CS+AI

COC257

B628327

Machine Learning Assisted Material Discovery

By

Joe Graham

Supervisors: Hossein Nevisi & Sina Saremi

Department of Computer Science

Loughborough University

April 2020

Machine learning assisted material discovery

The material discovery process is often a lengthy one, taking up to 15 years to find and experimentally test certain compounds. The purpose of this project is to determine whether machine learning techniques can help reduce the total time taken to discover new materials, specifically looking at ABO3 perovskites, and trying to develop an understanding of the relationship between the band gap in such and the tolerance factor along with other control variables by using linear and SVR regression algorithms.

Content

1. Introduction

1.1 Overview

The material discovery process refers to the steps undertaken to discover new materials. Material discovery can be extremely important in various industries, as it helps discover compounds with certain desired properties. For instance, the vehicle industry is constantly trying to discover new alloys which have slightly better rust resistance, or durability than the ones currently in use. Lithium batteries have also benefited from the material discovery process, improving drastically over the last few decades due to new discoveries with improved properties.

This said, there are several limitations when it comes to discovering new materials, the most notable one is the time taken for new material discovery, averaging 10-15 years. There are several industries trying to tackle these limitations by applying machine learning techniques to the process to speed up the overall time of discovery. Yet this too has its limitations, being that the majority of companies are performing such techniques behind closed doors, in other words, they are not sharing their data, making it harder for smaller companies, which do not have access to vast amounts of data, to replicate the speed undertaken by the larger corporations. Also, individuals who do publish their data/results often leave out the failed experiments, leaving a slightly biased set of data, from which mistakes cannot be learned.

The focus of this project is looking into simple ABO3 perovskites, defined as any material with the same type of crystal structure as calcium titanium oxide ($CaTiO_3$), trying to develop an understanding of the relationships between the different variables associated with such, looking at what data is available online and then using such data to develop machine learning algorithms capable of predicting values which may not have been discovered yet.

1.2 Aims & objectives

1.2.1 Initial project brief

The initial project brief proposed finding and collecting a vast amount of data, to then use a variety of machine learning techniques to find any relationships present within the data set. Thus, reducing the total time taken for the material discovery process. I mentioned the materials project database, a database composed of a variety of research and experimental data mostly from American companies, for which I created an account with the intention of using this data. During this project brief, I also spoke about several tools to be used. Latex was proposed to be used to write the report, sublime was the programming software mentioned along with word as the word processing software of choice. A record of useful links and pages was to be saved on a web-based storage tool. In order to download data from the material project website, the python materials genomics (pymatgen) library was proposed for use.

1.2.2 Altered project brief

The project developed, has not ventured very far from the initial project brief, as it was quite broad and allowed room for changes/developments throughout. During the development of this project, the key focus was still to determine whether machine learning techniques could help reduce the total time taken for the material discovery process. This said, there were several changes made during this project. Initially I was going to use the materials project database, yet a more detailed dataset, with a larger number of entries was found in the development stages. This then became the dataset that was used. Which also lead to a change in the variables to be found. Initially, the variable looked at trying to find a relationship for, was curie temperature, based on tolerance factor and other control variables. After finding this data set, the focus shifted to finding a relationship for bandgap, using the valence factors, tolerance factor and other control variables.

The tools mentioned during the initial project brief, quickly became redundant after further research, and development of the project. Instead of using sublime as the programming tool, PyCharm was used, this was an important change, as this software is specifically for python, allowing me to run the code on the console. In order to call the machine learning methods, the sci-kit learn library was used. The report has been written in word, instead of LaTeX and any images, such as result plots, were created using either the seaborn or matplotlib libraries. Another library used was pandas. This was used to create data frames, for loading and writing CSV files, and for creating graphs for plotting.

2. Requirements

The requirements section encompasses all the necessary information required for carrying out the development stages of this project. It involves the research stage at the initial stages of the project, and any research carried out throughout the rest of the assignment. This research has been gathered and displayed below in the form of a literature review, along with the data and tools available and required.

2.1 Literature Review

The literature review is split into different sections, where first we look at different machine learning algorithms, specifically classification and regression techniques, then ABO3 perovskites, and the material discovery process. Finally, we look at these two combined, where machine learning techniques have been used to assist the material discovery process, and more specifically, the ABO3 perovskites. The final section looks at any projects that are present which are identical or very similar to mine in terms of predicting the band gap using regression techniques.

2.1.1 Classification algorithms

Classification is a sub-section of machine learning; It involves a supervised learning approach in which the computer program learns from the input data and then uses this learning to classify new observations. (Sidana, 2020) This method can be extremely useful, as it allows the algorithm to learn which categories the data should be fit into, which may be instinctive for us humans. For instance, the MNIST data set is comprised of many hand-written digits in the form of images. Intuitively, it is relatively easy for us to place this data into categories. Yet the algorithm needs to learn certain rules for classifying this data. Yet once it has done so, will be able to classify such data at an extremely faster rate than any human can.

There are several different types of classification algorithms which all have their own benefits and downfalls depending on the data set involved. The most common algorithms include probabilistic methods, decision trees, rule-based methods, instance-based methods, support vector machine methods, and neural networks. (Aggarwal, 2014). For the purpose of this project, I am going to review the probabilistic and support vector machine methods.

