



Université Mohamed Khider de Biskra  
Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie  
Département des Sciences de la Matière

# MÉMOIRE DE MASTER

Sciences de la Matière  
Chimie  
Chimie des matériaux

Réf. :

---

Présenté et soutenu par :  
**Abdesselam Yassine**

Le:

## **Using Data Mining to Search for Perovskite Materials with Higher Specific Surface Area**

---

**Jury :**

<b>Sriti Fatima Zohra</b>	<b>MC/A</b>	<b>Université Mohamed Khider de Biskra</b>	<b>Président</b>
<b>Djani Faical</b>	<b>MC/A</b>	<b>Université Mohamed Khider de Biskra</b>	<b>Rapporteur</b>
<b>Kenouche Samir</b>	<b>MC/A</b>	<b>Université Mohamed Khider de Biskra</b>	<b>Examineur</b>

Année universitaire : 2021/2022

## **Acknowledgments**

*I would like to express my gratitude to My supervisor Dr. Djani F., for his enthusiasm for the project, for his support, encouragement and patience, I would also like to thank very much, Ms. F.Z.Sriti (M.C.A at the University of Biskra) for agreeing to chair the jury, as well as the jury member Mr. S.Kenouche (M.C.A at the University of Biskra) for having done me the honour and for taking an interest in my work in judging him. I would also like to thank my friends and family who supported me and offered deep insight into the study.*

*"Anything worth doing well is worth doing poorly at first."*

*Ray Congdon*

# Summary

General introduction.....	1
References.....	4

## Chapter 01: Bibliographic Study

I. What is Machine learning? .....	5
I.1.Introduction.....	5
I.2.General Definitions .....	5
• I.2.1.Machine Learning.....	5
• I.2.2.Machine Learning categories.....	6
➤ I.2.2.1.Supervised Learning.....	6
➤ I.2.2.2.Unsupervised Learning.....	7
➤ I.2.2.3.Reinforcement Learning.....	7
• I.2.3.Machine Learning tasks.....	8
➤ I.2.3.1.Classification and Regression.....	8
• I.2.4.Training and test data.....	9
• I.2.5.Models.....	9
• I.2.6.Machine Learning Algorithms.....	10
➤ I.2.6.1.Support Vector Machine Algorithm.....	10
➤ I.2.6.2.Methods of Support vector machine.....	11
➤ I.2.6.3.Support vector machine principle.....	12
• I.2.7. Under-fitting and Over-fitting: Problems of Machine Learning.....	13

II. What is Data mining? .....	14
II.1.Introduction.....	14
II.2.General Definition.....	15
II.3. Data Mining Tasks.....	15
II.4. Data Mining Process.....	17
II.5.The Basic Data Types.....	19
III. Relationship between Data mining and Machine Learning.....	20
IV. Perovskite.....	21
IV.1.Introduction.....	21
IV.2.Perovskite ideal structure.....	22
IV.3. Tetragonal perovskite.....	24
IV.4. Rhombohedral perovskites.....	24
IV.5. Orthorhombic perovskites.....	24
IV.6. Monoclinic and triclinic perovskites.....	24
IV.7. General properties and application of perovskite materials.....	25
IV.7.1. Magnetic properties.....	25
IV.7.2.Optical properties.....	25
IV.7.3.Piezoelectricity.....	25
References.....	26

## **Chapter 02: Data mining and Machine learning methods For Materials Discovery and Optimization**

I. Introduction.....	28
II. Why Perovskite materials?.....	29
II.1.The traditional way to develop materials.....	30
II.2. Methods of synthesis and characterization of mixed oxides.....	30
• II.2.1. Methods of synthesis.....	30
• II.2.2.Characterization methods.....	34
III. Applying Machine learning and Data mining methods in perovskite materials design and discovery.....	41
III.1.The workflow of Machine Learning and Data Mining.....	41
• III.1.1.Data preparation.....	41
• III.1.2.Feature generation and Feature selection.....	44
• III.1.3.Model selection.....	45
• III.1.4.Model evaluation.....	46
• III.1.5. Model application.....	48
IV. Application of machine learning and data mining in perovskite materials.....	49
References.....	50

## **Chapter 03: Using Data Mining to Search for Perovskite Materials with Higher Specific Surface Area**

I. Introduction.....	56
I.1.What is Weka ?.....	57
I.2.How to use Weka? .....	58
II. Executing My Data mining workflow.....	62
II.1.Data preparation.....	62
II.2.Feature selection.....	71
II.3.Model selection.....	81
II.4.Model application.....	93
References.....	95

## **Chapter 04 : Synthesis and characterization of some of the Perovskite samples**

I. Introduction.....	97
II. LaFeO <sub>3</sub> , LaMgO <sub>3</sub> , LaMg <sub>0.6</sub> Fe <sub>0.4</sub> O <sub>3</sub> , LaFe <sub>0.8</sub> Mg <sub>0.2</sub> O <sub>3</sub> , and LaFe <sub>0.7</sub> Mg <sub>0.3</sub> O <sub>3</sub> preparation by sol-gel method.....	98
III. Characterization of the prepared samples.....	99
III.1. X-ray Powder diffraction (XRD).....	99
III.2. TGA analyse for LaFeO <sub>3</sub> , LaMgO <sub>3</sub> , LaMg <sub>0.6</sub> Fe <sub>0.4</sub> O <sub>3</sub> , and LaFe <sub>0.8</sub> Mg <sub>0.2</sub> O <sub>3</sub> .....	100
III.3. FTIR analyse for LaFeO <sub>3</sub> , LaMgO <sub>3</sub> , LaMg <sub>0.6</sub> Fe <sub>0.4</sub> O <sub>3</sub> , LaFe <sub>0.8</sub> Mg <sub>0.2</sub> O <sub>3</sub> , and LaFe <sub>0.7</sub> Mg <sub>0.3</sub> O <sub>3</sub> .....	101
III.4 Specific area measurement by the BET method.....	102
General Conclusion.....	103

# List of Figures

## Chapter 01

Figure.I.1	A generic machine learning method.....	6
Figure.I.2	Example of a prediction model.....	7
Figure.I.3	Simple diagram illustrates the different ML algorithm, along with the categories.....	8
Figure.I.4	Simple diagram representing a. Classification, and b. Regression.....	9
Figure.I.5	A diagram representing a classified dataset into two different classes.....	10
Figure.I.6	Simple graphs representing a. 2 datasets, and b. separated datasets into 2 classes.....	12
Figure.I.7	The best boundary (hyper plane) found by the SVM algorithm.....	12
Figure.II.1	Four of the core data mining tasks.....	16
Figure.II.2	The data processing pipeline.....	17
Figure.III.1	Visual representation of the relationship between data-related fields.....	20
Figure.IV.1	Structure of a perovskite with general chemical formula $ABX_3$ . The red spheres are X atoms (usually oxygens), the blue spheres are B atoms (a smaller metal cation, such as $Ti^{4+}$ ), and the green spheres are the A atoms (a larger metal cation, such as $Ca^{2+}$ ).....	22
Figure.IV.2	The idealised perovskite structure of $SrTiO_3$ : (a) atom positions with $Sr^{2+}$ at cell origin; (b) $TiO_6$ octahedral coordination polyhedron; (c) atom positions with $Ti^{4+}$ at cell origin; (d) $TiO_6$ octahedral polyhedron framework with $Sr^{2+}$ at the cell centre; (e) cuboctahedral cage site.....	23

## Chapter 02

Figure.II.1	Number of published papers. (a) On keyword of ‘perovskite’ (from 1961 to December 2020). (b) On key words of ‘machine learning and material’ and ‘machine learning and perovskite’ (from 2002 to December 2020).....	29
-------------	--	----

Figure.II.2	The basic process of sol-gel method.....	30
Figure.II.3	Type diffractometer: BRUCKER-D8 ADVANCE.....	35
Figure.II.4	Differential Thermal Analysis (DTA) / Thermogravimetric Analysis (TG) Device.....	37
Figure.II.5	FT-IR workflow.....	38
Figure.II.6	BET Surface Area Analyzer quantachrome .....	40
Figure.III.1	The general workflow of ML in perovskite materials.....	41
Figure.III.2	Commonly used machine learning algorithms in materials science.....	46

### **Chapter 03**

Figure.I.1	Weka graphic interface.....	58
Figure.I.2	The Explorer interface.....	59
Figure.I.3	The Explorer interfaces preprocess section.....	59
Figure.I.4	The Explorer interfaces classify section.....	60
Figure.I.5	The Explorer interface Select attributes section.....	60
Figure.I.6	The Explorer interface Visualize section.....	61
Figure.II.1	My DM and ML workflow.....	62
Figure.II.2	A histogram representing the selected features (attributes).....	75
Figure.II.3	A simple figure representing the method.....	78
Figure.II.4	A histogram representing the best performed features.....	78
Figure.II.5	The classify section in Weka where we will perform all the tasks.....	81
Figure.II.6	Choosing Remove percentage Filter.....	82
Figure.II.7	Configuring remove percentage filter.....	82
Figure.II.8	The training dataset with 40 samples (instances).....	83
Figure.II.9	Creating the testing set with 10 samples (instances).....	83



Figure.II.10	Selecting the algorithm.....	<b>84</b>
Figure.II.11	Different metrics evaluation for testing the model performance.....	<b>84</b>
Figure.II.12	Testing the model on the testing set.....	<b>85</b>
Figure.II.13	A plot visualizing classifiers error on the training set for the y (predicted SSA) and x (actual SSA) using the visualize tool JMathtools.....	<b>86</b>
Figure.II.14	Scatter plot representing Actual SSA vs. predicted SSA for perovskite samples by SVR model using Origin software.....	<b>88</b>
Figure.II.15	A plot visualizing classifiers error on the training set for the y (predicted SSA) and x (actual SSA) for the second model using the visualize tool JMathtools.....	<b>89</b>
Figure.II.16	A Scatter plot representing Actual SSA vs. predicted SSA for perovskite samples by SVR second model using Origin software.....	<b>91</b>
Figure.II.17	5-folds CV and LOOCV tests on Weka.....	<b>91</b>
Figure.II.18	The removed values of SSA in Weka.....	<b>93</b>
Figure.II.19	Preparing the model for predictions.....	<b>94</b>

## **Chapter 04**

Figure.II.1	Synthesis Flowchart of (a). $\text{LaFeO}_3$ , (b). $\text{LaMgO}_3$ , (c). $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ , $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$ , and $\text{LaFe}_{0.7}\text{Mg}_{0.3}\text{O}_3$ .....	<b>98</b>
Figure.III.1	XRD pattern of the four samples.....	<b>99</b>
Figure.III.2	TGA analysis curve for A ( $\text{LaFeO}_3$ ), B ( $\text{LaMgO}_3$ ), C ( $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ ), D ( $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$ ).....	<b>100</b>
Figure.III.3	FTIR analyse for the five samples.....	<b>101</b>

# List of Tables

## Chapter 2

Table.II.1	Comparative study of the various methods of synthesis.....	33
Table.III.1	Publicly accessible databases of various materials.....	43

## Chapter 3

Table.I.1	The four perovskite-type materials screened out using Li Shi, Dongping Chang, Xiaobo Ji, and Wencong Lu.Journal model.....	57
Table.II.1	The dataset complete with 50 samples (instances) and 24 candidate features.....	63
Table.II.2	The list of 24 candidate features.....	67
Table.II.3	Feature selection results using Wrapper method.....	74
Table.II.4	Feature selection results with 10 folds cross validation results.....	75
Table.II.5	The reduced data well normalized and ready.....	77
Table.II.6	The normalized dataset for the new selected feature.....	80
Table.II.7	The selected features in the two datasets.....	81
Table.II.8	The predicted values and the actual values of the SSA of the samples....	87
Table.II.9	The predicted values and the actual values of the SSA of the samples for the second model.....	90
Table.II.10	A simple table representing all the results of the evaluation of the first and the second model.....	92
Table.II.11	The results of the prediction of the second SVR model.....	94

## Chapter 4

Table.III.1	Specific surface area of the synthesized and the collected samples.....	102
-------------	---	-----

The background features a large, abstract geometric design on the left side, consisting of a black triangle pointing downwards and a blue triangle pointing upwards, meeting at a diagonal line. The rest of the page is white with a pattern of light blue and grey hexagons and lines, some containing small blue dots, resembling a molecular or network structure.

# General

## Introduction

We are overwhelmed with data, the amount of data in the world and in our lives seems to be constantly increasing, and there is no end in sight. Pervasive computers make it too easy to save things we would have previously thrown away. Cheap disks and online storage make it too easy to postpone decisions on what to do with it all - we just get more memory and keep it all. Pervasive electronics record our decisions, our choices at the supermarket, our financial habits, our comings and goings. We go around the world, each time a record passes through a database. The World Wide Web (WWW) overwhelms us with information; meanwhile, every choice we make is recorded. And all of these are just personal choices - they have countless counterparts in the world of commerce and industry. We could all witness the growing gap between the generation of data and our understanding of it.

As the volume of data grows, inexorably the proportion of it that people understand drops alarmingly. All this data hides potentially useful information that is rarely explained or exploited. There is nothing new about finding patterns in data; people have been looking for patterns in data since the very beginning of human life. Hunters look for patterns in the migration behaviour of animals, farmers look for patterns in the growth of crops, politicians look for patterns in voter opinion, and lovers look for patterns in the responses of their partners. A scientist's job (like a child's) is to make sense of data, discover the patterns that govern how the physical world works, and encapsulate them into theories that can be used to predict what will happen in new situations. The job of the entrepreneur is to identify opportunities, that is, patterns of behaviour that can be transformed into a profitable business and to exploit them.

In data mining, data is stored electronically and research is automated, or at least augmented, by the computer. It's not particularly new, either. Suppose, to take a well-established example, that the problem is inconsistent customer retention in a highly competitive market. A database of customer choices, along with customer profiles, holds the key to this problem. Behavioural patterns of past customers can be analyzed to identify the distinguishing characteristics of those who can switch and those who can remain loyal. Once these characteristics are identified, they can be leveraged to identify current customers who are likely to abandon ship. This group may be targeted for special treatment, treatment that is too costly to apply to all customers. More positively, the same techniques can be used to identify customers who might be attracted to another service provided by the company, which they are not currently using, to direct them to special offers promoting that service. In today's

highly competitive, customer-centric and service-driven economy, data is the raw material that fuels business growth, if only it can be harnessed. [1]

Now let us talk about machine learning, According to Charles Green, the Director of Thought Leadership at Belatrix Software: “It’s a huge challenge to find data scientists, people with machine learning experience, or people with the skills to analyze and use the data, as well as those who can create the algorithms required for machine learning. Secondly, while the technology is still emerging, there are many ongoing developments. It’s clear that AI is a long way from how we might imagine it.” [2]

According to Arthur Samuel machine learning is a subfield of computer science that gives computers the ability to learn without being explicitly programmed [4] although not directly mentioned in Arthur Samuel's definition, a key feature of machine learning is the concept of self-learning. This refers to the application of statistical models to recognize patterns and improve performance based on data and empirical information; all without programming instructions directly. This is what Arthur Samuel described as the ability to learn without being explicitly programmed. But it doesn't mean machines make decisions without initial programming. In contrast, machine learning relies heavily on computer programming. Instead, Samuel observed in that the machines do not need a direct input command to perform a specified task, but rather input data. [3]

In recent years, data mining has been widely used to solve problems in chemical, material, and engineering processes based on data collected from experiments or simulations. Mentioning Predicting Selective Compounds using Machine learning models [5] Accelerated search for perovskite materials with higher Curie temperature based on the machine learning methods [6] search for materials with targeted properties by adaptive design [7] and in many worldwide pressing issues, such as greenhouse gas capture , catalytic materials design and optimization, and renewable energy studies. Data mining has shown predictive power to extract relationships between intrinsic and extrinsic properties. Typically, the task of a data mining process is to predict (or generate) those variables that are difficult to capture from experiments/simulations, using the simple variables that can be captured as inputs. A well-fitting nonlinear form allows predicted variables to be generated quickly from the inputs of these independent variables. In other words, a data mining process powered by machine learning can accelerate:

- (i) The optimization of technical processes,
- (ii) The discovery of new functional materials,
- (iii) The understanding of chemical processes.

Despite a number of studies published over the past decade, there is no established philosophy that provides a standard guideline for performing data mining. A general and simple but useful data-mining strategy for these scientific and application processes will ultimately benefit to the standard development of knowledge-based data-mining through a machine learning modelling process. [8]

In this work a predictive model will be created on the basis of machine learning and data mining methods by providing already established data base filled with  $ABO_3$  perovskite-type materials samples and try to predict the higher specific surface area of these samples.

The first chapter represent a simple Bibliographic Study about machine learning and data mining thus a reminder about perovskite-type materials.

The second chapter talk about how to apply Data mining and Machine learning methods in the field of Materials Discovery and Optimization plus mentioning some the traditional ways to develop materials.

The third chapter consist of the workflow of data mining and machine learning techniques and an explanation about the data mining software Weka which been used to execute these techniques besides the results and the interpretations.

