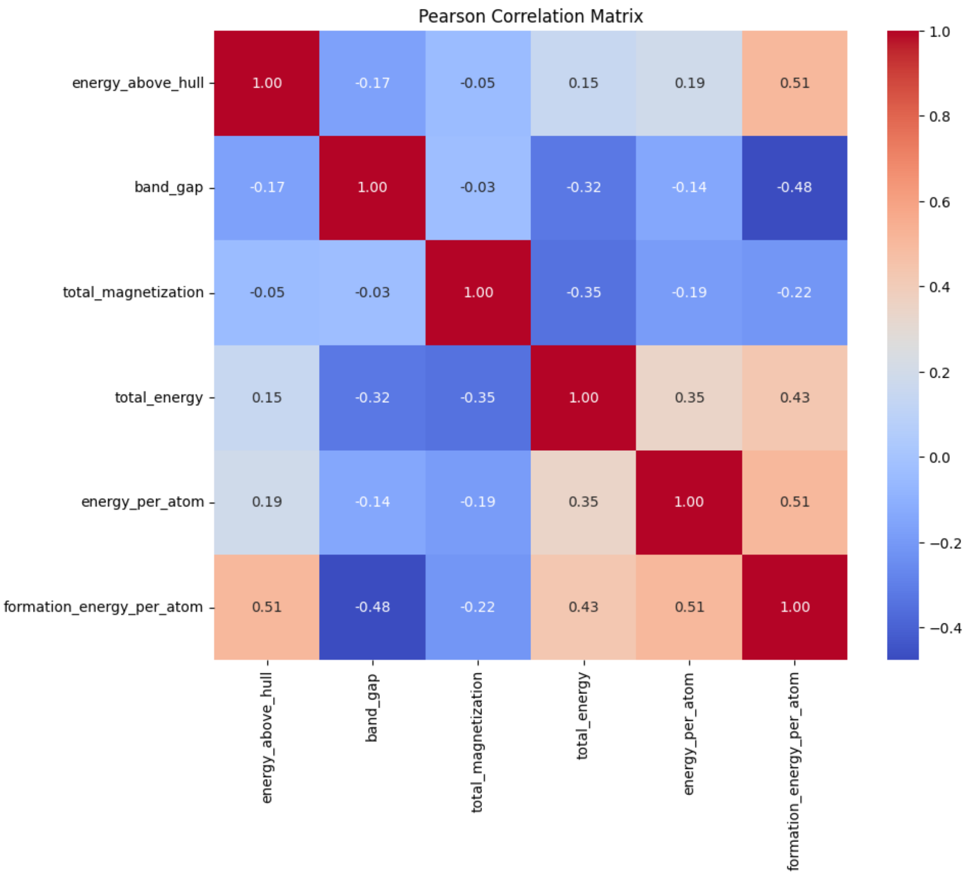
After univariate analysis, we can proceed to plot the pairwise scatter plots for the 6 features. Again, the pairplots include features from both MWE and MWOE



We can see that all features are not collinear in any way, perhaps only except energy above hull and formation energy per atom, where there is upper region where they have a clear linear relationship. To quantitatively verify the correlation between each variable besides the pairplot above, we can also calculate the pairwise Pearson correlation matrix. The Pearson coefficient formula is given as follows

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$$

where $cov(X,Y)$ is the covariance of X and Y, and σ_X and σ_Y are the standard deviations of X and Y



There is a strong positive correlation (0.51) between the energy above hull and the formation energy per atom, showing that materials requiring more energy to stabilize also tend to have higher formation energy. Additionally, the band gap shows moderate negative correlation (-0.48) with the formation energy, showing different types of bonding and electronic structures in insulators versus conductors. There is also a relatively high positive correlation (0.43) between total energy and formation energy, suggesting that when the stability of a material's structure increases (lower total energy), it requires more energy to form. Other than these, most other correlations are relatively weak.

4. Multivariate analysis

4.1 Selection of methods

Even though the project requirements only ask for one method, I proceed to conduct three multivariate statistical methods, one from each type to uncover as much knowledge as possible.

Dimension reduction with PCA: In the first stage, I use PCA to reduce the dimensions of the 6 features down to 3 principal components and visually verify if the MWE and MWOE have different scatters in the reduced dimensions. If they indeed have, then there we can weakly assume that the 6 physical properties can somehow differentiate MWE/MWOE.

Clustering with K-Means: Following dimensionality reduction, I used the K-Means clustering algorithm to test the hypothesis that materials can be categorized into two distinct groups with respect to their elastic properties. Particularly, I am trying to verify if the 2 components would result in the best clustering. If the answer is yes, then I proceed to hypothesize that these two clusters are indeed based on whether materials have the elastic properties or not.