Probabilistic classifiers and the archetypical naive Bayes classifier, are among the most popular classifiers used in the machine learning community and increasingly in many applications. (Garg and Roth, n.d.) The naïve Bayes method is one of the simplest and oldest analysis techniques still present today, developed in the 1700s; it was used in the aid of several historical events, including the Dreyfus affair, in which it was used as a way of providing new evidence in the court. Another well-documented example of use is the famous deciphering of the enigma code. (Zeger, 2012) The study of probabilistic classification is the study of approximating a joint distribution with a product distribution. Bayes rule is used to estimate the conditional probability of a class label y, and then assumptions are made on the model, to decompose this probability into a product of conditional probabilities. (Garg and Roth, n.d.) In other words, it makes use of conditional probability properties to predict the outcome as highlighted in figure 1.

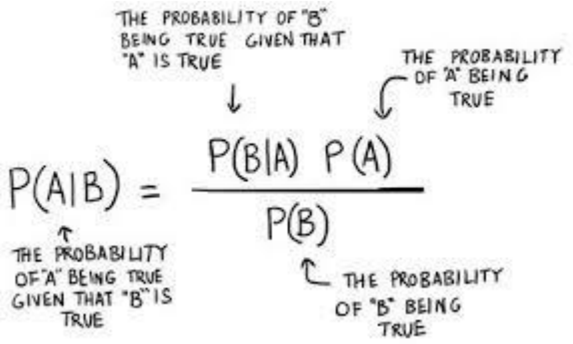
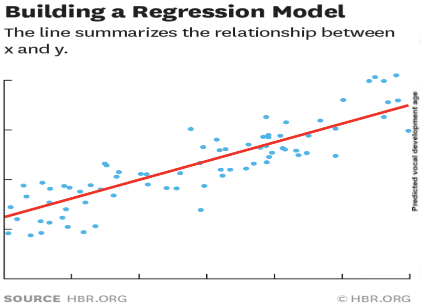

THE PROBABILITY OF "B" BEING TRUE GIVEN THAT "A" IS TRUE
THE PROBABILITY OF "A" BEING TRUE
P(A|B) = P(B|A) P(A) / P(B)
THE PROBABILITY OF "A" BEING TRUE GIVEN THAT "B" IS TRUE
THE PROBABILITY OF "B" BEING TRUE

Figure 1 (Naive Bayes, 2020).

Support vector machine, SVM, is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. (Noble, 2006) In simpler terms, it performs complex data transformation techniques, then determines the best way to separate the data base on the labels provided. SVMs are among the best (and many believe are indeed the best) “off-the-shelf” supervised learning algorithms. (Andrew Ng) Given the research, the algorithm to be used shall be the SVM algorithm, with a linear kernel given the data to be used.

2.1.2 Regression algorithms

The term regression has been defined in statistical analysis for centuries as the branch of statistics in which a dependent variable of interest is modelled as a linear combination of one or more predictor variables, together with a random error. (Bingham, Fry and Fry, 2010) In other words, Regression is the process of determining a relationship between two or more entities that manages to fit all the data given. Figure 2 represents an example of using linear regression to determine the line of best fit which summarises the relationship between unknown variables x and y. This process enables us to make a prediction for the x or y value given the value of the other attribute. Another form of regression is multi-linear, where more than one attribute is responsible for making a prediction, figure 3 shows an example of such.



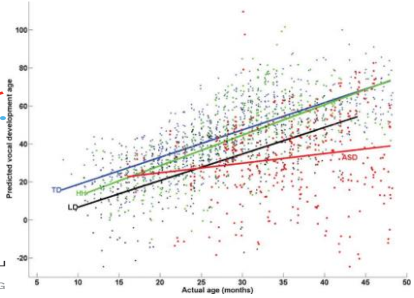


Figure 2 (HBR, 2020) *Figure 3 (Vandam, 2020)*

There have been multiple forms of regression algorithms been developed other the years, the most common of these include linear, logistic, polynomial and stepwise. (Launcher, 2020) For the purpose of this project I will look at linear and polynomial regression techniques given the data to be used.

Linear regression techniques try to recognize any relationships between the data, which are considered linear, in other terms, directly correlated. This can include, amongst many, Bayesian linear techniques or percentage techniques. These techniques can prove to be extremely useful for data which is considered uniform, where it can make very accurate predictions on unseen data. This said, it also comes with its downfalls, as the unseen data can sometimes not match the original data, where a threshold is met and the relationship changes such as seen in figure 4.