The forth chapter contain the synthesis of five of the perovskite samples (3 samples were from the collected data and two samples were screened out using Li Shi, Dongping Chang, Xiaobo Ji, and Wencong Lu. Journal model) via Sol-Gel method and characterize them using XRD, TGA analyze, Fourier transforms infrared spectroscopy analysis (FTIR), and Specific area measurement by the BET method.

## Bibliography

### References:

- 1 Witten, I. H., Frank, E., & Mark, A. Hall, 2011, Data Mining Practical Machine Learning Tools and Techniques, Burlington MA 01803 USA.
- 2 Charles Green. BBC, Will A Robot Take My Job? 2015, (<http://www.bbc.com/news/technology-34066941>).
- 3 Theobald, O. (2017). Machine learning for absolute beginners: a plain English introduction (Vol. 157). Scatterplot press.
- 4 Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. IBM Journal of research and development, 44(1.2), 206-226.
- 5 Ning, X., Walters, M., & Karypisxy, G. (2012). Improved machine learning models for predicting selective compounds. Journal of chemical information and modeling, 52(1), 38-50.
- 6 Zhai, X., Chen, M., & Lu, W. (2018). Accelerated search for perovskite materials with higher Curie temperature based on the machine learning methods. Computational Materials Science, 151, 41-48.
- 7 Xue, D., Balachandran, P. V., Hogden, J., Theiler, J., Xue, D., & Lookman, T. (2016). Accelerated search for materials with targeted properties by adaptive design. Nature communications, 7(1), 1-9.
- 8 Li, H., Zhang, Z., & Zhao, Z. Z. (2019). Data-mining for processes in chemistry, materials, and engineering. Processes, 7(3), 151.

# Chapter 01

## Bibliographic Study





## I. What is Machine learning

### I.1.Introduction

The term machine learning refers to the automated recognition of meaningful patterns in data. Over the past two decades, it has become a common tool for almost any task that requires extracting information from large datasets. We are surrounded by technology based on machine learning: search engines learn how to give us the best results (while placing profitable ads), anti-spam software learns to filter our emails, and credit card transactions are processed by software protected, the learns to recognize fraud. Digital cameras are learning to recognize faces, and smart personal assistant apps on smart phones are learning to recognize voice commands. Cars are equipped with accident prevention systems built using machine learning algorithms. Machine learning is also commonly used in scientific applications such as bioinformatics, medicine, and astronomy. A common feature of all these applications is that, in contrast to more traditional uses of computers, in these cases, due to the complexity of the patterns to be recognized, a human programmer cannot provide an explicit and detailed specification of how such tasks must be performed. Using intelligent beings as an example, many of our skills are acquired or honed by learning from our experience (rather than following explicit instructions given to us). Machine learning tools are concerned with giving programs the ability to "learn" and adapt. [1]

### I.2.General definitions

#### I.2.1.Machine learning

Machine learning is about extracting knowledge from data. It is a research field at the intersection of statistics, artificial intelligence, and computer science and is also known as predictive analytics or statistical learning. [2]The goal of Machine Learning (ML) is to construct computer programs that can learn from data. ML approaches can be distinguished in terms of representation and adaptation. A machine learning system needs to store the learned information in some knowledge representation structure which is called (an inductive) hypothesis and is typically of the form of a model. The hypothesis should generalize the training data giving preference for the simplest hypothesis; to obtain valid generalization, the hypothesis should be simpler than the data itself. A learning algorithm specifies how to update

the learned hypothesis with new experience such that the performance measure with regard to the task is being optimized. [1]

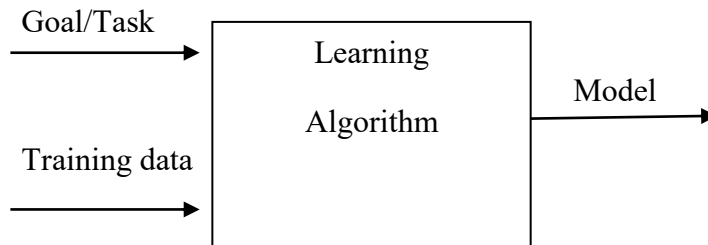


Figure.I.1. A generic machine learning method. [1]

## I.2.2. Machine learning categories

Machine learning incorporates several hundred statistical-based algorithms and choosing the right algorithm or combination of algorithms for the job is a constant challenge for anyone working in this field. It is important to understand the three overarching categories of machine learning, supervised, unsupervised, and reinforcement.

### I.2.2.1. Supervised learning:

As the first branch of machine learning, supervised learning focuses on learning patterns by connecting the relationship between variables and known outcomes and working with labeled datasets. Supervised learning works by feeding the machine pattern data with various functions (represented as "X") and returning the correct value of the data (represented as "y"). The fact that the output and feature values are known qualifies record as tagged. The algorithm then decodes the patterns present in the data and creates a model that can reproduce the same underlying rules with new data.

After the machine deciphers the rules and patterns of the data, it creates what is known as a model: an algorithmic equation for producing an outcome with new data based on the rules derived from the training data. Once the model is prepared, it can be applied to new data and tested for accuracy. After the model has passed both the training and test data stages, it is ready to be applied and used in the real world.

Examples of supervised learning algorithms include regression analysis, decision trees,  $k$ -nearest neighbors, neural networks, and support vector machines. [3]

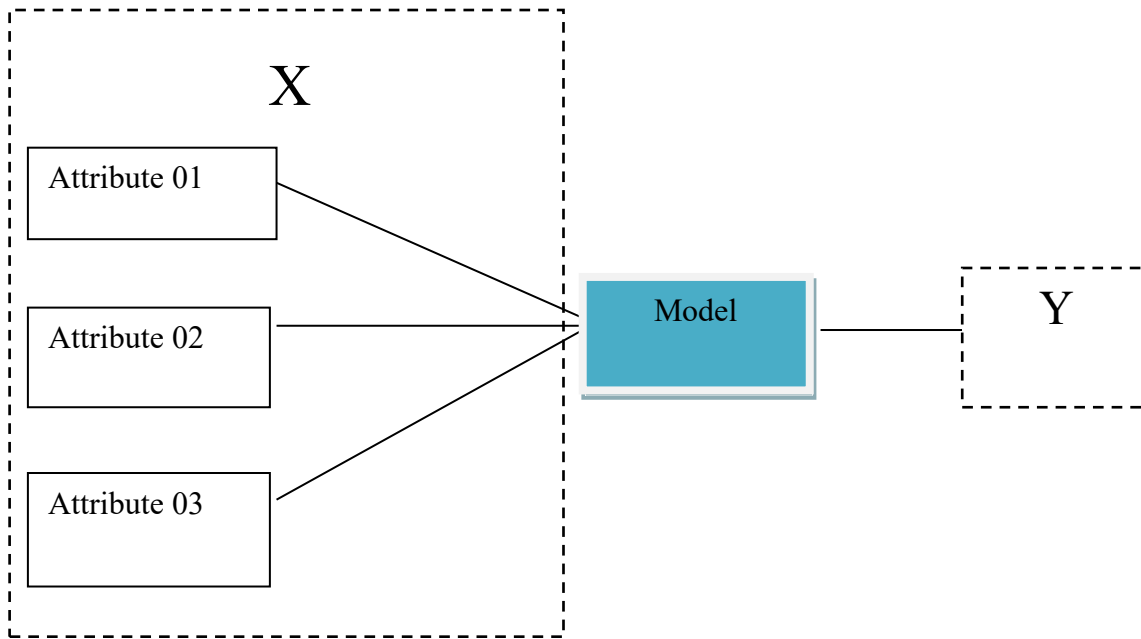


Figure.I.2. example of a prediction model [3]

#### I.2.2.2.Unsupervised learning:

In the case of unsupervised learning, not all variables and data patterns are classified. Instead, the machine should uncover hidden patterns and make labels through the employment of unsupervised learning algorithms.

The advantage of unsupervised learning is it enables you to discover patterns in the data that you were unaware existed. [3]

#### I.2.2.3.Reinforcement Learning

Reinforcement learning is that the third and most advanced algorithm category in machine learning. In contrast to supervised and unsupervised learning, reinforcement learning endlessly improves its model by leverage feedback from previous iterations. [3]

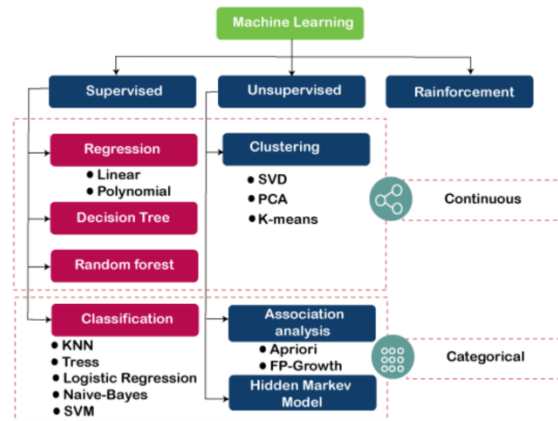


Figure.I.3. simple diagram illustrates the different ML algorithm, along with the categories [4]

## I.2.3. Machine learning tasks

### I.2.3.1. Classification and Regression

The tasks of classification and regression deal with the prediction of the value of one field (the target) based on the values of the other fields (attributes or features). If the target is discrete (e.g. nominal or ordinal) then the given task is called classification. If the target is continuous, the task is called regression. Classification or regressions normally are supervised procedures: based on a previously correctly labeled set of training instances, the model learns to correctly label new unseen instances. [1]

#### Classification:

Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters. In Classification, a computer program is trained on the training dataset and based on that training; it categorizes the data into different classes. The task of the classification algorithm is to find the mapping function to map the input(x) to the discrete output(y).

Classification Algorithms can be further divided into the following types:

- ❖ Logistic Regression
- ❖ Support Vector Machines
- ❖ Kernel SVM
- ❖ Decision Tree Classification

## Regression:

Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables. The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).

Regression Algorithms can be further divided into the following types:

- ❖ Simple Linear Regression
- ❖ Polynomial Regression
- ❖ Support Vector Regression
- ❖ Decision Tree Regression

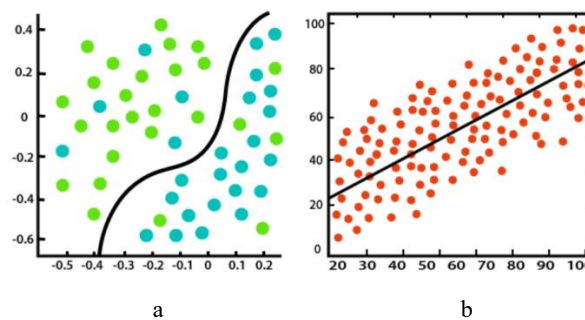


Figure.I.4. simple diagram representing (a) classification, (b) regression [4]

### I.2.4. Training and test data

In machine learning, data is split into training data and test data. The first split of data, the initial reserve of data you use to develop your model, provides the training data. After a successfully developed a model based on the training data and a satisfied accuracy, we'll be able to then test the model on the remaining data, referred to as the test data. [3]

### I.2.5. Models

Machine learning hypotheses might be available a spread of information representation forms, such as equations, decision trees, rules, distances and partitions, probabilistic and graphical models. Typical machine learning algorithms induce models that are hypotheses that characterize globally an entire data set. [1]

## I.2.6. Machine learning Algorithms

### I.2.6.1. Support Vector Machine Algorithm

Statistical learning theory was introduced within the late 1960's. Till the 1990's it absolutely was a strictly theoretical analysis of the matter of perform estimation from a given collection of data. Within the middle of the 1990's new kinds of learning algorithms (called support vector machines) based on the developed theory were proposed. [5]

A support vector machine or SVM is a supervised learning algorithm that can also be used for classification and regression problems. The goal of SVM is to create a hyperplane or decision boundary that can segregate datasets into different classes. The data points that help to define the hyperplane are known as support vectors, and hence it is named as support vector machine algorithm. [4] Consider the below diagram:

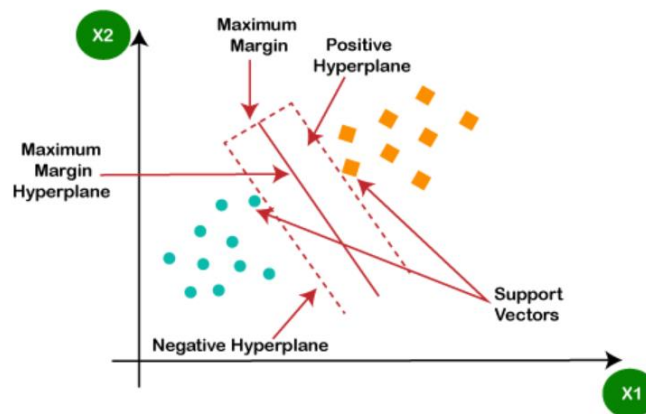


Figure.I.5. A diagram representing a classified dataset into two different classes.

### I.2.6.2. Methods of Support vector machine

The foundations of SVM are developed by Vapnik and are gaining popularity because of several enticing features, and promising empirical performance. The term SVM will refer to each SVC and SVR methods, which can be used for solving qualitative and quantitative issues respectively.

#### Support vector classification (SVC)

SVC has been recently planned as a really effective technique for resolution classification problems, which can be restricted to consideration of the 2-class problem while not loss of generality. During this problem the goal is to separate the two classes by a classifier evoked from available examples. It's expected that the classifier made has good performance on unseen examples, it generalizes well.

#### Support vector Regression (SVR)

In SVR, the basic idea is to map the data  $x$  into a higher dimensional feature space  $F$  via a nonlinear mapping  $U$  and then to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set

$$G = \{(x_i, d_i)\}_i^l$$

Where  $x_i$  is the input vector, and  $d_i$  is the desired value).SVR approximates the function in the following form:

$$y = \sum_{i=1}^l w_i \phi(x_i) + b$$

Where  $\{\phi(x_i)\}_{i=1}^l$  is the set of mappings of input features, and  $\{w_i\}_{i=1}^l$  and  $b$  are coefficients. They are estimated by minimizing the regularized risk function  $R(C)$ :

$$R(C) = C \frac{1}{N} \sum_{i=1}^N L_\varepsilon(d_i y_i) + \frac{1}{2} \|w\|^2$$

Where  $L_\varepsilon(d_i y_i) = |d - y| - \varepsilon$  for  $|d - y| \geq \varepsilon$

Otherwise  $L_\varepsilon(d_i y_i) = 0$  [6]

### I.2.6.3. Support vector machine principle

The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features  $x_1$  and  $x_2$ . We want a classifier that can classify the pair  $(x_1, x_2)$  of coordinates in either green or blue. So as it is 2D space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes.

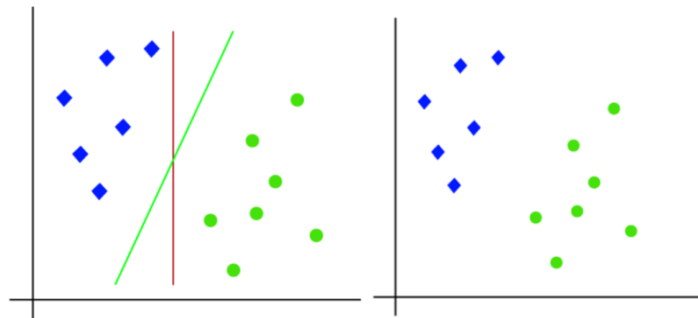


Figure.I.6. simple graphs representing a. 2 datasets and b. separated datasets into 2 classes

Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyperplane. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as margin. And the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane. On consider the image below:

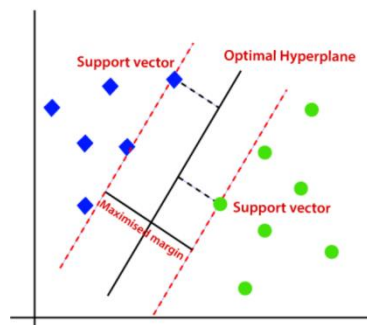


Figure.I.7. The best boundary (hyper plane) found by the SVM algorithm



If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. So to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions  $x$  and  $y$ , so for non-linear data, we will add a third dimension  $z$ . It can be calculated as:  $z^2 = x^2 + y^2$  [4]

### I.2.7. Under-fitting and Over-fitting: Problems of Machine Learning

According to statistical learning theory, the machine learning is a process of choosing an appropriate function from a given set of functions to correlate the data set. The set of functions used is called hypothesis functions or indicator functions\*. For example, in the process of linear regression or linear separation for different classes of samples, all linear functions are used as hypothesis functions. Since the appropriate function has to be chosen from the hypothesis functions only, the mathematical model built by using machine learning is always constrained within the scope of hypothesis functions used. For example, if a linear regression method is used as learning process, the mathematical model found shall be surely linear one, even if the actual data set exhibits some nonlinearity, because this nonlinearity has been treated as noise or residue and eliminated in the process of machine learning. [7]

If our algorithm works well with the training dataset but not well with test dataset, then such problem is called Over-fitting. And if our algorithm does not perform well even with training dataset, then such problem is called under-fitting. [4]

## I. What is Data mining

### II.1.Introduction

The trendy technologies of computers, networks, and sensors have created data collection and organization an almost easy task. However, the captured data has to be converted into information and knowledge from recorded data to become useful. Traditionally, the task of extracting useful information from recorded data has been performed by analysts; however, the increasing volume of data in modern businesses and sciences demand computer-based strategies for this task.