Classification with Fisher LDA: Finally, for the classification task, I used Fisher's Linear Discriminant Analysis (LDA). Fisher LDA focuses on finding a linear combination of features that has the best separation between the classes. Spoiler is that it is indeed true K-means return the best number of components as 2, otherwise I would not have tried this method. The classification is binary version, where 0 is for MWOE and 1 is MWE.

4.2 Technical implementation

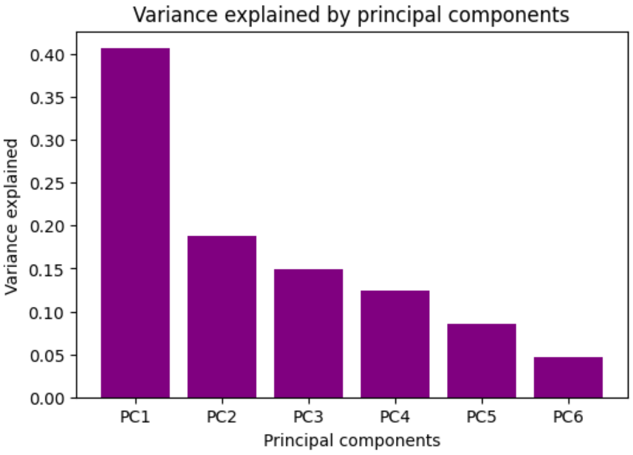
Dimension reduction with PCA: I used PCA from the scikit-learn library and applied it to the 6 standardized features to ensure each feature contributed equally to the analysis. Then, I proceed to plot the first 3 principal components that have the largest explained variance.

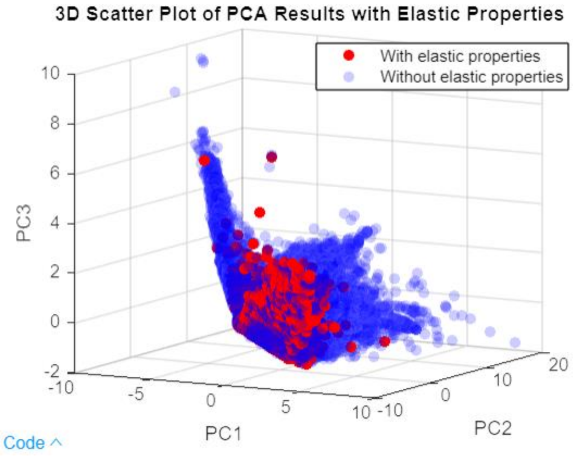
Clustering with K-Means: it was performed on the original standardized 6 features, not on the PCA-transformed data. I chose two clusters as the initial number of components, with increasing number of clusters until 10, to see which number of clusters has the best clustering results based on three metrics score: the silhouette score, the Davies-Bouldin score and the Calinski–Harabasz score.

Classification with Fisher LDA: it was applied to the original standardized dataset. I used the LDA classifier from scikit-learn with 80/20 ratio for training and testing. However, since the classes are imbalanced with number of MWOE 10 times larger than MWE, so I used a well-known over-sampling method called SMOTE [2], which samples synthetic points from minority class (MWE). Configurations are sampling ratio of 0.7 (ratio of MWE/MWOE), and number of nearest neighbors is 1000.