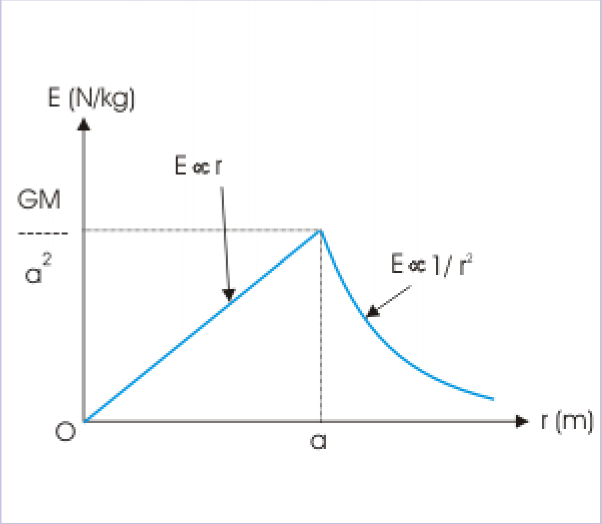


Figure 4 (cnx.org, 2020)

Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. (Python | Implementation of Polynomial Regression - GeeksforGeeks, 2020) Therefore, polynomial regression is very similar to linear regression in terms of trying to model a relationship between two or more variables, yet it can deal with multiple curvatures depending on the nth degree used.

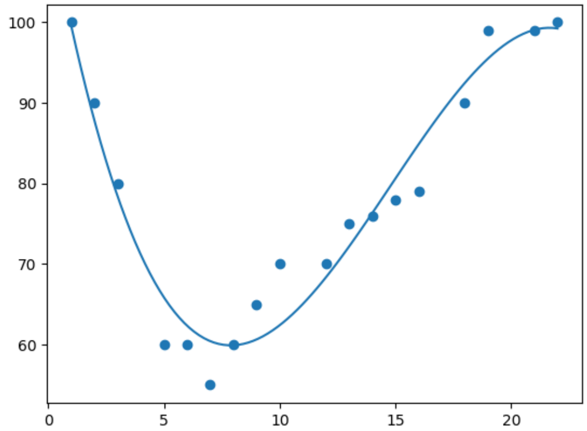


Figure 5 (W3 schools, 2020)

As a clear relationship between the features and the bandgap intended for use throughout the project was not clear, a mixture of regression models will be used and compared, to determine said relationship.

2.1.3 ABO3 material discovery

As previously stated, an ABO3 perovskite is the name given to a group of materials with general formula ABO3 having the same structure as the mineral calcium titanate (CaTiO3). There have been many applications for such compounds taking advantage of several properties, this is since the perovskite structure is one of the most extensively studied structures in materials science. (Mailadil T. Sebastian, 2008)

They can be useful in the production of solar cells, where their light absorption property is exploited; “Excitons are created after light absorption in the perovskite material, which is then separated into holes and electrons to be collected at the metal electrode and the FTO-coated glass electrode, respectively”. (Grace and Shapter, 2017)

$LaAlO_3$ perovskite was theoretically proposed as the best proton-conductor for solid electrolyte of solid oxide fuel cell. (Taku Onishi, 2015) These properties for advanced fuel cells has been exploited by various companies, which even NASA has taken part of. (Copeland-Johnson, 2020)

The varied and complex properties of ABO3 perovskites are the reason for the vast research being performed on such. Many industries, as seen above, have been trying to exploit such properties to suit their needs.

2.1.4 Machine learning assisted material discovery

In recent years, there have been a surplus in efforts to apply machine learning techniques to the material discovery process, mainly to reduce the total time of discovery, this is especially the case in the ABO3 perovskite search space, due to its popularity. Some of the projects and their outcomes are detailed below. (Yue, Tianlu, 2017)

Relating to the above material discovery applications, a machine learning model, specifically a regression technique was used to predict properties for identifying potentially stable solar cell materials. (Kailkhura et al., 2019) They did so by performing a machine learning pipeline, making use of various algorithms.

2.1.5 Regression techniques to predict band gap

After some research into this topic, it was clear that there had previously been similar projects to the one proposed, where the band gap is predicted based on a feature set close to the one proposed for use in this project. For instance, the band gap has been predicted using the valence vector sum of all three, A, B and O sites along with the formation energy. (Li et al., 2020)

Some of the regression models used include, linear ridge and random forest algorithms. In one project, they discuss how the linear ridge model “surprisingly” performs the best, which they associate with the basic features selected. Yet the random forest regression technique performed better on the data set. (Anantha, et. al.) Other models used include “decision trees, kernel ridge regression, extremely randomized trees, AdaBoost, and gradient boosting. The best performance is achieved by kernel ridge regression and extremely randomized trees.” (Gladkikh et al., 2020)

There seems to be a theme throughout most of the literature found, where randomized trees outperform the other models by some distance and linear regression models perform surprisingly well despite no known linear relationship between the features used and the band gap. Another running theme is the constant mention of insufficient data to predict values accurately, with most having to adapt or compile multiple data sets to be able to make sufficiently accurate predictions.

2.3 Data available

Despite the lack of digital data regarding the material project, in particular ABO3 perovskites as discussed throughout the literature, there were several databases with useful features found. At the initial stages of the project, the American, MPDB was intended for use. The Chinese rival of such database is known as the matmatch database, found at www.matmach.com. In order to develop a regression model to predict the curie temperature, as was the objectives at the start of the project, There was a database with various curie and Neel temperatures found online, on the magneticmaterials.org page following the following link http://52.56.120.52/data/. As only the curie temperature for ABO3 perovskites was needed, this needed to be selected from the data available, reducing the overall records from 39,000 to around 500 values. The data set which was used for the regression techniques throughout the project was found from a project aimed at calculating high-throughput DFT calculations as an online journal, the database associated with this data was found at figshare. (Emery, 2020)

2.2 Tools available/required

There is a large variety of tools available for use throughout this project, the tools needed are as follows, a tool for holding/collecting the data found. A tool for writing the code and the programming language to be used. Libraries for loading the data, performing machine learning methods and displaying the results. For writing the report, a word processing software needs to be used.