As data sets have grown in size and complexity, thus there had been an inevitable shift far from direct hands-on data analysis toward indirect, automatic data analysis during which the analyst works via a lot of complicated and complex tools. The whole method of applying computer based methodology, as well as new techniques for information discovery from data, is commonly known as data mining.

The importance of data mining arises from the very fact that the modern world is a data-driven world. We are encircled by data, numerical and otherwise, that must be analyzed and processed to convert it into information that informs, instructs, answers, or otherwise aids understanding and decision-making.

The new discipline of data mining has developed particularly to extract valuable information from such large data sets. In recent years there has been an explosive growth of strategies for locating new knowledge from raw data. this is be not shocking given the proliferation of low-priced computers (for implementing such methods in software), low-cost sensors, communications, and database technology (for collecting and storing data) and highly computer-literate application specialists who can create “interesting” and “useful” application problems.

Data mining isn't a brand new technology. The idea of extracting information and knowledge discovery from recorded data is a well-established concept in scientific and medical studies. What's new is that the convergence of many disciplines and corresponding technologies that have created a singular chance for data mining in scientific and corporate world. [8]

## II.2. General definition

It is no surprise that data mining, as a very knowledge base subject, may be outlined in many alternative ways. Even the term data mining doesn't really present all the foremost parts within the picture. To refer to the mining of gold from rocks or sand, we are saying gold mining rather than rock or sand mining.

Analogously, data mining should have been more appropriately named “knowledge mining from data,” which is unfortunately somewhat long. However, the shorter term, knowledge mining may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer carrying both “data” and “mining” became a popular choice. In addition, many other terms have a similar meaning to data mining for example, knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging [9]

In summary Data mining is an essential process where intelligent methods are applied to extract data patterns.

## II.3. Data Mining Tasks

Data mining tasks are generally divided into two main groups: Descriptive tasks data mining and predictive data mining.

**Predictive tasks:** The goal is to predict the value of a specific attribute based on the values of other attributes. The attribute to be predicted is called the target or dependent variable, the attributes used to make the prediction are known as the independent variables.

**Descriptive tasks:** The goal here is to obtain patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in the data. Descriptive data mining tasks are often exploratory in nature and often require post-processing techniques to validate and explain the results.

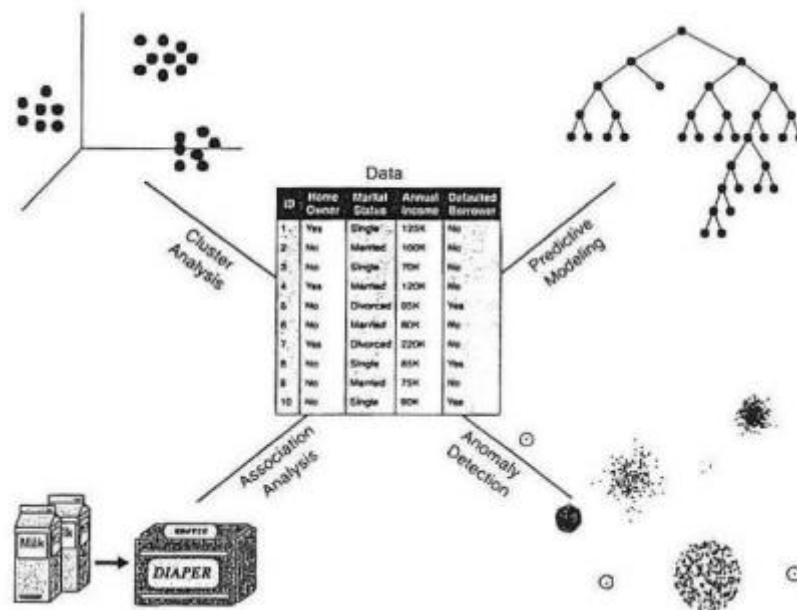


Figure.II.1: Four of the core data mining tasks.

**Predictive modeling** describes the task of building a model for the target variable using the independent variables. There are two types of predictive modeling tasks: classification, and regression the first is used for discrete target variables, and the last is used for continuous target variables.

**Association analysis** used to discover patterns that describe features associated with strength in the data. The patterns discovered are typically presented in the form of implication rules or subsets of features. Because of the exponential size of your search space, the goal of association analysis is to efficiently extract the most interesting patterns.

**Cluster analysis** attempts to find closely related groups of observations called clusters, such that observations belonging to the same clusters are more similar to each other than observations belonging to different groups.

**Anomaly detection** the task is to identify observations whose characteristics differ significantly from the rest of the data. These observations are called anomalies or outliers. The goal of an anomaly detection algorithm is to discover the actual anomalies and avoid falsely flagging normal objects as anomalous. [10]

## II.4. Data Mining Process

The data mining process is a pipeline containing many phases such as data cleaning, feature extraction, and algorithmic design. In this section, we will study these different phases. The workflow of a typical data mining application contains the following phases (Fig.II.2).

### The Data collection phase:

Data collection may require the use of specialized hardware or software tools like document crawling engine for collecting documents. Although this phase is very application specific and often outside the purview of the data mining analyst, it is critical because good decisions at this phase can significantly impact the data mining process. After the collection phase, the data is often stored in a database or more generally in a data warehouse for processing.

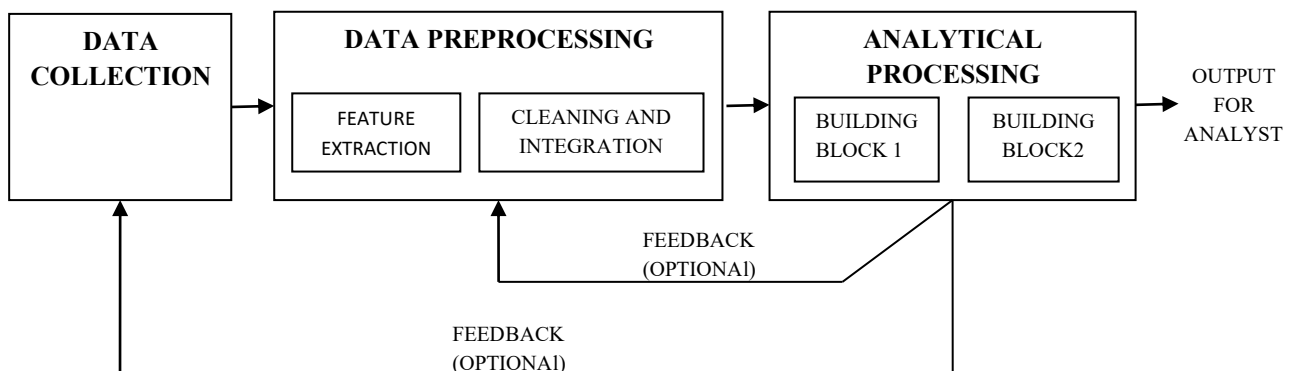


Figure.II.2: The data processing pipeline

### The Data Pre-processing Phase:

The data pre-processing phase is perhaps the most crucial one in the data mining process. Yet, it is rarely explored to the extent that it deserves because most of the focus is on the analytical aspects of data mining. This phase begins after the collection of the data, and it consists of the following steps:

- **Feature extraction:** An analyst may be faced with large volumes of raw documents, or system logs, with little guidance on how to turn that raw data into meaningful database functions for processing. This phase relies heavily on the analyst's ability to abstract the features that are most relevant to a particular application. Therefore, extracting the right features requires an understanding of the specific application domain involved.
- **Data cleaning:** The extracted data may contain incorrect or missing entries. Therefore, it may be necessary to delete some records or estimate missing entries. Inconsistencies may need to be eliminated.
- **Feature selection and transformation:** When the data is of very large dimensions, many data mining algorithms do not work effectively. Also, many of the high-dimensional features are noisy and can add errors to the data mining process. Therefore, a variety of methods are used to remove irrelevant features or transform the current set of features into a new data space that is more suitable for analysis. Another related aspect is data transformation, where a data set with a specific set of attributes can be transformed into a data set with a different set of attributes of the same or different type.

### The Analytical Phase:

A major challenge is that each data mining application is unique, making it difficult to create common and reusable techniques across applications. However, many data mining formulations are used over and over again in the context of different applications. These correspond to the most important "super problems" or building blocks of the data mining process.

## II.5.The Basic Data Types

One of the interesting aspects of the data mining process is the wide variety of data types that are available for analysis. There are two broad types of data, of varying complexity, for the data mining process:

- Nondependency-oriented data: This typically refers to simple data types such as multidimensional data or text data. These data types are the simplest and most commonly encountered. In these cases, the data records do not have any specified dependencies between either the data items or the attributes.
- Dependency-oriented data: In these cases, implicit or explicit relationships may exist between data items, like web graphs, Chemical compound databases...etc. In general, dependency-oriented data are more challenging because of the complexities created by preexisting relationships between data items. Such dependencies between data items need to be incorporated directly into the analytical process to obtain contextually meaningful results. [11]

### III. Relationship between Data mining and Machine Learning

Relevant disciplines can also be difficult to tell apart at first glance, such as “machine learning” and “data mining.” Machine learning, data mining, computer programming, and most relevant fields (excluding classical statistics) derive first from computer science, which encompasses everything related to the design and use of computers. Within the all-encompassing space of computer science is the next broad field: data science. Narrower than computer science, data science comprises methods and systems to extract knowledge and insights from data through the use of computers.

As mentioned, machine learning also overlaps with data mining. A popular algorithm, such as *k*-means clustering, association analysis, and regression analysis, are applied in both data mining and machine learning to analyze data. But where machine learning focuses on the incremental process of self-learning and data modelling to form predictions about the future, data mining narrows in on cleaning large datasets to glean valuable insight from the past. Both data mining and machine learning appear similar, and they do use many of the same tools. [3]

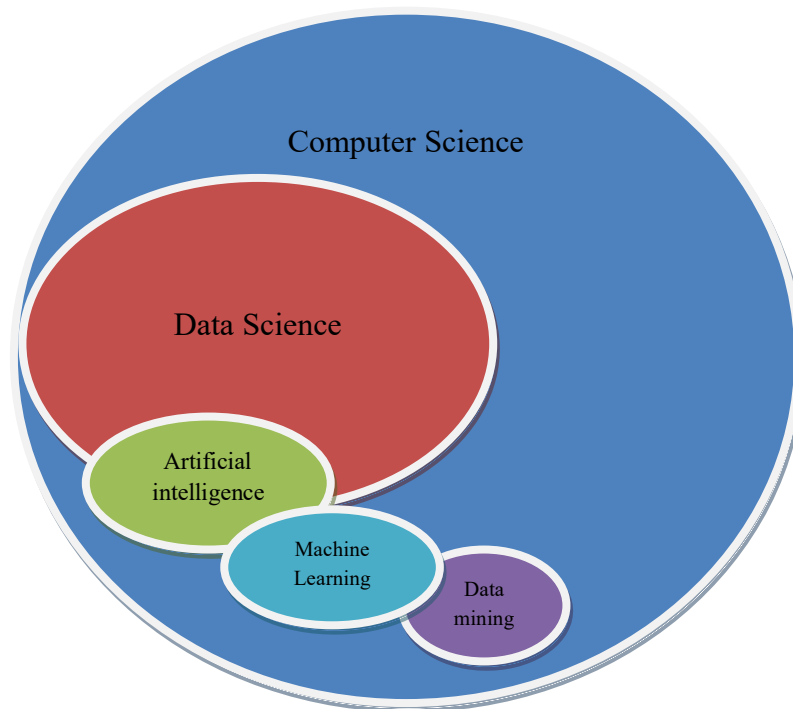


Figure.III.1: Visual representation of the relationship between data-related fields



## IV. Perovskite

### IV.1. Introduction

Discovered by a Russian scientist, Gustav Rose, in 1839 in mineral deposits in the Ural Mountains and the research was conducted by the Russian mineralogist Lev Perovski, for whom this mineral was named perovskite. Perovskites are a class of compounds with similar structures to the mineral perovskite,  $\text{CaTiO}_3$ , and can be viewed as being derived from a parent phase with the general formula  $\text{ABX}_3$ . They have been intensively studied since the middle of the 20th century because of their innate properties: initially dielectric, piezoelectric and ferroelectric. This range of behaviour has been extended to areas such as magnetic ordering, multiferroic properties, electronic conductivity, superconductivity, and thermal and optical properties. In addition to these purely physical aspects, the phases have a large number of chemical properties. Many perovskite phases exhibit useful redox and catalytic behaviour, often dependent on the presence of chemical defects in the phase. The importance of perovskite became apparent when the valuable dielectric and ferroelectric properties of barium titanate,  $\text{BaTiO}_3$ , were discovered in the 1940s. This material quickly found applications in electronics in the form of capacitors and transducers. In the following decades, attempts to improve the material properties of  $\text{BaTiO}_3$  led to intensive research into the structure-property relationships of a large number of phases related to the nominally ionic ceramic perovskite with the general composition  $\text{ABO}_3$ , with the result that a large number of new phases developed synthesized <sup>[12]</sup>

## IV.2. Perovskite ideal structure

The ideal  $ABX_3$  perovskite structure is described together with some of the structural variations that occur which have significance for chemical and physical properties and which make precise structure determination a difficult task. [12]

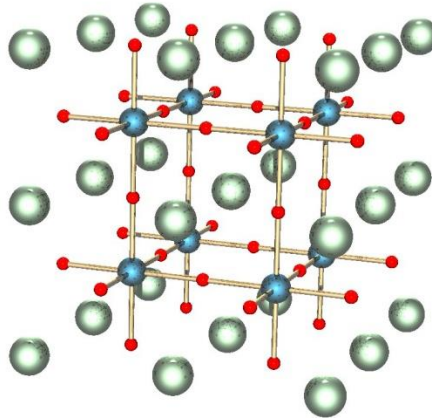


Figure.IV.1: Structure of a perovskite with general chemical formula  $ABX_3$ . The red spheres are X atoms (usually oxygens), the blue spheres are B atoms (a smaller metal cation, such as  $Ti^{4+}$ ), and the green spheres are the A atoms (a larger metal cation, such as  $Ca^{2+}$ ). [14]

Generally we are talking about a considerable number of mixed oxides which are referred to as perovskite oxide distinguish by a unique formula  $ABO_3$  where A represent a large radius cation with a coordination number 12 (e.g. Ca, Pb, Rb, Sr, Na, K...), B a lower radius cation, higher load with a coordination number 6 (e.g. Ti, Sn, W, Zr, Nb, Ta, ...), and finally the oxygen ion. [13]

The idealized perovskite structure is cubic and is adopted by  $SrTiO_3$  at room temperature.

There are two general ways of listing the atoms in the cubic unit cell.

- ✓ The standard crystallographic description places the choice of origin at the Sr atom:

$SrTiO_3$ : cubic;  $a = 0.3905$  nm,  $Z = 1$ ; space group,  $Pm\bar{3}m$

Atoms positions:

- Sr : 1(a) 0, 0, 0
- Ti : 1(b)  $\frac{1}{2}$ ,  $\frac{1}{2}$ ,  $\frac{1}{2}$
- O : 3(c)  $\frac{1}{2}$ ,  $\frac{1}{2}$ , 0 ;  $\frac{1}{2}$ , 0,  $\frac{1}{2}$  ; 0,  $\frac{1}{2}$ ,  $\frac{1}{2}$

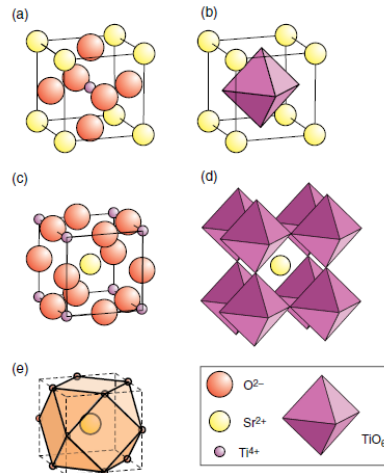


Figure.IV.2: The idealised perovskite structure of  $\text{SrTiO}_3$ : (a) atom positions with  $\text{Sr}^{2+}$  at cell origin; (b)  $\text{TiO}_6$  octahedral coordination polyhedron; (c) atom positions with  $\text{Ti}^{4+}$  at cell origin; (d)  $\text{TiO}_6$  octahedral polyhedron framework with  $\text{Sr}^{2+}$  at the cell centre; (e) cuboctahedral cage site [15]

The  $\text{Sr}^{2+}$  ions lie at the corners of the unit cell. The  $\text{Ti}^{4+}$  ions lie at the cell centre and are surrounded by a regular octahedron of  $\text{O}^{2-}$  ions (Figure a and b). For some purposes it is useful to translate the cell origin to the  $\text{Ti}^{4+}$  ions:

Atoms positions:

- Ti : 1(a) 0, 0, 0
- Sr : 1(b)  $\frac{1}{2}$ ,  $\frac{1}{2}$ ,  $\frac{1}{2}$
- O : 3(d)  $\frac{1}{2}$ , 0, 0 ; 0,  $\frac{1}{2}$ , 0 ; 0, 0,  $\frac{1}{2}$

The large  $\text{Sr}^{2+}$  ions are coordinated to 12  $\text{O}^{2-}$  ions and are now situated at the unit cell centre (Figure c). For a discussion of the chemical and physical properties of this (and other) perovskite, it is convenient to think of the structure as built-up from an array of corner sharing  $\text{TiO}_6$  octahedra (Figure d). The large  $\text{Sr}^{2+}$  ions are located at the unit cell centre and are surrounded by a cuboctahedral cage of  $\text{O}^{2-}$  ions (Figure e).