4.3 Result presentation and interpretation

Dimension reduction with PCA

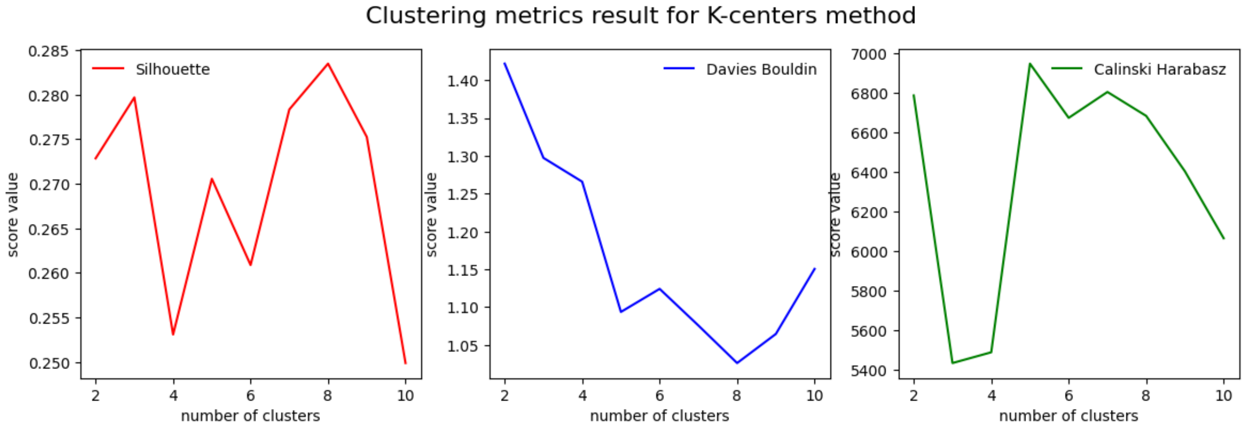




From the above figures, it does not appear that even in lower dimensions, the two classes are linearly separable. However, it is evident that MWE tends to only cluster at the center of the distribution of MWOE, which has a much wider range across three principal components. Additionally, it appears that MWE concentrates tightly along PC3, but MWOE spreads out much flatter. This can be a good sign that the classification task can be performed better than random metric performance.

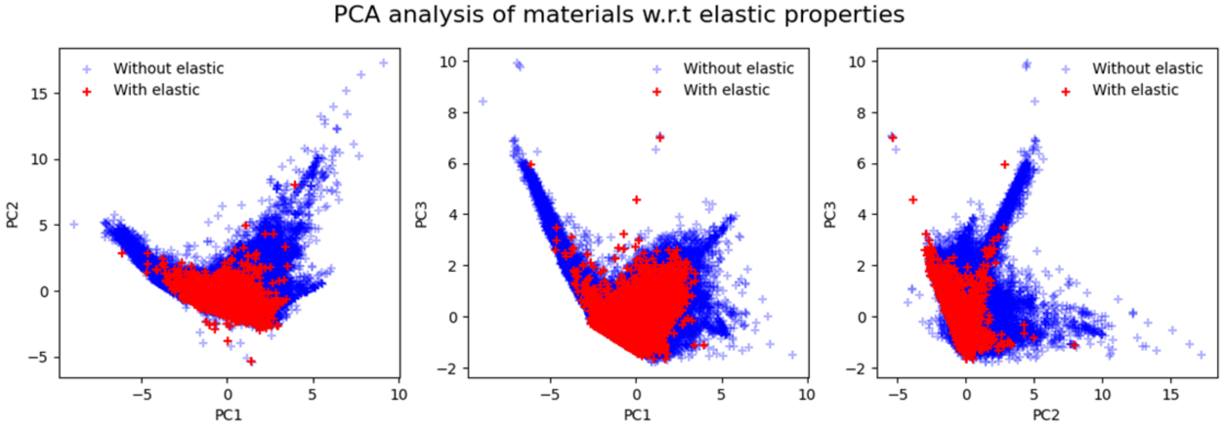
Clustering with K-Means

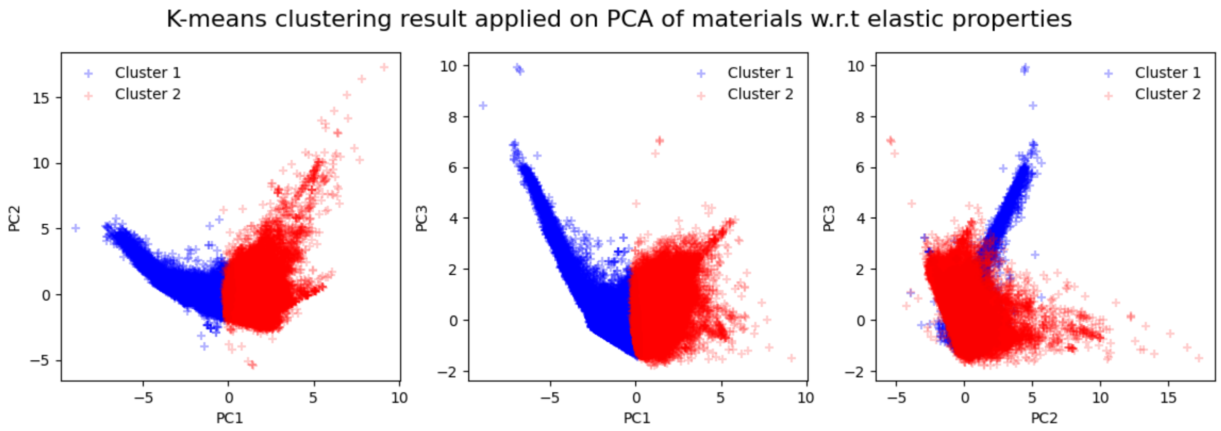
The reason that I used K-means and not others (spectral, hierarchical, depth-based, etc.), is because due to memory limitation, not because I do not suspect that they perform worse than k-means. For example, spectral clustering builds a similarity matrix of 83989 x 83989, which crashes my computer. Therefore, I test k-means on number of clusters from 2 to 10 to determine the optimal one.