For each of these uses, there are multiple tools available, yet, there are some which are considered the best ones for the applications given amongst the computer science community, for instance, Python is the preferred programming language for applying machine learning, as it has many libraries available for such and is comprehensive and adaptive, in contrast to R for example. In order to store the data, the optimum choice is to use a CSV file, as opposed to building a database, simply put, as it is not essential. Also, there are python libraries, which can handle CSV files very easily in contrast to databases. There are several programming software which claim to be one of the best for writing python, amongst such are Sublime 3, Eclipse, Vi/Vim and PyCharm. I opted for the use of PyCharm, as it claimed to be beneficial for performing machine learning methods, with auto-fill options and a comprehensive GUI, it has been extremely beneficial, as it has helped me with my programming skills in other languages also. The most common library for performing machine learning methods, is that of sci-kit learn, which I therefore opted to use. In order to plot the graphs, the most common plotting library was also used, namely matplotlib. In order to write the report, I opted for the word processing software, Microsoft word, as opposed to LaTeX, although I have used LaTeX before in a previous project, I was more comfortable with word, especially given the layout for the LaTeX document was given in the previous case.

3. Design

3.1 Design specification

As discussed above, this purpose of this project is to use a variety of machine learning techniques to predict the band gap of ABO3 perovskites. In order to do so, the data wanted must be collected, merged and subsequently cleansed. This can be achieved by downloading the data set as a CSV file, dealing with all unwanted or null values by either deleting them or replacing them with averages. Also, the tolerance factor should be calculated and added to the dataset.

Once the data has been collected, the data wanted for analysis should be selected from the data set. This should be the data which has been experimentally tested. After the selection procedure, a mixture of regression techniques should be employed to determine which works best on the data set, namely, linear, polynomial and RBF models. The model which works the best, based on the accuracy or score values, should then be used to predict the band gap of the non-experimentally tested ABO3 perovskites.

3.2 How to implement tools

The various tools mentioned above can be used throughout the duration of this project. In order to commence the development stage, the programming language needs to be selected, in this case, python is the best candidate, as it has several libraries available for performing machine learning techniques, along with compatibility with most other programming languages. Python requires a simple download to get it running on your computer. Once this has been achieved, a programming software needs to be selected, in this case PyCharm is used. To set up a project in PyCharm, a python environment needs to be created and linked to the project via an interpreter. To do so, I installed Anaconda, an environment management software, intended for python. Using the Anaconda command line, it is possible to create a new python environment. Once the environment has been set up, all the required tools need to be installed into that environment, for instance, the sci-kit learn library, which is home to various machine learning techniques, and the Seaborn and matplotlib libraries, responsible for displaying data in a graphical format. To do so, the Anaconda command line can be used, calling the corresponding installs whilst accessing the python environment. In order to load the data using python, a library such as pandas can be used, allowing it to be saved as a data frame for further use.

3.4 MPDB

The MPDB refers to the materials project database, which can be accessed after successful registration on their website. They use an application programming interface (API) for giving access, hence the need for registration. Once registered, the user is given an API key, specific to said user. This system is based on

Representational state transfer (REST) principles, which use uniform resource identifiers (URIs). In order to access the data form the storage, there are a variety of simple queries that can be made, there is also a library, namely pymatgen, which makes more complex queries, such as a search for ABO3 perovskites available.

3.5 CSV

A CSV file is a type of storage technique, where data is saved as comma-separated values, where each line of the file is an independent data record, and each record consists of one or more fields, separated by commas. This allows large sets of data to be stored in a relatively small sized file, without the need of a database or other storage services. The most popular way to access these files is using a spreadsheet type software, such as Microsoft excel. These files are extremely useful, as the data from such can be loaded or written over easily by using a python library such as pandas.

4. Implementation

4.1 Python/code

The code is split into four different python files, this section will look at the different methods and parameters used, the reasons for that choice, and any issues or improvements worth discussing.

4.2 Classification

This file was developed during the initial stages of the project, since then, the aims and objectives have changed, making this file redundant for the actual project, yet I thought it was worth mentioning. The initial section of code is responsible for importing all the required libraries, and their sub directories. Pandas allows CSV files to be loaded and written to, sklearn allows for classification algorithms to be called and pymatgen allows for queries to the MPDB. The next code section is responsible for connecting to the MPDB and initializing the query for ABO3 perovskites. Firstly, my API key is saved as a variable, then the materials project rester subdirectory from the pymatgen library is called using said API key. Two arrays are initialized for use later. The query is then initialized, searching for all ABO3 perovskites, retrieving the pretty formula, e above hull, dielectric properties and whether it is compatible for each data record.

