

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | Project final report | | | | |  | |
|  |  | | | | | | |  |
|  | | | |  |  | | | |
|  | | | | Nguyen Xuan Binh ID: 887799 |  | | | |
|  | | | | 13/04/2024—Multivariate statistical analysis MS-E2112 |  | | | |
|  | | |  | | |  | | |

# 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 column simply acts as an ID, it is not a feature. Then, the next 6 columns are physical properties of the 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 condition of 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.

# 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 |
| skew ness | 5.920774 | 1.34649 | 4.908275 | -2.227180 | 0.318152 | 0.530711 |
| kurto sis | 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 band gap 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 degrees of magnetization among the materials. The data is positively skewed (4.908275) and has a high kurtosis (40.911982), showing many materials with near-zero magnetization.

4. Total energy: The total energy 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).

A screenshot of a graph

Description automatically generated

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.

# Bivariate analysis

After univariate analysis, we can proceed to plot the pairwise scatter plots for the 6 features. Again, the pairplots include both MWE and MWOE, but in general, the pairplots should also look like the one below for individual case of MWE and MWOE.

A group of blue graphs

Description automatically generated with medium confidence

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 a 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

where is the covariance of and , and and are the standard deviations of and

A screenshot of a graph

Description automatically generated

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.

# 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 from this dataset. This is how I structured my approach in answering the scientific question.

***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?***

* Dimension reduction with traditional 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 at least provide some information on whether the material have elastic properties or not.
* 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.

All these statistical methods are conducted on the standardized version of the 6 features, as the features have vastly different value scales.

## 4.2 Technical implementation

Dimension Reduction with PCA Implementation

In the implementation of Principal Component Analysis (PCA), we standardized our dataset to ensure each feature contributed equally to the analysis. Utilizing a PCA algorithm from the scikit-learn library, we reduced our high-dimensional data to a lower-dimensional space that retained the most significant variance. This process enabled us to visualize the data more effectively and provided a simplified input for subsequent clustering and classification tasks.

K-Means Clustering Implementation

The K-Means clustering was performed on the PCA-transformed data. We specified two clusters as our initial parameter, aligning with our hypothesis regarding the division of materials based on elastic properties. The algorithm iteratively assigned data points to the nearest centroid and recalculated centroids based on current cluster memberships until convergence was achieved. The result was a partition of the dataset into two groups, which was then evaluated for its alignment with the known elastic properties of the materials.

Fisher LDA Classification Implementation

For classification, Fisher's Linear Discriminant Analysis (LDA) was applied to the original dataset with class labels indicating the presence or absence of elastic properties. We employed the LDA classifier from scikit-learn, which was trained on a subset of the data reserved for training purposes. The model's hyperparameters were fine-tuned using cross-validation to avoid overfitting. The classifier's performance, including its ability to generalize to unseen data, was quantified using metrics such as accuracy, precision, recall, and F1-score on a holdout test set.

## 4.3 Result presentation and interpretation

A graph of a number of components

Description automatically generated

A graph of a red and blue graph

Description automatically generated

A graph with a line

Description automatically generated

Contingency Table:

[[35388 40925]

[ 5904 1772]]

Chi-squared Test Statistic: 2602.0087847538152

P-value: 0.0

Phi Coefficient: 0.17601236322330074

The number of materials with elastic properties: 7676

The number of materials without elastic properties: 76313

Result without SMOTE

Accuracy: 0.9073

Precision: 0.4

Recall: 0.0287

F1 Score: 0.0535

Confusion Matrix:

[[15197 66]

[ 1491 44]]

Result with SMOTE

Accuracy: 0.7718

Precision: 0.2269

Recall: 0.6221

F1 Score: 0.3325

Confusion Matrix:

[[12009 3254]

[ 580 955]]

# Critical evaluations

# 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