- Silhouette score is the similarity of each point is to points in its own cluster compared to points in other clusters, with a range from -1 (incorrect cluster) to +1 (highly dense cluster).
- Davies-Bouldin score is the average similarity between each cluster with its most similar cluster, where lower score is better clustering with less overlap between clusters.
- Calinski-Harabasz score is the ratio of the sum of between-clusters dispersion to within-cluster dispersion for all clusters, where a higher score indicates better defined, distinct clusters.

We can see that the optimal number of clusters is 2, where its scores are non-dominated by any other number of clusters. As a result, I chose the number of clusters to be 2, and retrieved the clustering indices and project that classification result on the original PCA obtained previously.





Interpretation: it appears that the clusters by k-means and the clusters by elastic properties do not coincide in anyway. My conclusion is that optimal clusters number of 2 is purely random, because the original dataset does not have any clusters in the first place, so 2 simply results in the best scores.

Classification result with Fisher LDA

- The number of materials with elastic properties: 7676 (train: 6141 + test: 1535)
- The number of materials without elastic properties: 76313 (train: 61050 + test: 15263)

The classification metrics and confusion matrices are reported below with and without using SMOTE.

	Fisher LDA result without SMOTE	Fisher LDA result with SMOTE
Accuracy	0.9073	0.7718
Precision	0.4	0.2269
Recall	0.0287	0.6221
F1 score	0.0535	0.3325

Results without SMOTE	Predicted as MWOE	Predicted as MWE
Actual MWOE	15197	66
Actual MWE	1491	44

Results with SMOTE	Predicted as MWOE	Predicted as MWE
Actual MWOE	12009	3254
Actual MWE	580	955

From the results, the accuracy of 0.9 without SMOTE is misleading because it fails to detect the MWE and predict every material as MWOE. As a result, it has a very low recall of 0.028 and F1 of 0.05. The imbalance between the classes makes it hard to balance precision and recall. When SMOTE is used, the accuracy has dropped to 0.77, but it generally performs much better with the higher F1 score of 0.3325. Therefore, oversampling is crucial in improving Fisher LDA at classifying MWE and MWOE.

5. Conclusions and critical evaluations

Here are my conclusive answers to the research question posed in the introduction section

1. ***What insights can be derived from the distributions and relationships of electronic, magnetic, and mechanical properties within various materials?***

Distribution characteristics from univariate analysis: energy above hull and band gap are skewed towards 0, showing that most materials are close to thermodynamic stability and lower energy states, which have potential applications that require stable and energy-efficient materials.

Correlations between material properties from bivariate analysis: The positive correlation between energy above hull and formation energy per atom shows that materials requiring more energy to stabilize also have higher energy changes during formation. In contrast, the negative correlation between band gap and formation energy per atom shows the difference in conductors versus insulators. Materials with larger band gaps (insulators) generally require less energy for formation, which are useful in electronics where high resistance to electron flow is needed.

2. ***How do these properties influence the likelihood of materials possessing elastic characteristics from multivariate analysis?***

Dimensional Insights from PCA: MWE tended to cluster in the center of MWOE, suggesting that materials with elastic properties have a narrower range of physical properties.

Clustering Analysis via K-Means: two clusters provided the most meaningful separation, even though these clusters did not agree with elastic properties. It shows that while the six physical properties can help categorize materials, the way they classify does not really reveal the presence of elasticity.

Classification with Fisher LDA: the model has a good recall in SMOTE version, showing that when classes are balanced, the physical properties can be effective in predicting presence of elasticity.

However, there lacks a quantitative answer on the direct correlation between physical properties and presence of elasticity. The studies above are only purely qualitative to explore prediction of elasticity.

Critical evaluations:

Here are some sources of bias that I may have missed out in the analysis without testing them yet

1. Dimensionality reduction bias: Using PCA before clustering and classification can miss out on non-linear relationships between the 6 features for distinguishing between classes. A possible nonlinear dimensionality reduction method could be ISOMAP, aka geodesic distance.

2. Sampling bias in SMOTE: While SMOTE improves class balance and recall metric, it may create sampling bias by creating synthetic samples that are not logical for real-world materials (gas and fluids), which can make the model overfits on the minority class MWE.

3. Model assumption bias: Fisher LDA assumes that the data are normally distributed within each class and different classes share the same covariance matrix. It appears that the two materials are not normally distributed and would not share the same covariance matrix. A better classifying model without assumptions could be Random Forest.

References

[1] Jain, A.; Ong, S.; Hautier, G.; Chen, W.; Richards, W.; Dacek, S., et al. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials, 1(1), 011002. Report #: ARTN 011002. http://dx.doi.org/10.1063/1.4812323 Retrieved from https://escholarship.org/uc/item/3h26p692

[2] Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.