The next section of code filters this queried data, based on two criteria, taking values with 0 e above hull values that are also compatible. These values are then appended to the two arrays initialized above. After this, the arrays of data records, are passed as a data frame, and then exported to a blank CSV file already created.

The final section of code looked at creating a classification algorithm for predicting whether a compound was stable or not. Before being able to perform the classification techniques, I first found the data set, and added the tolerance factor to each, based on the equation t = (rA + rO)/sqr2(rB + rO), where r is the radius. I then labeled the data based on the data-sets guidelines stating that a compound with a e above hull value within 0.25 of 0 was stable, given by the label 1, and all others were given a label of 0. I then used a simple classification technique, predicting the value for the label, based on the calculated tolerance factor, given in the data set. This performed extremely well, due to the way the labelling was done. It made it redundant, as a filter could be used for values within 0.25 of 0 for the e above hull field. Therefore, rendering this classification algorithm non-essential for this project, yet could still be useful for perovskites not in the dataset.

4.3 Regression

This section looks at the different regression models used for trying to predict the band gap given the data set. There are several repeated sections of code throughout the three python files as I wanted to keep these separate for easier comprehensive and to manipulate them individually to see the results. This will only be discussed once in the first instance of such. Also discussed, are issues raised during the implementation of the models

4.3.1 Linear regression

The first section of code is responsible for importing all the required libraries for use throughout the file. The pandas library allows the data to be loaded. The sci-kit learn library was used again, this time for calling the regression methods. In order to display the data, the seaborn and matplotlib libraries were used.

The following section of code loads the data from a CSV file using pandas. This is loaded as a data frame, then the first 10 values are displayed to inspect the data is as expected. A graph is also plotted and displayed using the seaborn library of the most common values for the band gap field.

The following section is responsible for choosing the columns wanted as the control variables, in this case, valence A and B, the tolerance factor, the stability and formation energy. The data and target values are then initialised as desired, in this case, the target is the band gap. The data is scaled and normalized before being split into training and test sets, in this case, a split of 0.2 was used, in other words, 20% of the data was kept for testing.

Once the data was split into training and testing sets, the linear regression model could be called and fit to the data. In this case, a feature selection algorithm was called, taking only features below 80% of the variance threshold. This needed to be called and fit to the data. The values for the target data are then predicted.

The next section of code displays the performance of the algorithm, firstly, it displays the coefficient for each of the features used and the mean squared error. The difference between the actual and predicted values are then compiled into a data frame and displayed as a simple bar chart.

4.3.2 Polynomial regression

As the previous algorithm, the first section of code imports the necessary libraries, in this case, the seaborn library is not used, yet all the others are the same.

As before, the next section allows the data to be loaded, and the first 10 values displayed. The values for the band gap are not displayed in this file.

The next section of code initializes the polynomial function with degree 2. The data and target values are then initialised, and normalized based on l2, i.e. a squared error algorithm. For this regression model, a feature selection method, taking the top 60% of features was called and fit to the data, then the polynomial function is called and fit to the data.

Once the polynomial feature function has been fitted to the data, the data can be split, and the linear model called as previously, in this instance, due to the polynomial function, it will predict the target values based on a polynomial algorithm.

As previously, the coefficients and mean squared error are displayed, as well as a bar chart of the predicted compared to the actual values by first saving this as a pandas data frame.

4.3.3 Radial basis function regression

Again, the first section of code imports all the required libraries, in this case, the same libraries as the previous model are used, with differences in the sub directories.

As before, the next section allows the data to be loaded, and the first 10 values displayed. The values for the band gap are also not displayed in this file.

The data and target values are initialised and normalized based on the l2 method. A feature selection method similar to the previous model was used, where the best k features are selected, in this case, k is equal to 3. The data is then split into training and testing sets.

The SVR algorithm is then called, using the radial basis function kernel. The parameters selected were C = 100, gamma = 0.1, and epsilon = 0.1. I kept the default values for gamma and epsilon, as changing them did not improve the accuracy of the model, yet C seemed to perform better with a higher value. The target values are then predicted using this model.

The coefficients are not displayed as it is not possible due to it not being a type of linear function. Therefore, just the mean squared error is displayed. A data frame is created of the actual and predicted values, and displayed as a bar chart, as with the previous models.

5. Evaluation and discussion

In this section, the results of the classification and regression models, in the form of accuracy will be displayed, compared and then a discussion about the reasons behind these differences will ensue.

5.1 Model accuracies

The accuracies for the different regression models are displayed in several ways, the mean squared error, determines how far of the predicted values were on average from the actual values. A simple bar chart is then displayed showing the actual vs the predicted values, for a more visual analysis of the different models. The accuracy of the classification algorithm, although not essential for the purpose of this project, is also detailed below

5.1.3 Classification

The classification algorithm achieved an accuracy of 0.96

Table 1

Entry number	Actual	Predicted
3293	0	0
1315	0	0
1839	0	0
249	0	0

As you can see in table 1, most of the labels for the data were equal to 0, meaning they were mostly unstable. This may have contributed to a loss of prediction power on unknown data, where the algorithm biasedly selects 0 as the label for most compounds. This said, the classification accuracy was very high, and performed well on the known dataset.

5.1.2 Linear regression

The linear regression algorithm performed with an average mean squared error of 1.32.