The  $\text{TiO}_6$  framework is regular and the octahedra are parallel to each other. All the  $\text{Ti}^{4+}\text{—O}^{2-}$  bond lengths are equal and the six  $\text{O}^{2-}\text{—Ti}^{4+}\text{—O}^{2-}$  bonds are linear.

### IV.3. Tetragonal perovskite

The best known example of a tetragonal perovskite is probably the room temperature form of the ferroelectric BaTiO<sub>3</sub>, with  $a = 3.944 \text{ \AA}$ , and  $c = 4.038 \text{ \AA}$  and  $Z = 1$ . In this case the TiO<sub>6</sub>-octahedra are somewhat distorted (one Ti-O bond at  $1.86 \text{ \AA}$ , four at  $2.00 \text{ \AA}$  and one longer at  $2.17 \text{ \AA}$ ). Barium is coordinated by four oxygens at  $2.80 \text{ \AA}$ , four at  $2.83 \text{ \AA}$  and four more at  $2.88 \text{ \AA}$ . A number of other tetragonal perovskites (PbHfO<sub>3</sub>, SrPbO<sub>3</sub>, SrZrO<sub>3</sub>, AgTaO<sub>3</sub>, KCoF<sub>3</sub>, CsPbCl<sub>3</sub>, CsPbBr<sub>3</sub>, etc.) are isotypic with BaTiO<sub>3</sub> and possess unimolecular cells.

### IV.4. Rhombohedral perovskites

In several materials the cubic cell may have a small deformation to rhombohedral symmetry. If this deformation does not enlarge the unit cell, it is possible to index it on a unit cell containing either one or two formula units with rhombohedral angles  $\alpha \sim 90^\circ$  or  $\alpha \sim 60^\circ$  respectively. However, the anions are generally displaced as requires the larger unit cell with  $\alpha \sim 60^\circ$ . Examples of rhombohedral perovskites are LaAlO<sub>3</sub>, PrAlO<sub>3</sub>, LaNiO<sub>3</sub> and LaCoO<sub>3</sub>.

### IV.5. Orthorhombic perovskites

The GdFeO<sub>3</sub> structure is probably the most common of all the orthorhombically distorted perovskites. Its space group is Pbnm and the cell constants are:  $a = 5.346 \text{ \AA}$ ,  $b = 5.616 \text{ \AA}$  and  $c = 7.666 \text{ \AA}$  with  $Z = 4$ . These constants are related to the cubic pseudocell  $a'$  by  $a \sim b \sim \sqrt{2} a'$  and  $c \sim 2a'$

In this structure the FeO<sub>6</sub> octahedra are distorted and tilted. Also the GdO<sub>12</sub>-polyhedron is severely distorted, showing (8+4) coordination.

Other materials adopting this orthorhombic-distorted structure are NaUO<sub>3</sub>, NaMgF<sub>3</sub>, LaYbO<sub>3</sub> and a great number of lanthanide compounds of the type LnCrO<sub>3</sub>, LnGaO<sub>3</sub>, LnFeO<sub>3</sub>, LnMnO<sub>3</sub>, LnRhO<sub>3</sub>, etc.

### IV.6. Monoclinic and triclinic perovskites

Monoclinic (AgCuF<sub>3</sub>, CsPbI<sub>3</sub>, PbSnO<sub>3</sub>, BiCrO<sub>3</sub>, etc.) or triclinic (BiMnO<sub>3</sub>, BiScO<sub>3</sub>) unit cells have been reported in several cases. However, in many cases these cells have proved to be pseudocells of a real multiple cell. For example, GdFeO<sub>3</sub>-type phases have been frequently indexed on the bases of a monoclinic pseudocell with  $a \sim b \sim c \sim a'$  and  $B \sim 90^\circ$ .

## IV.7. General properties and application of perovskite materials

The  $ABX_3$  perovskites exhibit several interesting properties such as ferromagnetism, ferroelectricity, pyro- and piezoelectricity, superconductivity (both, classic and high  $T_c$ ), large thermal conductivity, fluorescence and catalytic activity.

### IV.7.1. Magnetic properties

A number of different and interesting magnetic properties have been reported for perovskite-like materials. In some of them, the outer d-electrons are localized and spontaneously magnetic, in some they are itinerant and spontaneously magnetic and in others, Pauli paramagnetism has been found. Which of these properties is stabilized depends on the number of d-electrons per transition metal B cation and the strength of the B-O-B interactions.

### IV.7.2. Optical properties

The measurement of optical properties has often been used to characterize perovskite materials and also to identify phase transitions. The electro optical properties of different oxide materials have also been analyzed.

### IV.7.3. Piezoelectricity

A piezoelectric material develops an electric polarization when it is mechanically stressed along an appropriate direction. In the converse effect, an applied electric field produces a mechanical distortion in the material. Among the 32 crystal classes, 11 are centrosymmetric and therefore do not possess polar properties. Of the 21 non-centrosymmetric classes, 20 of them exhibit piezoelectricity whereas the remaining one (the cubic class 432) has a set of symmetry elements that combine to exclude piezoelectric character. [15]

## Bibliography

References:

- 1 Ławrynowicz, A., & Tresp, V. (2014). Introducing machine learning. Perspectives on Ontology Learning; Lehmann, J., Voelker, J., Eds, 35-50.
- 2 Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc.".
- 3 Theobald, O. (2017). Machine learning for absolute beginners: a plain English introduction (Vol. 157). Scatterplot press.
- 4 Sonoo Jaywalk. (2011). (<https://www.javatpoint.com/machine-learning-algorithms>).
- 5 Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE transactions on neural networks, 10(5), 988-999.
- 6 Lu, W. C., Ji, X. B., Li, M. J., Liu, L., Yue, B. H., & Zhang, L. M. (2013). Using support vector machine for materials design. Advances in Manufacturing, 1(2), 151-159.
- 7 Chen, N. (2004). Support vector machine in chemistry. World Scientific.
- 8 Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.
- 9 Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
- 10 Tan, P. N., Steinbach, M., & Kumar, V. (2005). Chapter on Cluster Analysis. pdf. Introduction to data mining. Boston: Pearson Addison Wesley.
- 11 Aggarwal, C. C. (2015). Data mining: the textbook (Vol. 1). New York: springer.
- 12 Tilley, R. J. (2016). Perovskites: structure-property relationships. John Wiley & Sons.

- 13 M. Lebid. (2016). Study of the physico-chemical properties of oxides based on lanthanum, iron and magnesium.(Doctoral Thesis Univ. Biskra).Retrieved from [http://thesis.univ-biskra.dz/2350/1/Chimi\\_d1\\_2016.pdf](http://thesis.univ-biskra.dz/2350/1/Chimi_d1_2016.pdf)
- 14 Navrotsky, A. (1998). Energetics and crystal chemical systematics among ilmenite, lithium niobate, and perovskite structures. *Chemistry of Materials*, 10(10), 2787-2793.
- 15 Baran, E. J. (1990). Structural chemistry and physicochemical properties of perovskite-like materials. *Catalysis Today*, 8(2), 133-151.

# Chapter 02

Data mining and Machine  
learning methods For  
Materials Discovery and  
Optimization





## I. Introduction

The field of computational chemistry has become increasingly predictive in the 21st century, with activity in applications ranging from the development of catalysts for the conversion of greenhouse gases, to the discovery of materials for harvesting and storing energy, to computational drug design. Expected connection (with reasonable accuracy) even before it is made in the lab. [1]

Data-driven research (data mining) and machine learning (ML) have emerged as promising new drivers in chemistry and materials [2]. Recently, the materials community has put a renewed focus on collecting and organizing large datasets for research, materials design, and eventual application of statistical or "machine learning" techniques. For example, searching composite databases through density functional theory (DFT) calculations has been used to identify battery materials [3,4] to aid the design of metal alloys [5,6] and so much other applications.

These datasets presents a big opportunity to form and develop models with reasonable accuracy with the help of machine learning and data mining techniques Rather than manually designing and programming such models, such techniques produce predictive models by learning from a series of examples. Machine learning models have been shown to predict the properties of crystalline materials much faster than DFT [7-10], estimate properties that are difficult to access via other computational tools [11,12], and guide the search for new materials [13-17]. With the continued development of general-purpose data mining methods for many types of materials data [18-20] and the proliferation of material property databases [21], this emerging field of "materials informatics" is positioned to have a continued impact on materials design, Alexander Dunn [22]

This Chapter includes a discussion about how to apply Machine learning (ML) methods and Data mining (DM) techniques in Chemistry in the materials categories in the goal of finding and discovering new materials with promising properties, specifically perovskite like materials.

## II. Why Perovskite materials

Perovskite materials have attracted much attention in many scientific fields for the composition diversity, easily available synthetic conditions and a variety of attractive properties<sup>[23,24]</sup>. The  $ABO_3$ -type perovskite oxide has bit by bit become an exploration hotspot in trendy industrial catalysis and thermoelectricity for the manageable structure, outstanding stability and low cost. <sup>[25,26]</sup> Inorganic double perovskite has aroused associate interest in solar cells and light-emitting diodes thanks to adjustable photoelectrical properties. <sup>[27, 28]</sup>

There are so many papers and articles published about perovskite especially from the year 2013 since the solar cell where proposed plus the machine learning application in this field show an alarming increase (figure II.1) <sup>[29]</sup>

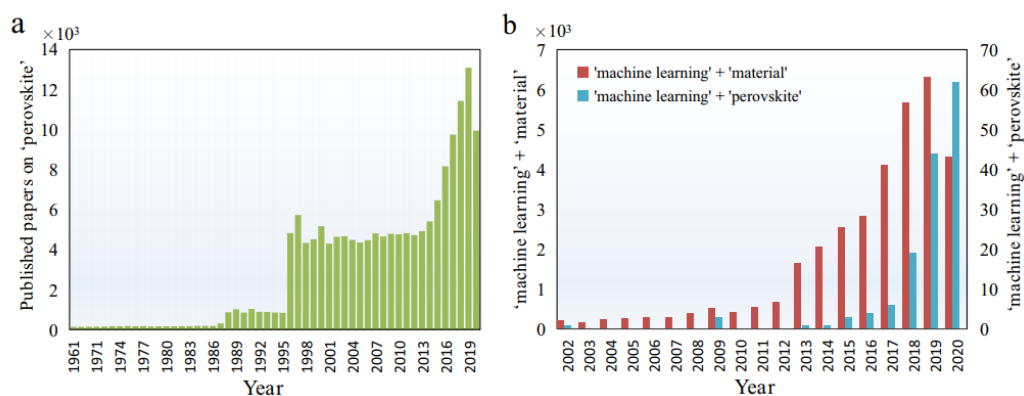


Figure II.1. Number of published papers. **(a)** On keyword of ‘perovskite’ (from 1961 to December 2020). **(b)** On key words of ‘machine learning and material’ and ‘machine learning and perovskite’ (from 2002 to December 2020)

## II.1. The traditional way to develop materials

The traditional way to develop materials is sometimes based on trial and error, continuous synthesis and characterization keep attempting till the properties of virtual materials meet the target. The method needs a long-time study on a restricted amount of materials and complex experimental procedures, which may be a long and high-ticket endeavor. Underneath this limitation, vital scientific progress typically comes from the researchers' expertise and intuition or maybe was discovered by accident. Plus the discovery of high-performance materials needs a long cycle from experimental design to commercialization. [29]

## II.2. Methods of synthesis and characterization of mixed oxides

### II.2.1. Methods of synthesis

Several methods have been described for the preparation of perovskite-type oxides:

**Sol-Gel method:** Sol-gel technique (S-G) is a method, for material preparation beneath delicate condition, of solidifying a compound containing an extremely chemically active component through a solution, sol, or gel, and then heat-treating an oxide or other compound. This highly chemically active component is used as a precursor uniformly mix these raw materials in the liquid phase, and perform hydrolysis and condensation chemical reactions to form a stable transparent sol system in solution. The sol slowly polymerizes between the aged colloidal particles to form a gel with a three-dimensional network structure. The gel network is filled with a solvent that loses fluidity to forma gel. The gel is dried, sintered and solidified to prepare molecular and even nano-substructure materials. The chemical process of the sol-gel method is to first disperse the raw materials in a solvent, and then undergo a hydrolysis reaction to form an active monomer. The active monomer is polymerized and begins to become a sol, and then a gel with a certain spatial structure is formed. After drying and heat treatment Preparation of required materials. [30]

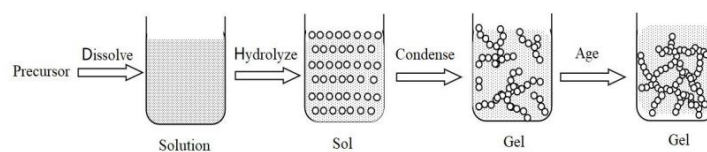


Figure.II.2. the basic process of sol-gel method

✓ **terminology:**

**Sol:** Sol is a colloidal system with liquid characteristics. The dispersed particles are solid or macro-molecules. The particle size is between 1 and 100 nm (somebody says 1-1000 nm), and the particles are evenly distributed in the dispersion medium.

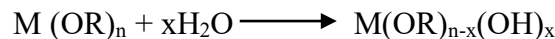
**Gel:** The colloidal particles or polymers in the sol or solution are connected to each other under certain conditions to form a spatial network structure, and the structural voids are filled with liquid as a dispersion medium in xerogel, it can also be gas, xerogel is also called aerogel, such a special dispersion system is called a gel. [30]

The basic reaction steps of the S-G method are as follows:

**Solvation:** The metal cation  $M^{z+}$  attracts water molecules to form the solvent unit  $M(H_2O)_n^{z+}$ . In order to maintain its coordination number, it has a strong tendency to release  $H^+$ .

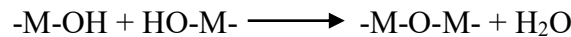


**Hydrolysis reaction:** Non-ionizing molecular precursors, such as metal alkoxide  $M(OR)_n$ , react with water :

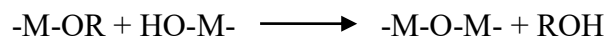


**Polycondensation reaction:** According to the type of molecules removed, it can be divided into two categories:

a) Dehydration polycondensation:



b) Dealcoholization polycondensation:



**Co-precipitation method:** The Co-precipitation synthesis method, proposed by Wackowski and his collaborators uses ammonium nitrate, added to the precursor solution of perovskite. The resulting product is decomposed at 300°C and then calcinated in oxygen at 500°C. Perovskites with specific surfaces of 30 m<sup>2</sup>/g are obtained in this way.

The precursors of sites A and B of the perovskite structure (acetate, chloride, and nitrate) are mixed in water. All species are then precipitated at basic pH in the form of oxalate or hydroxide, after the intermediate stages of settling, rinsing and filtration the precipitate undergoes a washing intended to break the agglomerates. The chemical qualities (stoichiometry, homogeneity) and physical qualities (particle size, grain shape) of these powders are good. The following parameters are of major significance:

- pH control
- stirring time
- Order of introduction of reagents into the basic solution.
- Room temperature control. [31]

**Solid-state reactions:** The synthesis of oxides (perovskite) by solid-state reactions is one of the most widely used methods in solid-state chemistry. The basis of this process is the reaction by heat treatment between two or more substances in solid form, which are first mixed. Reagents, oxides and/or carbonates in powder form are weighed in stoichiometric quantities and mixed intimately by grinding in a mortar.

Obtaining a homogeneous mixture of small particles then facilitates the kinetics of the reaction. The powder is then subjected to successive heat treatments until a single phase is obtained, the temperature being generally maintained at about 1000°C.

- a. **Raw materials:** They are composed of oxides, carbonates, nitrates, etc. An ideal powder can be described as consisting of small grains (of the order of 1  $\mu\text{m}$ ) of regular shape with a very narrow particle size distribution. Purity and possible additives are checked. The main problem of basic raw materials, which are in the form of powders, is the difficulty of assessing the fundamental parameters which reflect the reactivity of the material in relation to the others with which it reacts, the thermal history of the material, it therefore plays a very important role.
- b. **Mixing and grinding:** This is one of the essential phases of the production cycle of a solid with a perovskite structure. A uniform distribution of precursors is also obtained during this process. The powders are weighed according to the stoichiometric quantities given by the reaction equation.
- c. **Calcination:** The materials are subjected to a thermal cycle under controlled atmosphere during which, due to diffusion phenomena, they react in the solid phase

and form the desired phase. During this reaction, carbon dioxide or oxygen dioxide and water vapour are released.