Table 2

Valence A	Valence B	Tolerance factor	Formation energy	Stability
0.44323583	-0.44323583	0.79416486	4.70807147	-1.6350165

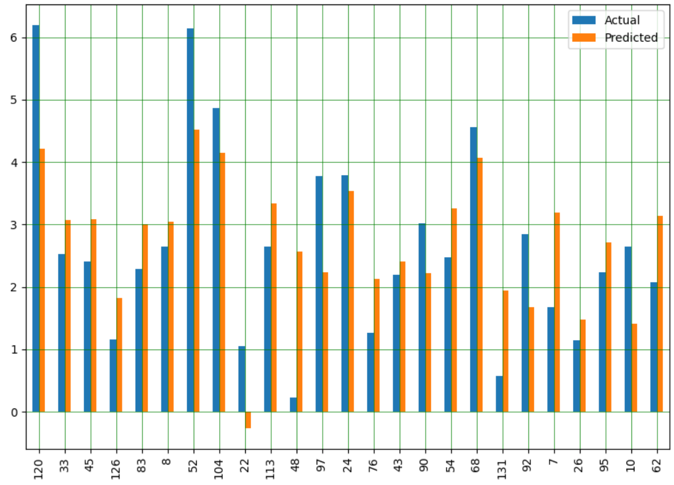


Figure 6

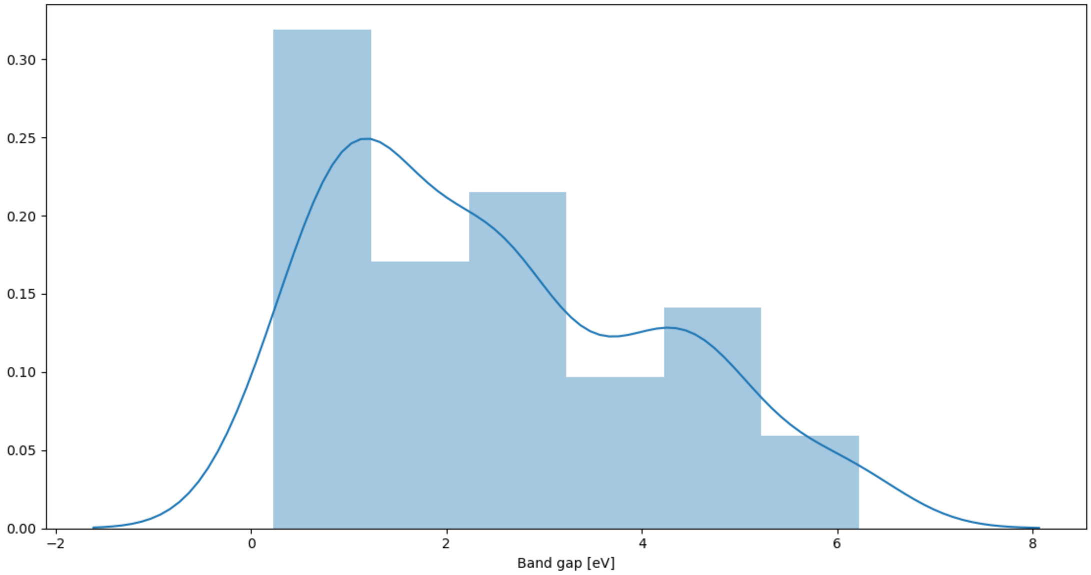


Figure 7

5.1.3 Polynomial regression

The polynomial model performed with an average mean squared error of 1.20.

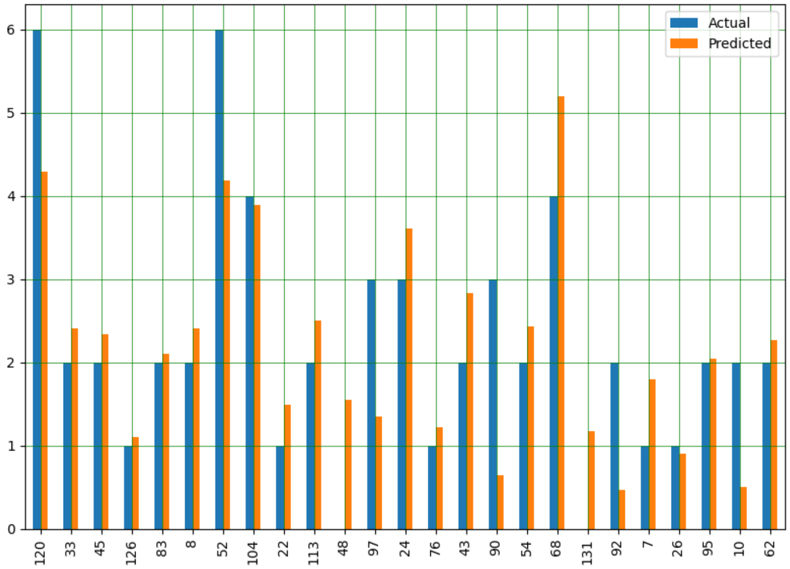


Figure 8

5.1.4 SVR radial basis function regression

The radial basis function regression algorithm performed with an average mean squared error of 2.03.

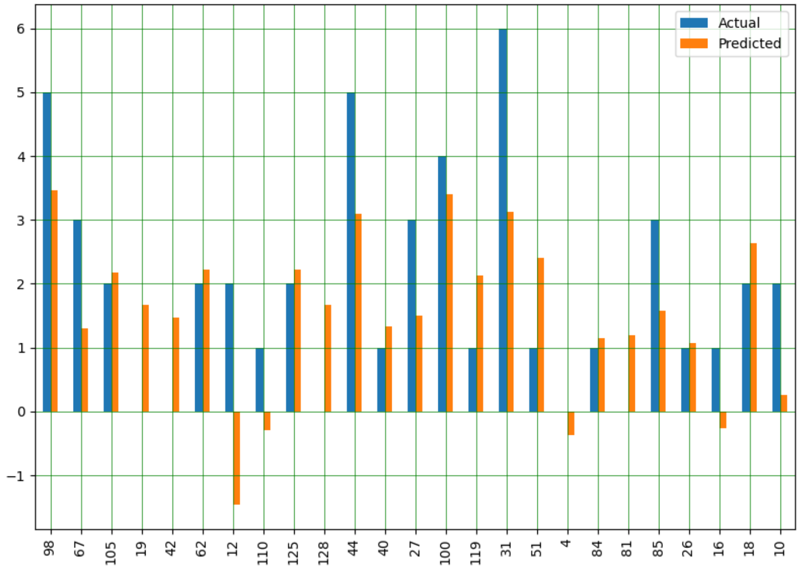


Figure 9

5.2 Difference between models

The polynomial model performed the best on the dataset when taking the mean squared error into account, whereas the RBF model performed the worse. It is also interesting to note that the linear model was not that far behind the polynomial model. Table 2 shows the coefficient values for the different features, the most correlated features are the formation energy and stability, by a considerable margin. This is probably since they measure the eV per atom, and the band gap is a measure of the eV.

Before applying the regression models, I believed that the RBF algorithm would perform the best, therefore, the results came as a surprise to me. These show that there is a polynomial relationship of second degree between the features and the band gap. Therefore, this regression model should be used to predict the values on the unexperimented data set.

5.3 Problems encountered

Several issues were encountered during the development of this project leading to key decisions being bade ultimately changing large portions of the project. After creating a file to connect to the MPDB, load the data, filter such data and save it to a CSV file, a larger data set was found. This happened during the development stage rather than the research phase, leading to a decision being made to opt out of using the original MPDB, and instead use the larger data set. In order to create a classification algorithm, I had to label the data available to me. To do so, the web page for the data set stated that all compounds within 0.25 of zero, are stable. This led to a naïve algorithm being created, which performed extremely well on the data, yet the reason for this was down to the simple labeling method, which was redundant, as a simple filter could be used instead. Therefore, a key decision was made to not use the classification model.

When developing the regression models, the original target data was looking at curie temperature, for which a separate database was found, and compiled to the larger data set. After performing a simple linear regression algorithm on such data, it was clear that there were insufficient data points to make an accurate prediction and resulted in another key change. This time, opting to change the target data to the band gap, which was a feature in the large data set. This was a risky change, as this topic was not encountered during the research phase and had to be researched during the development stage. Therefore, there was no way of knowing if a relationship was going to be present between the given features and the band gap before analysis.

During the analysis of the data, there were several changes which took place. I opted to delete all records with a value of zero for the band gap in the data, as it seemed to skew the results. Also, I did not take into consideration the different crystal shapes present within the data, which may have caused issues on the performance of the different models. Something worth mentioning also, was the changes made to increase the accuracy of the models. Many different feature selection models were used to assess which one worked best, along with the parameters for each method and the normalization and scaling functions. These functions helped improve the accuracy of the models by a significant amount.

5.4 How to improve project

There are several ways in which this project could be improved. Firstly, the data; It is well known that for accurate machine learning predictions, there must be a large quantity of data points for the machine to effectively ‘learn’. The data set I used was large, yet I only trained the models on those values that were experimentally tested. Therefore, to improve the dataset, more experimentally tested values should be recorded.

The regression models used could also have been improved in various ways; A larger range of models could have been used and compared rather than only three as there could be a better performing model not visited in this project. Also, the best performing algorithm was that of the polynomial kernel. This algorithm should have been taken and explored in different ways to maximise efficiency. Given that it is a polynomial model, all degrees should have been tested, with a variety of normalization functions using a range of parameters. Another issue with the regression models was the lack of consistency in the model chosen. All three models should have been SVR models using the different corresponding kernels for an accurate comparison.

6. Conclusion

This closing section intends to summarise all the points mentioned above, as well as drawing any valid conclusions from the data available.