- d. **Regrind**: The powder is ground again to reduce the granulometry, homogenize it and increase its reactivity. The powder is then subjected to a high temperature heat treatment to obtain the desired phases. [31]

Table.II.1. Comparative study of the various methods of synthesis.

Method	Advantages	Disadvantages
<b>Sol-Gel</b>	Flexible, dispersed homogeneous, technology mature	Solvent, carbon residues
<b>Co-precipitation</b>	High surfaces, low C contamination, thermal stability	Solvents, dependent method perovskite
<b>Solid-state reactions</b>	thermal stability	lower activity

## II.2.2.Characterization methods

### X-ray diffraction:

- Principle and equipment

X-ray diffraction is a tool to study the fine structure of matter. This technique originated with von Laue's discovery in 1912 that crystals diffract X-rays, revealing the diffraction shape of the crystal structure. At first, X-ray diffraction was only used for crystal structure determination. Later, other applications developed, and today the method is used not only for structure determination, but also for problems as diverse as chemical analysis and stress measurement, to study equilibria of phase and particle size measurement, to determine the orientation of a crystal, or the set of Orientations in a polycrystalline aggregate. [32]

The device used is a diffractometer BRUCKER-D8 ADVANCE which has a theta: theta geometry with a copper sealed tube x-ray source producing Cu  $k\alpha$  radiation at a wavelength of 1.5406 Å from a generator operating at 40 kV and 40 mA. A parallel beam of monochromatic x-ray radiation is produced by the use of a Göbel mirror optic (primary optic). The diffracted x-rays are recorded on a scintillation counter detector located behind a set of long Soller slits/parallel foils. The sample remains flat throughout the measurement but can be rotated to allow for better sampling and removal of preferred orientation effects.

This machine operates using a Copper Line Focus X-ray tube producing  $K\alpha$  radiation ( $K\alpha_1 = 1.540598$  Å,  $K\alpha_2 = 1.544426$  Å,  $K\alpha$  ratio 0.5,  $K\alpha_{\text{avg}} = 1.541874$  Å).

Data collections using detector scans at a grazing incidence angle of  $3^\circ$  were undertaken with a scan range from  $10$  to  $90^\circ$  at  $0.05^\circ$  step 8 s/step. [33]

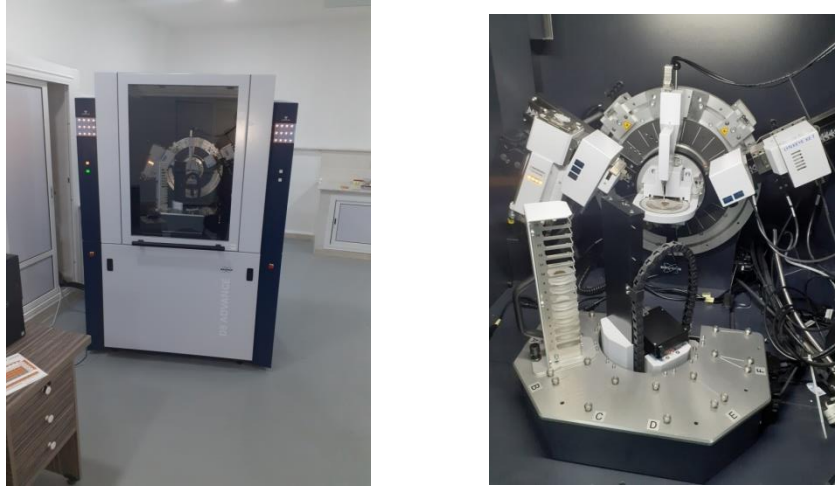


Figure.II.3. Type diffractometer BRUCKER-D8 ADVANCE

- **X-ray powder diffraction analysis**

X-ray powder diffraction is most widely used for the identification of unknown crystalline materials (e.g. minerals, inorganic compounds), is a non-destructive characterization method that identifies crystallized phases present in any material compared to a reference file, updated annually, which currently contains data on 69,500 compounds (Dossier J.C.P.D.S: Joint Committee for Powder Diffraction Standards. A careful analysis of diffractograms gives access to various properties of a crystallized material.

- **The position:** determining the positions of the lines allows the identification of the crystalline phase and the calculation of its lattice parameters.
- **The form:** the shape of the lines provides information about the size of the coherent diffraction domains and the rate of structural defects present in the sample.
- **Relative intensity:** the determination of the relative intensities of the lines allows conclusions to be drawn about the position of the various atoms in the crystal lattice.

- **Principle of obtaining spectra:**

The powder, which consists of an infinite number of (crystalline) grains, is bombarded with a monochromatic X-ray beam of known wavelength generated by a copper anticathode. The emitted radiation is defined by a system of sources (Sollers slits) and windows located before and after the sample. The latter is placed on a sample holder that rotates with a uniform motion around an axis (angular circle) lying in its plane, which increases the number of possible orientations of the lattice planes (hkl). Because the particles are randomly oriented,



there will always be a family of planes that lead to diffraction, for which the BRAGG relationship is verified. [31]

$$2d_{hkl}\sin\theta = n\lambda$$

$\lambda$  : Incident X-ray beam wavelength

$\theta$ : Diffraction angle

$d_{hkl}$ : Inter-reticular distance characterizing the family of planes detected by the indices h,k,l.

**n**: integer

### Differential thermal analysis (DTA) / Thermogravimetric analysis (TG)

Thermal analysis is the analysis of a change in a property of a sample that is related to an applied temperature change. The sample is normally in a solid state and the changes that occur upon heating include melting, phase transition, sublimation and decomposition.

- **Thermogravimetric analysis (TG)**

The analysis of the change in the mass of a sample on heating. TG measures mass changes in a material as a function of temperature (or time) under a controlled atmosphere. Its principal uses include measurement of a material's thermal stability and composition. TG is most useful for dehydration, decomposition, desorption, and oxidation processes.

- **Differential thermal analysis (DTA)**

The most widely used thermal method of analysis In DTA, the temperature of a sample is compared to that of an inert reference material during a programmed temperature change. The temperature must remain the same until a thermal event occurs, such as melting, decomposition or a change in crystal structure. When an endothermic event occurs within the sample, the sample temperature lags behind the reference temperature and a minimum is observed on the curve. On the other hand, when an exothermic event occurs, the temperature of the sample exceeds the reference temperature and a maximum is observed on the curve. The area under the endotherm or oxotherm is related to the enthalpy of the thermal event,  $\Delta H$ .

For many problems it is advantageous to use both DTA and TG since DTA events can be classified into those that do or do not involve a mass change. [34,35]

TG-DTA modes can be used to determine the following:

Melting points	Thermal and oxidative stability
Glass transition temperatures	Purity
Cristallinity	Transformation temperatures
Moisture/ volatile content	



Figure.II.4. Differential Thermal Analysis (DTA) / Thermogravimetric Analysis (TG) Device

- **Fourier transforms infrared spectroscopy analysis (FTIR):**

Infrared spectroscopy has been a workhorse for materials analysis in the laboratory for more than 70 years. An infrared spectrum represents a fingerprint of a sample with absorption peaks that correspond to the vibrational frequencies between the bonds of the atoms that make up the material. Because each different material is a unique combination of atoms, no two compounds produce exactly the same infrared spectrum. Therefore, infrared spectroscopy can lead to a positive identification (qualitative analysis) of each different type of material. The size of the peaks in the spectrum is also a direct indication of the amount of material present. With modern software algorithms, infrared is an excellent tool for quantitative analysis.

Fourier Transform Infrared (FTIR) spectrometry was developed to overcome the limitations encountered with dispersive instruments. The main difficulty was the slow scanning process. A method was needed to measure all infrared frequencies simultaneously rather than individually. A solution was devised using a very simple optical device called an interferometer. The interferometer produces a unique type of signal in which all infrared

frequencies are encoded. The signal can be measured very quickly, typically on the order of a second or so.

This reduces the amount of time per scan to a few seconds instead of several minutes. The interferometer is the result of the mutual "interference" of these two beams. The resulting signal is called an interferogram and has the unique property that each data point (a function of the position of the moving mirror) that makes up the signal contains information about each infrared frequency coming from the source.

Since the analyst needs a frequency spectrum (a plot of the intensity at each individual frequency) for identification, the measured interferogram signal cannot be directly interpreted. A means of "decoding" each frequency is needed. This can be accomplished through a well-known mathematical technique called the Fourier Transformation. This transformation is performed by the computer, which then presents the user with the desired spectral information for analysis. [36]

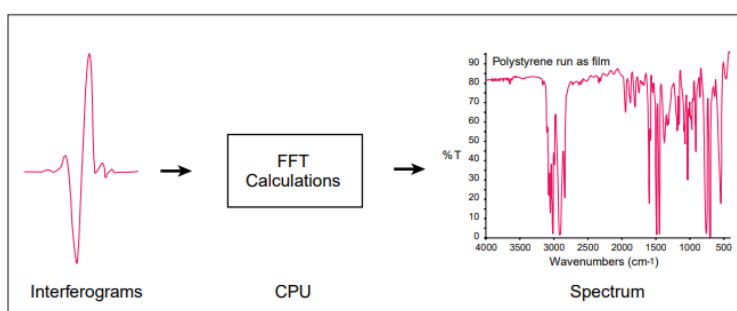


Figure.II.5. FT-IR workflow

- **Specific area measurement by the BET method**

The BET method was developed by Brunauer, Emmett and Teller in 1938 and allows specific surface areas to be measured by gas adsorption. It is based on the determination of gas quantity necessary to cover the external surface and internal pores of a solid by a complete monolayer. The method is applicable on powdered solid sample which particle diameter does not exceed 2 mm and which specific surface area is greater than  $0.2 \text{ m}^2 \cdot \text{g}^{-1}$ .

The sample is placed in an oven at  $105^\circ\text{C}$ , crushed and put into a glass sample holder. In order to empty the sample porosity of water and air that it may contain and enable fixation of  $\text{N}_2$  gas, the powdered sample is degassed at  $105^\circ\text{C}$  for 120 minutes and cooled in a bath of liquid nitrogen at a temperature of 77 K, to avoid gas condensation with increasing

temperature. Helium, a gas that will not fix on the sample surface, is injected into the sample holder to measure the volume which is not occupied by the sample. After helium evacuation, nitrogen is injected by successive steps, enabling the apparatus to measure the pressure in the sample holder. The partial pressure regularly measured and in the range of 0 to 0.995 enables to determine the quantity of adsorbed nitrogen. Results are processed using the equation of Brunauer, Emmett and Teller: [37]

$$\frac{\frac{P_S}{P_0}}{n_a(1-\frac{P_S}{P_0})} = \frac{1}{n_m C} + \left(\frac{C-1}{n_m C}\right) P_S/P_0 \quad *$$

$P_S[Pa]$ : is the pressure of adsorption gas in equilibrium with the adsorbate gas.

$P_0[Pa]$ : is the saturation vapour pressure of the adsorption gas.

$\frac{P_S}{P_0}$ : is the relative pressure of the adsorption gas.

$n_a[\text{mol}\cdot\text{g}^{-1}]$ : is the specific adsorbed gas quantity.

$n_m [\text{mol}\cdot\text{g}^{-1}]$ : is the molecular coverage capacity, quantity of adsorbed gas necessary to cover a unit surface with a complete monolayer.

$C$ : is the BET constant.

$\frac{\frac{P_S}{P_0}}{n_a(1-\frac{P_S}{P_0})}$ : is represented in function of relative pressure  $P_S/P_0$ . When  $P_S/P_0$  is in the range 0.05 to 0.35.

Equation \* is a linear function  $y= ax+b$  with slope  $a = \left(\frac{C-1}{n_m C}\right)$

and y-intercept  $b = \frac{1}{n_m C}$

BET constant writes:  $C = \frac{a}{b} + 1$

And the monolayer volume is given by:

$$V_m = \frac{1}{a + b}$$

The corresponding specific surface area is deduced with the following relation

$$A_s = \frac{V_M}{V_m V_{sample}} S_{Adsorbate} N_A$$

Where  $A_s$  [ $\text{m}^2 \cdot \text{g}^{-1}$ ]: is the specific surface area of the solid.

$V_M$  [ $\text{cm}^3$ ]: is the volume of the adsorbed gas monolayer.

$S_{Adsorbate}$  [ $\text{m}^2$ ]: is the area of the efficient section per adsorbate molecule.

$V_M$  [ $22414 \text{ cm}^3 \cdot \text{mol}^{-1}$  at  $P = 1 \text{ atm}$  and  $T = 25^\circ\text{C}$ ]: is the volume of a molecular gram.

$M_{sample}$  [ $\text{g}$ ]: the mass of the sample after degassing.

$N_A$  [ $6.022 \cdot 10^{23} \text{ atomes} \cdot \text{mol}^{-1}$ ]: is the Avogadro constant.



Figure.II.6.BET Surface Area Analyzer quantachrome

### III. Applying Machine learning and Data mining methods in perovskite materials design and discovery

In this section we will talk about the successful application of ML and DM techniques in properties prediction and stability assessment of perovskite material.

#### III.1.The workflow of Machine Learning and Data Mining

The most common application of ML is to construct a statistical model used for data analysis and prediction. The main purpose of ML aims at evaluating or predicting the objects after training the model with historical data and specific conditions [38]



Figure.III.1. the general workflow of ML in perovskite materials

- The general form and layout of the workflow may change depending on the Data and its features.

##### III.1.1.Data preparation:

The dataset used for ML sometimes contains dependent and independent variables related to the materials. independent variables, also called features ,descriptors or Attributes, refer to the representative info involving the structure and characteristics of materials, together with the chemical composition, atomic or molecular parameters, structural parameters, additionally because the technological conditions for synthesis process. The dependent variables refer to the target property of the materials affected by the independent variables, also known as the target variables. [39,40]

There are few important notes to take while preparing the data:

- ✓ The quantity and quality of data are key factors within the discovery of materials.
- ✓ A general rule of thumb is that a reasonable ml model needs the quantity of data over thrice of descriptors at least.

- ✓ The quality of the data depends on the spatial coverage of the target properties and the uncertainties associated with the data.
- ✓ Data with a standard distribution is best for ML, insufficient data of specific target or poor coverage of specific properties might not form an appropriate data distribution for ML.
- ✓ Data uncertainty, such as experimental error or calculation error, might have an effect on the quality of data. The roughness of the modeling data directly determines the consequences and results of the created model.
- ✓ The prediction error of the model is higher than the error of the training data.
  - To reduce the roughness of the data there are several steps to do we included some of them:
    - ✓ The deletion of missing values
    - ✓ The completion of experimental conditions
    - ✓ Data normalization and scaling
    - ✓ Data standardization and (can improve model accuracy and convergence speed)
      - We can retrieve and collect the data from known and valid sources like available and authorize databases, research papers
- ✓ The use of autonomous workflows to generate data in a convenient and fast way but risk the quality of data obtained cause it will be inferior to the data obtained from databases
- ✓ The dataset could also be generated through lab-scale calculations performed by many data mining for materials packages like Materials Studio (MS), Vienna Ab initio Simulation Package (VASP)...etc.
  - ML models created by the calculated data have comparatively sensible evaluation metrics. However, the calculations of complicated materials might take up too many calculation resources and take an extended time.
  - The results of many ml algorithms will vary with whether or not any standardization or scaled. It's value noting that each feature variables and target variables can be normalized or scale.
  - Data need to be reformatted into a single tabular form, imputed missing values, eliminated erroneous or incomparable data points, and normalized and rescaled the data. [29]

Table.III.1. Publicly accessible databases of various materials

Database	Brief description	URL
Materials Project (MP)	Calculation data of properties of known and hypothetical materials	<a href="https://materialsproject.org">https://materialsproject.org</a>
The Inorganic Crystal Structure Database (ICSD)	Experimental characterization data of inorganic crystal structure	<a href="https://icsd.fiz-karlsruhe.de/index.xhtml">https://icsd.fiz-karlsruhe.de/index.xhtml</a>
Cambridge Structural Database (CSD)	The structure database of small molecules and metal-organic molecular crystals based on X-ray and neutron diffraction experiments collected by the Cambridge Crystallographic Data Centre	<a href="https://www.ccdc.cam.ac.uk/">https://www.ccdc.cam.ac.uk/</a>
Aflow-Automatic-FLOW for Materials Discovery (AFLOW)	A data repository of structure and property of inorganic materials from high-throughput ab initio calculations	<a href="http://www.aflowlib.org">http://www.aflowlib.org</a>
Crystallography Open Database (COD)	Structures data of organic, inorganic, and metal-organic compounds and minerals	<a href="http://cod.ensicaen.fr">http://cod.ensicaen.fr</a>
Open Quantum Materials Database (OQMD)	Theoretical simulation calculation data of mostly hypothetical materials	<a href="http://www.oqmd.org/">http://www.oqmd.org/</a>
Materials Platform for Data Science (MPDS)	Peer-reviewed crystal structure, phase diagram, or physical property	<a href="https://mpds.io/#modal/menu">https://mpds.io/#modal/menu</a>
Springer Materials	The world's largest material data resource, a unique, high-quality numerical database	<a href="https://materials.springer.com">https://materials.springer.com</a>
Materials Cloud	Structural calculation data of candidate two-dimensional materials	<a href="https://www.materialscloud.org/discover/2dstructures/dashboard/ptable">https://www.materialscloud.org/discover/2dstructures/dashboard/ptable</a>
Materiae	Topological material database	<a href="http://materiae.iphy.ac.cn/">http://materiae.iphy.ac.cn/</a>



### III.1.2.Feature generation and Feature selection

We will need a set of candidate features or attributes to form a valid benchmark data set to train and test the model. These features are usually derived from known properties of the constituent elements such as atomic radius and electronegativity. The quantity of features should be less than that of dataset samples for effectively training and avoiding overfitting.