To conclude, Despite the lack of data available to the public regarding materials, specifically ABO3 perovskites, in recent years, there has been several attempts and collecting data for such, for machine learning purposes. This project manages to capture some of the data available and perform three regression algorithms, namely, linear, polynomial and RBF, on such data to determine if it is possible to predict the values for perovskites still undiscovered. The polynomial algorithm with degree two performed the best, achieving a mean squared error of only 1.20. Although this figure is still relatively high, it shows that some sort of relationship is present, and given further probing, a more accurate prediction could be made which would be reliable enough for use in the material discovery process. Thus, proving the hypothesis that machine learning techniques can help shape the future of said process by reducing the overall time of discovery, making it a very valuable tool for many industries concerned in finding new improved materials such as the ones mentioned above. Hopefully in the future, more companies will start to adopt these techniques, which will also lead to more data being made available to the public, for more accurate predictions to be made, therefore, creating an exponential increase in material discovery development.

References

1. *Sidana, M., 2020. Intro To Types Of Classification Algorithms In Machine Learning. [online] Medium. Available at: <https://medium.com/@Mandysidana/machine-learning-types-of-classification-9497bd4f2e14> [Accessed 29 April 2020].*
2. *Aggarwal, C., 2014. Data Classification.*
3. *Garg, A. and Roth, D., n.d. Understanding Probabilistic Classifiers. University of Illinois.*
4. *Zeger, S., 2012. The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy, Sharon Bertsch McGrayne, Yale U. Press, New Haven, CT, 2011. $27.50 (320 pp.). ISBN 978-0-300-16969-0. Physics Today, 65(7), pp.54-56.*
5. *Figure 1. 2020. Naive Bayes. [image] Available at: <https://mc.ai/naive-bayesnb-classifier/> [Accessed 29 April 2020].*
6. *Noble, W., 2006. What is a support vector machine?. Nature Biotechnology, 24(12), pp.1565-1567.*
7. *Andrew Ng, CS229 Lecture notes part V, Support Vector Machines*
8. *Bingham, N., Fry, J. and Fry, J., 2010. Regression. London: Springer.*
9. *HBR, 2020. Linear Regression. [image] Available at: <https://hbr.org/2015/11/a-refresher-on-regression-analysis> [Accessed 26 March 2020].*
10. *Vandam, M., 2020. Multi Linear Regression. [image] Available at: <https://www.researchgate.net/figure/Multiple-linear-regression-MLR-model-predictions-for-individual-observations-ie_fig1_270961898> [Accessed 26 March 2020].*
11. *Launcher, C., 2020. CL Data School. [online] Machine Learning Internship. Available at: <https://www.careerlauncher.com/machine-learning/internship/4-common-regression-technique-in-machine-learning-you-need-to-check-out.html> [Accessed 29 April 2020].*
12. *cnx.org, 2020. [image] Available at: <https://socratic.org/questions/if-we-draw-a-potential-energy-u-vs-distance-r-curve-for-a-mass-m-in-the-gravitat> [Accessed 29 April 2020].*
13. *GeeksforGeeks. 2020. Python | Implementation Of Polynomial Regression - Geeksforgeeks. [online] Available at: <https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/> [Accessed 29 April 2020].*

14. *W3 schools, 2020. [image] Available at: <https://www.w3schools.com/python/python_ml_polynomial_regression.asp> [Accessed 29 April 2020].*

15. *Grace, T. and Shapter, J., 2017. Use of Carbon Nanotubes in Third-Generation Solar Cells. Industrial Applications of Carbon Nanotubes, 2017,.*

16. *Taku Onishi, Quantum chemistry in proton-conductors, Advances in quantum chemistry, 2015*

17. *Mailadil T. Sebastian, ABO3 type perovskites Dielectric materials for wireless communication, 2008*

18. *Copeland-Johnson, T., 2020. Development of Perovskite-Based Photovoltaic Cells For Extraterrestrial Energy Generation. [online] NASA. Available at: <https://www.nasa.gov/spacetech/strg/2013_johnson.html> [Accessed 29 April 2020].*

19. *Yue Liu, Tianlu Zhao, Wangwei Ju, Siqi Shi, Material discovery and design using machine learing 2017 pages 159-177*

20. *Kailkhura, B., Gallagher, B., Kim, S., Hiszpanski, A. and Han, T., 2019. Reliable and explainable machine-learning methods for accelerated material discovery. npj Computational Materials, 5(1).*

21. *Li, C., Hao, H., Xu, B., Zhao, G., Chen, L., Zhang, S. and Liu, H., 2020. A progressive learning method for predicting the band gap of ABO3 perovskites using an instrumental variable. Journal of Materials Chemistry C, 8(9), pp.3127-3136.*

22. *Anantha Natarajan S, R Varadhan, Ezhilvil ME, Band gap estimation using machine learning techniques, department of metallurgical and materials engineering.*

23. *Gladkikh, V., Kim, D., Hajibabaei, A., Jana, A., Myung, C. and Kim, K., 2020. Machine Learning for Predicting the Band Gaps of ABX3 Perovskites from Elemental Properties. The Journal of Physical Chemistry C, 124(16), pp.8905-8918.*

24. *Emery, A., 2020. High-Throughput DFT Calculations Of Formation Energy, Stability And Oxygen Vacancy Formation Energy Of ABO3 Perovskites. [online] figshare. Available at: <https://figshare.com/articles/High-throughput_DFT_calculations_of_formation_energy_stability_and_oxygen_vacancy_formation_energy_of_ABO3_perovskites/5334142> [Accessed 29 April 2020].*