The properties of each material depend on a specific set of features; therefore feature selection should reduce the dimension of input space as much as possible without losing important information. In particular, redundant and high self-correlation features should be removed to guarantee the efficiency and accuracy of models. Known feature selection methods are:

Exhaustive search: Exhaustive search can get the best subset, but it is only suitable for small datasets because it consumes a lot of computer time.

Heuristic search: It is suitable for medium-sized data sets.

Non-deterministic search: like search with the genetic algorithm and others. It can be used for large datasets.

Reasonable material features should meet the following three conditions:

- Perfect representation of material properties.
- Sensitive to target properties.
- Easy to discover and obtain. [29]

### III.1.3. Model selection

ML algorithms could be briefly divided into two categories: supervised learning and unsupervised learning. Supervised learning is the process of using a set of samples with known labels to adjust the parameters of the models and achieve the required performance, which is further divided into regression and classification. If the target property is a continuous value, the process is called regression. If the target is a discrete value, the process of searching the prediction function is called classification (Figure III.2)

Generally, the best model is obtained by comparing multiple algorithms. The criteria of algorithm selection are mainly based on the results of cross validation and independent test. The commonly used evaluation metrics include mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), determination coefficient ( $R^2$ ), and correlation coefficient (R) for regression.

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}$$

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y - \hat{y})^2}{N}}$$

$$R = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

There other evaluation metrics such as confusion matrix, precision, recall, receiver operating characteristic curve (ROC), and area under ROC curve (AUC) for classification. [29]

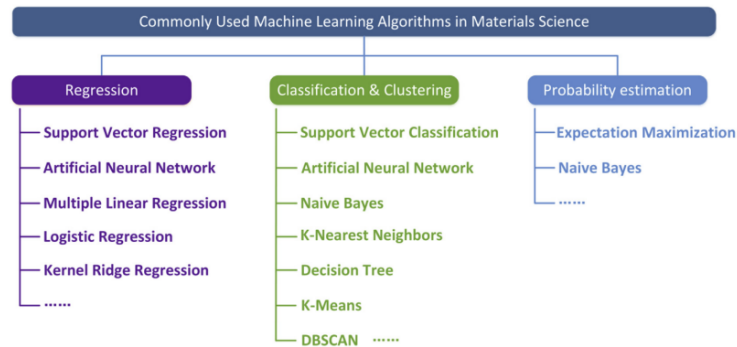


Figure.III.2. Commonly used machine learning algorithms in materials science

Explaining some of the ML algorithms:

**Support vector machine:** Support vector machine (SVM) includes support vector classification (SVC) and support vector regression (SVR). The main idea of SVC is to establish an optimal decision hyperplane to maximize the distance between the two kinds of samples closest to the plane on both sides of the plane. The basic idea of SVR is to map the data  $X$  into a higher-dimensional feature space  $F$  via a nonlinear mapping  $\Phi$  and then to do linear regression in this space. SVM provides good generalization ability for classification and regression tasks.

**Artificial Neural Networks (ANN):** falls in the category of regression and classification tasks. A neural network is composed of a large number of connected nodes (neurons). Samples are classified or regressed according to different connection modes and connection signals (weights) between nodes.

**Multiple Linear Regressions:** Solve the regression problem when the relationship between multiple independent variables and one dependent variable is linear. [29]

### III.1.4. Model evaluation

After model selection, it is usually necessary to tune the internal hyper-parameters of the model algorithm to for balance over-fitting and under-fitting, in other word optimize the selected model.

Even well-trained ML models can contain errors due to noise in the training data, measurement limitations, computational uncertainties, or simply outliers or missing data. Poor model performance usually indicates high bias or high variance. [29,41]

➤ **Terminology:**

**Model optimization:** When the learner (or set of learners) has been chosen and predictions are being made, the newly constructed model should be evaluated to permit for optimization and supreme choice of the most effective model. 3 principal sources of error arise and must be taken into account: model bias, model variance, and irreducible errors.

$$\text{Total Error} = \text{Bias} + \text{Variance} + \text{Irreducible Errors}$$

**Bias:** It is the error of wrong assumptions in the algorithm and can cause the model to lack underlying relationships.

**Variance:** is sensitivity to small fluctuations in the training set

**High bias:** or under-fitting It occurs when the model is not flexible enough to adequately describe the relationship between the expected inputs and predicted outputs, or when the data is not detailed enough to allow appropriate rules to be discovered.

**High variance:** or over-fitting occurs when a model becomes too complex; typically this occurs when the number of parameters is increased. The diagnostic test for over-fitting is that a model's accuracy in representing training data continues to improve while its performance in estimating test data declines. [41]

➤ **model evaluation methods:**

There are three commonly used model evaluation methods: independent test, cross validation, and bootstrapping.

**Independent test method:** The model's generalization error can be assessed by testing, but the goal of the model is to predict unknown samples. Therefore, a set of tests is required to test generalizability. The error obtained with the test set can be taken as an approximation of the generalization error. The smallest error from the independent test generally indicates the strongest generalization of the available model. It's worth noting that the independent test set and the training set must be mutually exclusive.

**Cross validation (CV) or k-fold cross validation method:** this method is used to evaluate the reliability of the ML models. The input data is divided into k mutually exclusive subsets of the similar size, each subset is generated by 'stratified samples'. The union of k-1 subsets is used as the training set with the remaining one used as the testing set. After k times of training

and testing, all test results are averaged to represent the final ML performance. The stability and fidelity of the evaluation results of the CV method depend to a large extent on the value of  $k$ .

$K$  is a specified number; the commonly used values are 5, 10, and 20. When  $k$  is equal to the sample number of input data, this method is called leave-one-out cross validation (LOOCV).

LOOCV is not affected by random sample partitioning and the results are often considered to be more accurate.

**Bootstrap method:** represented by giving Given a dataset  $D$  containing  $m$  samples, then it will randomly copy a sample from  $D$  to the dataset  $D'$  at a time until  $D'$  contains  $m$  samples. Some data may be sampled repeatedly while some data may never be sampled.

$D$  is designated as the training set and  $D'$  is used as the testing set.

The number of training samples obtained by the bootstrapping method is equal to the original dataset.

The bootstrapping method is effective under the condition of a small dataset. [29]

### III.1.5. Model application

The purpose of ML is to generalize the hidden patterns between descriptors and material properties of existing data samples. The properties could be accurately predicted with the built model. Therefore, the developed ML model can be applied to high-throughput screening. First, many virtual patterns could be designed and then the properties could be predicted with the ML model. Finally, the materials with the desired properties would be selected from the hypothetical samples for the experiments. [29]

As an optional step we can develop the online prediction model for sharing. The network model enables more users to predict target properties this make model applications easier and faster. For example Furmanchuk and al. [42] developed an online application to predict the Seebeck coefficient of crystalline materials.

## IV. Application of machine learning and data mining in perovskite materials

The maximum thrilling success of materials studies is to discover a few new compounds with distinctive shape and high-quality properties new perovskite type materials with exciting properties this is achieved by using the support vector machine algorithm.

- Predicting the thermodynamic stability of around 230,000 possible  $ABX_3$  compounds and screen out the stable candidates using ERT algorithm with the prediction accuracy was set and evaluated with MAE scores 121 meV/atom. [43]
- 40 potential  $ABX_3$  perovskite halides with high perovskite crystal structure formation probability were determined using SVM algorithm with the training set scores 93.8% and the testing set 92.1%. [44]
- Taking an important step towards a basic understanding of the interfacial properties of perovskite, facilitating further breakthroughs in photovoltaic technology. Proposed two promising stable candidate materials,  $RbSnCl_3$  and  $RbSnBr_3$ , for future photovoltaic and related applications using also SVM with the training set scores 94% and the testing set 96% [45]
- Propose 11 undiscovered Li (Na) based perovskite materials with ideal bandgap and formation energy ranges for solar cell applications using RF algorithm mean score is 0.98 standard deviation is 0.002. [46]

## Bibliography

### References:

- 1 Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715), 547-555.
- 2 Haghightlari, M., Vishwakarma, G., Altarawy, D., Subramanian, R., Kota, B. U., Sonpal, A., Setlur, S., Hachmann, J. (2020). ChemML: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10(4), e1458.
- 3 Hautier, G. (2016, August). Prediction of new battery materials based on ab initio computations. In *AIP Conference Proceedings* (Vol. 1765, No. 1, p. 020009). AIP Publishing LLC.
- 4 Aykol, M., Kim, S., Hegde, V. I., Snyder, D., Lu, Z., Hao, S., Kirklin, S., Morgan, D., Wolverton, C. (2016). High-throughput computational design of cathode coatings for Li-ion batteries. *Nature communications*, 7(1), 1-12.
- 5 Nyshadham, C., Oses, C., Hansen, J. E., Takeuchi, I., Curtarolo, S., & Hart, G. L. (2017). A computational high-throughput search for new ternary superalloys. *Acta Materialia*, 122, 438-447.
- 6 Kirklin, S., Saal, J. E., Hegde, V. I., & Wolverton, C. (2016). High-throughput computational search for strengthening precipitates in alloys. *Acta Materialia*, 102, 125-135.
- 7 Ward, L., Liu, R., Krishna, A., Hegde, V. I., Agrawal, A., Choudhary, A., & Wolverton, C. (2017). Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B*, 96(2), 024104.

- 8 Rupp, M., Tkatchenko, A., Müller, K. R., & Von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5), 058301.
- 9 Carrete, J., Li, W., Mingo, N., Wang, S., & Curtarolo, S. (2014). Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Physical Review X*, 4(1), 011019.
- 10 Ward, L., & Wolverton, C. (2017). Atomistic calculations and materials informatics: A review. *Current Opinion in Solid State and Materials Science*, 21(3), 167-176.
- 11 Luo, J., Lezzi, P. J., Vargheese, K. D., Tandia, A., Harris, J. T., Gross, T. M., & Mauro, J. C. (2016). Competing indentation deformation mechanisms in glass using different strengthening methods. *Frontiers in Materials*, 3, 52.
- 12 Bucholz, E. W., Kong, C. S., Marchman, K. R., Sawyer, W. G., Phillpot, S. R., Sinnott, S. B., & Rajan, K. (2012). Data-driven model for estimation of friction coefficient via informatics methods. *Tribology Letters*, 47(2), 211-221.
- 13 Wang, A. Y. T., Murdock, R. J., Kauwe, S. K., Oliynyk, A. O., Gurlo, A., Brgoch, J., Persson, K.A., & Sparks, T. D. (2020). Machine learning for materials scientists: an introductory guide toward best practices. *Chemistry of Materials*, 32(12), 4954-4965.
- 14 PV, Y. R. L. Z. B. (2018). Xue D. Zhou Y. Ding X. Sun J. Xue D. Lookman T. *Adv. Mater*, 30, 1702884.
- 15 Mannodi-Kanakkithodi, A., Chandrasekaran, A., Kim, C., Huan, T. D., Pilia, G., Botu, V., & Ramprasad, R. (2018). Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Materials Today*, 21(7), 785-796.
- 16 Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A., & Armiento, R. (2016). Machine learning energies of 2 million elpasolite (A B C 2 D 6) crystals. *Physical review letters*, 117(13), 135502.



- 17 Ren, F., Ward, L., Williams, T., Laws, K. J., Wolverton, C., Hattrick-Simpers, J., & Mehta, A. (2018). Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science advances*, 4(4), eaaq1566.
- 18 Seko, A., Hayashi, H., Nakayama, K., Takahashi, A., & Tanaka, I. (2017). Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14), 144110.
- 19 Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(1), 1-13.
- 20 Al-Harbi, H. F., Landi, G., & Kalidindi, S. R. (2012). Multi-scale modeling of the elastic response of a structural component made from a composite material using the materials knowledge system. *Modelling and Simulation in Materials Science and Engineering*, 20(5), 055001.
- 21 Hill, J., Mulholland, G., Persson, K., Seshadri, R., Wolverton, C., & Meredig, B. (2016). Materials science with large-scale data and informatics: Unlocking new opportunities. *Mrs Bulletin*, 41(5), 399-409.
- 22 Logan Warda, Alexander Dunnc, Alireza Faghaniniac, Nils E.R. Zimmermann , Saurabh Bajaj,Qi Wang , Joseph Montoya , Jiming Chen, Kyle Bystrom, Maxwell Dylla , Kyle Charda ,Mark Asta, Kristin A. Persson, G. Jeffrey Snyder, Ian Foster, Anubhav Jain.152, (2018) 60-69.
- 23 Oró-Solé, J., Clark, L., Kumar, N., Bonin, W., Sundaresan, A., Attfield, J. P., Rao, C.N.R., & Fuytes, A. (2014). Synthesis, anion order and magnetic properties of RVO  $3-x$  N x perovskites (R= La, Pr, Nd;  $0 \leq x \leq 1$ ). *Journal of Materials Chemistry C*, 2(12), 2212-2220.

- 24 Shiogai, J., Chida, T., Hashimoto, K., Fujiwara, K., Sasaki, T., & Tsukazaki, A. (2020). Signature of band inversion in the perovskite thin-film alloys  $\text{BaSn}_{1-x}\text{Pb}_x\text{O}_3$ . *Physical Review B*, 101(12), 125125.
- 25 Ekström, E., Le Febvrier, A., Bourgeois, F., Lundqvist, B., Palisaitis, J., Persson, P. Å., ... & Eklund, P. (2020). The effects of microstructure, Nb content and secondary Ruddlesden–Popper phase on thermoelectric properties in perovskite  $\text{CaMn}_{1-x}\text{Nb}_x\text{O}_3$  ( $x=0-0.10$ ) thin films. *RSC advances*, 10(13), 7918-7926.
- 26 Sydorchuk, V., Lutsyuk, I., Shved, V., Hreb, V., Kondyr, A., Zakutevskyy, O., & Vasylechko, L. (2020).  $\text{PrCo}_{1-x}\text{Fe}_x\text{O}_3$  perovskite powders for possible photocatalytic applications. *Research on Chemical Intermediates*, 46(3), 1909-1930.
- 27 Li, L., Tian, G., Chang, W., Yan, Y., Ling, F., Jiang, S., Xiang, G., & Zhou, X. (2020). A novel double-perovskite  $\text{LiLaMgTeO}_6$ :  $\text{Mn}^{4+}$  far-red phosphor for indoor plant cultivation white LEDs: Crystal and electronic structure, and photoluminescence properties. *Journal of Alloys and Compounds*, 832, 154905.
- 28 Zhao, D., Wang, B., Liang, C., Liu, T., Wei, Q., Wang, S., Wang, K., Zhang, Z., Li, X., Peng, S., & Xing, G. (2020). Facile deposition of high-quality  $\text{Cs}_2\text{AgBiBr}_6$  films for efficient double perovskite solar cells. *Science China Materials*, 63(8), 1518-1525.
- 29 Tao, Q., Xu, P., Li, M., & Lu, W. (2021). Machine learning for perovskite materials design and discovery. *npj Computational Materials*, 7(1), 1-18.
- 30 Wang, X. (2020). Preparation, synthesis and application of sol-gel method. Vidyasirimedhi Institute of Science and Technology.
- 31 M. Lebid. (2016). Study of the physic o-chemical properties of oxides based on lanthanum, iron and magnesium. (Doctoral Thesis Univ. Biskra). Retrieved from [http://thesis.univ-biskra.dz/2350/1/Chimi\\_d1\\_2016.pdf](http://thesis.univ-biskra.dz/2350/1/Chimi_d1_2016.pdf)

- 32 Cullity, B. D. (1978). Elements of X-ray diffraction, Addison. Wesley Mass, 127-31.
- 33 Dr. John E. Warren & Mr. Gary Harrison(n.d.)  
(<https://xray.materials.manchester.ac.uk/equipment/BrukerD8Advance.html>)
- 34 Atkins, P. W. (2006). Shriver & Atkins Inorganic Chemistry: Solutions manual (Vol. 2). WH Freeman and Company.
- 35 West, A. R. (1999). Basic solid state chemistry. John Wiley & Sons Incorporated.
- 36 Introduction to Fourier Transform Infrared Spectrometry. (2001) Thermo Nicolet Corporation.
- 37 Yu et al. (2017). BET Method. hal.archives-ouvertes.
- 38 Goldsmith, B. R., Esterhuizen, J., Liu, J. X., Bartel, C. J., & Sutton, C. (2018). Machine learning for heterogeneous catalyst design and discovery.
- 39 Lu, W., Xiao, R., Yang, J., Li, H., & Zhang, W. (2017). Data mining-aided materials discovery and optimization. Journal of materiomics, 3(3), 191-201.
- 40 Wan, X., Feng, W., Wang, Y., Wang, H., Zhang, X., Deng, C., & Yang, N. (2019). Materials discovery and properties prediction in thermal transport via materials informatics: a mini review. Nano letters, 19(6), 3387-3395.
- 41 Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. Nature, 559(7715), 547-555.
- 42 Furmanchuk, A. O., Saal, J. E., Doak, J. W., Olson, G. B., Choudhary, A., & Agrawal, A. (2018). Prediction of seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach. Journal of computational chemistry, 39(4), 191-202.

- 43 Schmidt, J., Shi, J., Borlido, P., Chen, L., Botti, S., & Marques, M. A. (2017). Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chemistry of Materials*, 29(12), 5090-5103.
- 44 Pilania, G., Balachandran, P. V., Kim, C., & Lookman, T. (2016). Finding new perovskite halides via machine learning. *Frontiers in Materials*, 3, 19.
- 45 Jain, D., Chaube, S., Khullar, P., Srinivasan, S. G., & Rai, B. (2019). Bulk and surface DFT investigations of inorganic halide perovskites screened using machine learning and materials property databases. *Physical Chemistry Chemical Physics*, 21(35), 19423-19436.
- 46 Takahashi, K., Takahashi, L., Miyazato, I., & Tanaka, Y. (2018). Searching for hidden perovskite materials for photovoltaic systems by combining data science and first principle calculations. *ACS Photonics*, 5(3), 771-775.

# Chapter 03

Using Data Mining to  
Search for Perovskite  
Materials with Higher  
Specific Surface Area



## I. Introduction

The  $ABO_3$  perovskite have an interesting property known as the specific surface area (SSA) that is one of the vital and necessary properties related to photo catalytic ability [1]. In this work the main objective is to try to study the link between the SSA (in the vary of  $1-60 \text{ m}^2 \text{ g}^{-1}$ ) of perovskite and its features, together with technical parameters and chemical compositions with a help of data mining tasks and machine learning algorithms, basically will preparing a workflow consist of data mining crucial tasks (data preparation and pre-processing, generating and selecting features) and machine learning algorithms (generally classification and regression) in the goal of creating a predictive model capable of predicting and finding perovskite-type materials with higher specific surface area.

Once understanding each step of the workflow, a question will pop up where and how to execute these important steps, which platforms and tools that help you process your data make predictions and create your model. There are a lot of web applications and software that can help you execute ML and DM tasks like jupyter notebook powered by python a programming language which is the most effective in ML and DM projects, this language houses so many libraries such us NumPy, Pandas, and Scikit-learn which are a collection of pre-compiled programming routines frequently used in machine learning. Other software like Monkey Learn, Rapid Miner, Weka... etc. As a beginner using the data mining software Weka to perform the many DM and ML tasks, is an efficient and easy way.

This work is inspired by a group of researchers from the university of Materials Genome Institute, Shanghai University China who were able to create a data mining model capable of predicting the specific surface area ( $SSA \text{ m}^2.\text{g}^{-1}$ ) of perovskite-type materials ( $ABO_3$ ) using Genetic-support vector regression algorithm (Ga-SVR) implemented by a computational software called ExpMiner (a data mining software package) developed in their laboratory and then they screen out 5 perovskite-type materials using the Online Computation Platform for Materials Data Mining (OCPMDM) developed in their lab too which been found very hard to use because of it being in beta state and not that completely functional. They used specifically SVR algorithm (a supervised machine learning algorithm) By comparing it to two different machine learning algorithms, namely, partial least-squares (PLS)[2], and artificial neural network (ANN)[3] and they find that the optimal model is the SVR model[4].( as said in chapter 2 SVR algorithm been used a lot in materials chemistry).

By constructing a SVR model they found the correlation coefficient (R) between the predicted and experiment SSA as high as 0.986 for the training dataset and 0.935 for the leave one out cross validation (LOOCV). From they were able to predict and screen out 5 perovskite type materials with a higher specific surface area utilizing this model through the OCPMDM.

Table.I.1.The four perovskite-type materials screened out using Li Shi, Dongping Chang, Xiaobo Ji, and Wencong Lu.Journal model

Molecular formula	SSA (m <sup>2</sup> g <sup>-1</sup> )
LaFe <sub>0.8</sub> Mg <sub>0.2</sub> O <sub>3</sub>	57.70
LaFe <sub>0.7</sub> Mg <sub>0.3</sub> O <sub>3</sub>	58.09
LaFe <sub>0.9</sub> Co <sub>0.1</sub> O <sub>3</sub>	54.81
LaFe <sub>0.8</sub> Co <sub>0.2</sub> O <sub>3</sub>	54.82
LaFe <sub>0.7</sub> Co <sub>0.3</sub> O <sub>3</sub>	52.03

So returning to this work where an SVR model will be created but with a twist by using a little different set of features from the same dataset for each model that means through feature selection then picking the best model that gives the best results .Note that the researchers also worked with different set of features which are B-aff (electron affinity of the B position),B-Tm(the melting point of the B position),A-Tb(normal boiling point of the A position),CT(Calcinations temperature),and AH(calcinations time).

All the work will be done using Weka as said before. Further explanation here below:

### I.1.What is Weka:

Developed at the University of Waikato in New Zealand and named after a flightless bird found only on the islands of New Zealand, Weka is an open-source software for data mining it contains tools for data preparation, regression classification, clustering, association rules mining, feature selection, and visualization to help in various DM and ML tasks all this without writing any program code at all. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems.



Figure.I.1. Weka graphic interface

## I.2.How to use Weka

A simple way to use Weka is to go and upload your dataset into the software apply a learning method and analyze its output to learn more about the data. Also you can use learned models to generate predictions for new samples (instances). Another useful way is applying different learners and compares their performance to select one to predict. In the interactive WEKA interface, we can select the desired learning method from an easy access menu. Plus we can tune and modify a number of parameters of these methods through an object editor.

**The explorer option** is where all the work will be done from all the data mining tasks and machine learning techniques. Through it we can quickly read in a dataset from a file with different types ( csv, data, Libsvm...etc) with Weka standard file type comes with (.arff) extension. It guides very well by giving access to all of its facilities using menu selection and form filling.

**The experiment option** helps us answer which methods and parameter values work best for the given problem when applying classification and regression techniques.

**The Knowledge Flow** interface allows us to design configurations for streamed data processing.

The last one **the workbench** which is the most configurable s a unified graphical user interfaces that combines the other three. [5]

Much of the work will circle around the explorer section so an explanation is below with further details.



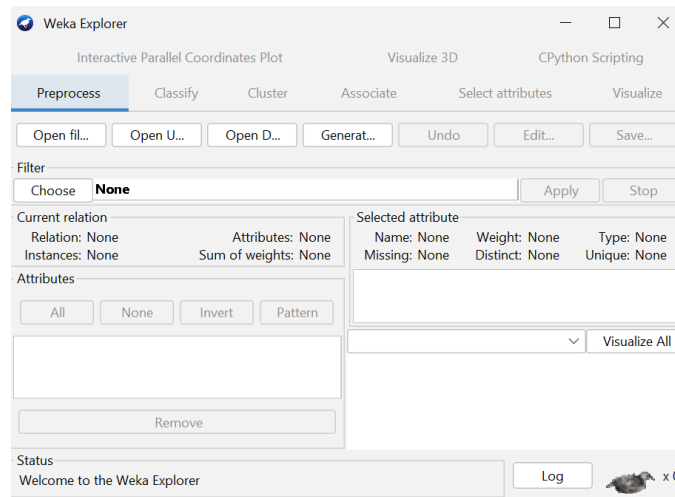


Figure.I.2. the Explorer interface.

**Pre-process section:** from this option you will be able to import the dataset and modify it in various ways. Only files whose names end in .arff appear in the file browser you can change the Format item in the file selection box. Weka support many data files such as

- Spreadsheet files with extension .csv
- Serialized instances with extension .bsi
- SVM-Light format files with extension .dat
- ASCII Matlab files with extension .m
- C4.5's native file format with extensions .names and .data
- LIBSVM format files with extension .libsvm
- XML-based ARFF format files with extension .xrff
- Excel extension files .xls and .xlsx

Also in this section you can import databases from the DB button generate and import data from a URL...etc and so much other useful tools.

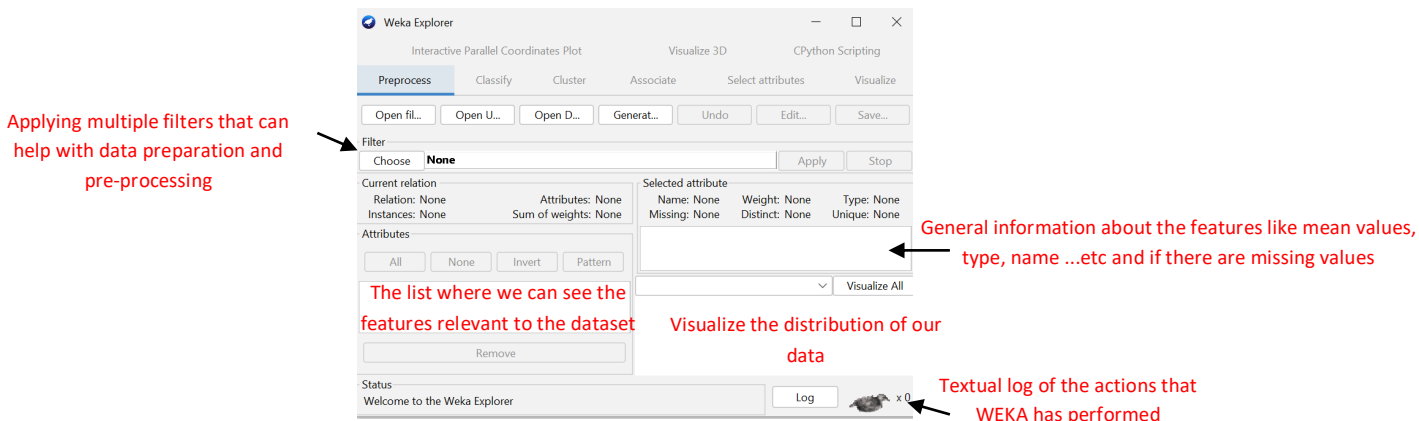


Figure.I.3. the Explorer interface pre-process section.

**Classify section:** the option lets you apply Variety of ML algorithms in order to perform classification or regression and evaluate them. Plus it let us train and test our data perform a lot of evaluation methods and more option to create an accurate models.

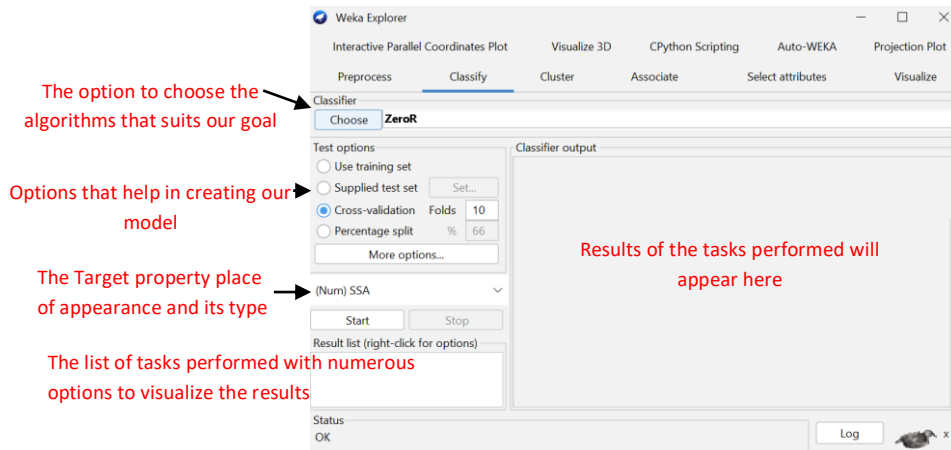


Figure.I.4. the Explorer interface classify section.

**Select attributes section:** where you can select key features which are the most relevant aspects of the dataset, it gives different methods for feature selection which are divided into a search method and attribute evaluators both are configurable.

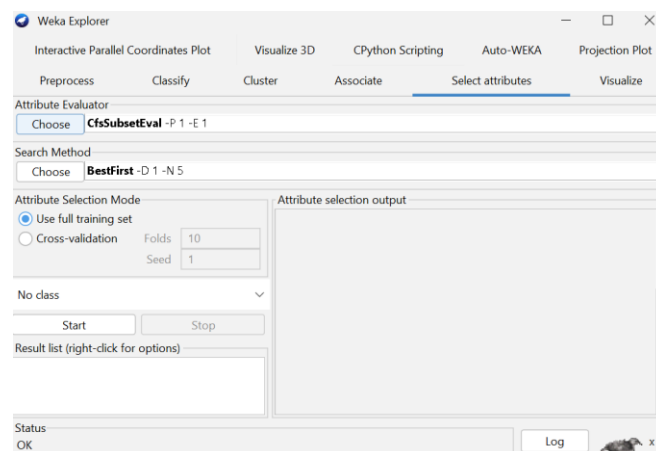


Figure.I.5. the Explorer interface Select attributes section.

**Visualize section:** View different two-dimensional plots of the data and interact with them.

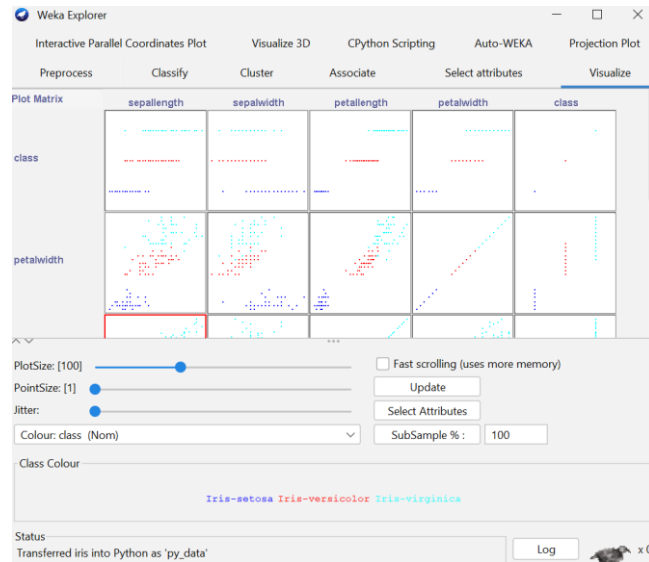


Figure.I.6.the Explorer interface Visualize section.

Other sections are: Cluster helps you Learn clusters from the dataset; Associate from you can Learn association rules for the data and evaluate them. [5]

## II. Executing My Data mining workflow:

Any Data mining and machine learning project need a workflow to better realize and achieve the goal of the task at hand so as mentioned in chapter two the workflow or the flowchart of data mining and machine learning that will be followed is:

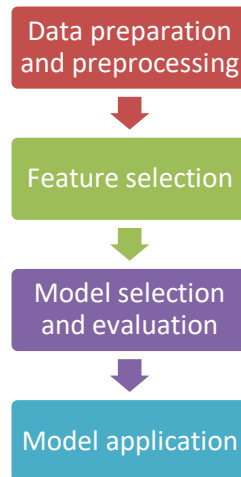


Figure.II.1.My DM and ML workflow

Beginning with creating the models following different steps in the first half of the workflow (data preparation and feature selection).

### II.1.Data preparation

#### Data n°01:

The data which been used consist of a total of 50 samples of Perovskite type-materials prepared via sol-gel method with their SSAs ranging from 1 to 60 m<sup>2</sup> g<sup>-1</sup> (Table.II.1) as support information collected by the same group of researchers from the literature [4], besides a list of a candidate features consisted of technical parameters and chemical compositions (Table.II.2). Deciding to work with this ready dataset in order to better understand how data mining and machine learning methods works in materials design and discovery and also compare the models results to theirs.

Table.II.1.The dataset complete with 50 samples (instances) and 24 candidate features

NO.	Molecular	SSA	Ra	Rb	Ea
1	ZnTiO <sub>3</sub>	1.05	74	67	1.65
2	LaFeO <sub>3</sub>	1.08	103.2	55	1.1
3	BiFeO <sub>3</sub>	0.7514	103	55	2.02
4	BiTi <sub>0.15</sub> Fe <sub>0.85</sub> O <sub>3</sub>	0.9507	103	56.8	2.02
5	LaCoO <sub>3</sub>	17	103.2	54.5	1.1
6	LaCo <sub>0.94</sub> Mg <sub>0.06</sub> O <sub>3</sub>	19	103.2	55.55	1.1
7	LaCo <sub>0.90</sub> Mg <sub>0.10</sub> O <sub>3</sub>	21	103.2	56.25	1.1
8	LaCo <sub>0.80</sub> Mg <sub>0.20</sub> O <sub>3</sub>	22	103.2	58	1.1
9	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	27.75	105	55	1.287
10	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	20.63	105	55	1.287
11	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	12.46	105	55	1.287
12	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	5.91	105	55	1.287
13	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	4.19	105	55	1.287
14	LaFeO <sub>3</sub>	11.39	103.2	55	1.1
15	LaMg <sub>0.2</sub> Fe <sub>0.8</sub> O <sub>3</sub>	15.07	103.2	58.4	1.1
16	LaMg <sub>0.4</sub> Fe <sub>0.6</sub> O <sub>3</sub>	17.63	103.2	61.8	1.1
17	LaMg <sub>0.6</sub> Fe <sub>0.4</sub> O <sub>3</sub>	24.41	103.2	65.2	1.1
18	LaMg <sub>0.8</sub> Fe <sub>0.2</sub> O <sub>3</sub>	13.32	103.2	68.6	1.1
19	LaMgO <sub>3</sub>	8.65	103.2	72	1.1
20	LaCrO <sub>3</sub>	3.95	103.2	61.5	1.1
21	LaMg <sub>0.2</sub> Cr <sub>0.8</sub> O <sub>3</sub>	8.42	103.2	63.6	1.1
22	LaMg <sub>0.4</sub> Cr <sub>0.6</sub> O <sub>3</sub>	29.71	103.2	65.7	1.1
23	LaMg <sub>0.6</sub> Cr <sub>0.4</sub> O <sub>3</sub>	18.41	103.2	67.8	1.1
24	LaMg <sub>0.8</sub> Cr <sub>0.2</sub> O <sub>3</sub>	14.46	103.2	69.9	1.1
25	PrFeO <sub>3</sub>	10.88	99	55	1.13
26	LaFe <sub>0.9</sub> Co <sub>0.1</sub> O <sub>3</sub>	51.2	103.2	54.95	1.1
27	LaFe <sub>0.1</sub> Co <sub>0.9</sub> O <sub>3</sub>	42.8	103.2	54.55	1.1
28	LaFeO <sub>3</sub>	8.5	103.2	55	1.1
29	SrTiO <sub>3</sub>	16.4	118	67	0.95
30	La <sub>0.002</sub> Sr <sub>0.998</sub> TiO <sub>3</sub>	19.7	117.9704	67	0.9503
31	La <sub>0.005</sub> Sr <sub>0.995</sub> TiO <sub>3</sub>	22.3	117.926	67	0.95075
32	La <sub>0.01</sub> Sr <sub>0.99</sub> TiO <sub>3</sub>	24.1	117.852	67	0.9515
33	La <sub>0.02</sub> Sr <sub>0.98</sub> TiO <sub>3</sub>	23.2	117.704	67	0.953
34	LaFeO <sub>3</sub>	9.5	103.2	55	1.1
35	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	25.8	103.2	55	1.1
36	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	22.55	103.2	55	1.1
37	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	20.04	103.2	55	1.1
38	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	8.5	103.2	55	1.1
39	La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	5.8	103.2	55	1.1
40	LaNiO <sub>3</sub>	14.1	103.2	56	1.1
41	LaNiO <sub>3</sub>	12.7	103.2	56	1.1
42	LaNiO <sub>3</sub>	11.8	103.2	56	1.1
43	LaNiO <sub>3</sub>	6.5	103.2	56	1.1
44	LaNiO <sub>3</sub>	15.1	103.2	56	1.1
45	LaNiO <sub>3</sub>	12.2	103.2	56	1.1
46	LaFeO <sub>3</sub>	21.9	103.2	55.0	1.1
47	LaFeO <sub>3</sub>	15.4	103.2	55.0	1.1
48	LaFeO <sub>3</sub>	10.1	103.2	55.0	1.1
49	LaFeO <sub>3</sub>	5.2	103.2	55.0	1.1
50	LaFeO <sub>3</sub>	1.1	103.2	55	1.1
NO.	Eb	TF	aO3	rc	Za
1	1.54	0.731	160.524	-21.762	9.394
2	1.83	0.882	132.552	11.855	5.577
3	1.83	0.881	132.55	11.727	7.286
4	1.7865	0.873	136.795	9.959	7.286
5	1.88	0.884	131.373	12.341	5.577
6	1.8458	0.879	133.849	11.318	5.577
7	1.823	0.876	135.5	10.631	5.577
8	1.766	0.869	139.627	8.898	5.577
9	1.83	0.888	132.569	13.005	6.031
10	1.83	0.888	132.569	13.005	6.031
11	1.83	0.888	132.569	13.005	6.031
12	1.83	0.888	132.569	13.005	6.031
13	1.83	0.888	132.569	13.005	6.031

Using Data mining to search for perovskite Materials with Higher Specific Surface Area

14	1.83	0.882	132.552	11.855	5.577
15	1.726	0.867	140.571	8.499	5.577
16	1.622	0.852	148.589	5.063	5.577
17	1.518	0.838	156.608	1.55	5.577
18	1.414	0.824	164.626	-2.036	5.577
19	1.31	0.811	172.644	-5.692	5.577
20	1.66	0.853	147.882	5.369	5.577
21	1.59	0.845	152.834	3.212	5.577
22	1.52	0.836	157.787	1.027	5.577
23	1.45	0.828	162.739	-1.186	5.577
24	1.38	0.819	167.692	-3.426	5.577
25	1.83	0.867	132.512	9.147	5.473
26	1.835	0.882	132.434	11.904	5.577
27	1.875	0.884	131.491	12.293	5.577
28	1.83	0.882	132.552	11.855	5.577
29	1.54	0.881	160.991	9.581	5.695
30	1.54	0.881	160.99	9.562	5.695
31	1.54	0.881	160.99	9.533	5.694
32	1.54	0.881	160.989	9.484	5.694
33	1.54	0.88	160.988	9.388	5.693
34	1.83	0.882	132.552	11.855	5.577
35	1.83	0.882	132.552	11.855	5.577
36	1.83	0.882	132.552	11.855	5.577
37	1.83	0.882	132.552	11.855	5.577
38	1.83	0.882	132.552	11.855	5.577
39	1.83	0.882	132.552	11.855	5.577
40	1.91	0.877	134.911	10.876	5.577
41	1.91	0.877	134.911	10.876	5.577
42	1.91	0.877	134.911	10.876	5.577
43	1.91	0.877	134.911	10.876	5.577
44	1.91	0.877	134.911	10.876	5.577
45	1.91	0.877	134.911	10.876	5.577
46	1.8	0.9	132.6	11.6	5.6
47	1.8	0.9	132.9	11.9	5.6
48	1.8	0.9	132.9	11.9	5.6
49	1.8	0.9	132.9	11.9	5.6
50	1.83	0.882	132.911	11.855	5.577
<b>NO.</b>	<b>Zb</b>	<b>Ra/Rb</b>	<b>Mass</b>	<b>A-aff</b>	<b>B-aff</b>
1	6.828	1.104	161.27	-58	7.6
2	7.902	1.876	242.75	48	15.7
3	7.902	1.873	312.85	91.3	15.7
4	7.741	1.813	311.6545	91.3	14.485
5	7.881	1.894	245.83	48	63.8
6	7.867	1.858	243.7528	48	57.632
7	7.858	1.835	242.368	48	53.52
8	7.834	1.779	238.906	48	43.24
9	7.902	1.909	248.054	36.46	15.7
10	7.902	1.909	248.054	36.46	15.7
11	7.902	1.909	248.054	36.46	15.7
12	7.902	1.909	248.054	36.46	15.7
13	7.902	1.909	248.054	36.46	15.7
14	7.902	1.876	242.75	48	15.7
15	7.851	1.767	236.442	48	4.76
16	7.8	1.67	230.134	48	-6.18
17	7.749	1.583	223.826	48	-17.12
18	7.697	1.504	217.518	48	-28.06
19	7.646	1.433	211.21	48	-39
20	6.767	1.678	238.9	48	64.3
21	6.942	1.623	233.362	48	43.64
22	7.118	1.571	227.824	48	22.98
23	7.294	1.522	222.286	48	2.32
24	7.47	1.476	216.748	48	-18.34
25	7.902	1.8	244.75	47	15.7
26	7.9	1.878	243.058	48	20.51
27	7.883	1.892	245.522	48	58.99
28	7.902	1.876	242.75	48	15.7
29	6.828	1.761	183.5	-29	7.6
30	6.828	1.761	183.60256	-28.846	7.6

Using Data mining to search for perovskite Materials with Higher Specific Surface Area

31	6.828	1.76	183.7564	-28.615	7.6
32	6.828	1.759	184.0128	-28.23	7.6
33	6.828	1.757	184.5256	-27.46	7.6
34	7.902	1.876	242.75	48	15.7
35	7.902	1.876	242.75	48	15.7
36	7.902	1.876	242.75	48	15.7
37	7.902	1.876	242.75	48	15.7
38	7.902	1.876	242.75	48	15.7
39	7.902	1.876	242.75	48	15.7
40	7.64	1.843	245.59	48	111.5
41	7.64	1.843	245.59	48	111.5
42	7.64	1.843	245.59	48	111.5
43	7.64	1.843	245.59	48	111.5
44	7.64	1.843	245.59	48	111.5
45	7.64	1.843	245.59	48	111.5
46	7.9	1.9	242.8	48	15.7
47	7.9	1.9	242.8	48	15.7
48	7.9	1.9	242.8	48	15.7
49	7.9	1.9	242.8	48	15.7
50	7.902	1.876	242.75	48.00	15.70
<b>NO.</b>	<b>A-Tm</b>	<b>B-Tm</b>	<b>A-Tb</b>	<b>B-Tb</b>	<b>A_Hfus</b>
1	419.53	1670	907	3287	108.1
2	918	1538	3464	2861	44.6
3	271.4	1538	1564	2861	53.3
4	271.4	1557.8	1564	2924.9	53.3
5	918	1495	3464	2927	44.6
6	918	1444.3	3464	2816.78	44.6
7	918	1410.5	3464	2743.3	44.6
8	918	1326	3464	2559.6	44.6
9	783.28	1538	2626.7	2861	66.82
10	783.28	1538	2626.7	2861	66.82
11	783.28	1538	2626.7	2861	66.82
12	783.28	1538	2626.7	2861	66.82
13	783.28	1538	2626.7	2861	66.82
14	918	1538	3464	2861	44.6
15	918	1360.4	3464	2506.8	44.6
16	918	1182.8	3464	2152.6	44.6
17	918	1005.2	3464	1798.4	44.6
18	918	827.6	3464	1444.2	44.6
19	918	650	3464	1090	44.6
20	918	1907	3464	2671	44.6
21	918	1655.6	3464	2354.8	44.6
22	918	1404.2	3464	2038.6	44.6
23	918	1152.8	3464	1722.4	44.6
24	918	901.4	3464	1406.2	44.6
25	931	1538	3520	2861	48.9
26	918	1533.7	3464	2867.6	44.6
27	918	1499.3	3464	2920.4	44.6
28	918	1538	3464	2861	44.6
29	777	1670	1382	3287	84.8
30	777.282	1670	1386.164	3287	84.7196
31	777.705	1670	1392.41	3287	84.599
32	778.41	1670	1402.82	3287	84.398
33	779.82	1670	1423.64	3287	83.996
34	918	1538	3464	2861	44.6
35	918	1538	3464	2861	44.6
36	918	1538	3464	2861	44.6
37	918	1538	3464	2861	44.6
38	918	1538	3464	2861	44.6
39	918	1538	3464	2861	44.6
40	918	1455	3464	2931	44.6
41	918	1455	3464	2931	44.6
42	918	1455	3464	2931	44.6
43	918	1455	3464	2931	44.6
44	918	1455	3464	2931	44.6

Using Data mining to search for perovskite Materials with Higher Specific Surface Area

45	918	1455	3464	2931	44.6	
46	918	1538	3464	2861	44.6	
47	918	1538	3464	2861	44.6	
48	918	1538	3464	2861	44.6	
49	918	1538	3464	2861	44.6	
50	918.00	1538.00	3464.00	2861.00	44.60	
<b>NO.</b>	<b>B_Hfus</b>	<b>D-A</b>	<b>D-B</b>	<b>CT</b>	<b>AH</b>	<b>DT</b>
1	295.6	7.14	4.51	900	2	120
2	247.3	6.15	7.87	900	4	150
3	247.3	9.79	7.87	900	4	150
4	254.545	9.79	7.366	900	4	150
5	272.5	6.15	8.86	750	4	110
6	277.084	6.15	8.4328	750	4	110
7	280.14	6.15	8.148	750	4	110
8	287.78	6.15	7.436	750	4	110
9	247.3	6.487	7.87	500	4	120
10	247.3	6.487	7.87	600	4	120
11	247.3	6.487	7.87	700	4	120
12	247.3	6.487	7.87	800	4	120
13	247.3	6.487	7.87	900	4	120
14	247.3	6.15	7.87	600	5	100
15	267.62	6.15	6.644	600	5	100
16	287.94	6.15	5.418	600	5	100
17	308.26	6.15	4.192	600	5	100
18	328.58	6.15	2.966	600	5	100
19	348.9	6.15	1.74	600	5	100
20	404	6.15	7.15	600	5	100
21	392.98	6.15	6.068	600	5	100
22	381.96	6.15	4.986	600	5	100
23	370.94	6.15	3.904	600	5	100
24	359.92	6.15	2.822	600	5	100
25	247.3	6.77	7.87	700	5	90
26	249.82	6.15	7.969	750	10	110
27	269.98	6.15	8.761	750	10	110
28	247.3	6.15	7.87	700	3	110
29	295.6	2.64	4.51	650	10	110
30	295.6	2.64702	4.51	650	10	110
31	295.6	2.65755	4.51	650	10	110
32	295.6	2.6751	4.51	650	10	110
33	295.6	2.7102	4.51	650	10	110
34	247.3	6.15	7.87	700	4	90
35	247.3	6.15	7.87	500	2	130
36	247.3	6.15	7.87	600	2	130
37	247.3	6.15	7.87	700	2	130
38	247.3	6.15	7.87	800	2	130
39	247.3	6.15	7.87	900	2	130
40	290.3	6.15	8.9	600	2	130
41	290.3	6.15	8.9	700	2	130
42	290.3	6.15	8.9	800	2	130
43	290.3	6.15	8.9	900	2	130
44	290.3	6.15	8.9	600	4	130
45	290.3	6.15	8.9	600	6	130
46	247.3	6.15	7.87	500	4	120
47	247.3	6.15	7.87	600	4	120
48	247.3	6.15	7.87	700	4	120
49	247.3	6.15	7.87	800	4	120
50	247.30	6.15	7.87	900.00	4.00	120.00



Table.II.2.The list of 24 candidate features

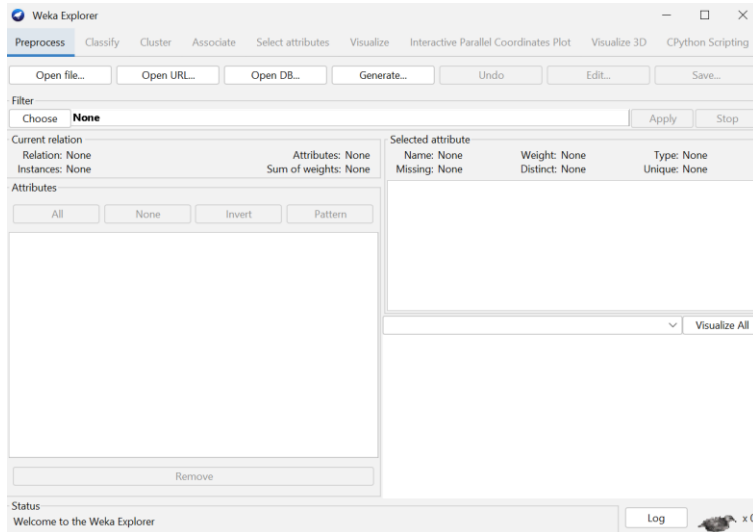
1	Atomic radius of the A position (Ra)
2	Atomic radius of the B position (Rb)
3	Electronegativity of the A position (Ea)
4	Electronegativity of the B position (Eb)
5	Unit cell lattice edge ( $aO3$ ).
6	Critical radius (rc)
7	Ionization potential of the A position (Za)
8	Ionization potential of the B position (Zb)
9	Ratio of the atomic radii of the A and B positions (Ra/Rb)
10	Molecular mass (mass)
11	Electron affinity of the A position (A-aff)
12	Electron affinity of the B position (B-aff)
13	Melting point of the A position (A-Tm)
14	Melting point of the B position (B-Tm)
15	Normal boiling point of the A position (A-Tb)
16	Normal boiling point of the B position (B-Tb)
17	Enthalpy of fusion at the melting point of the A position (A-Hfus)
18	Enthalpy of fusion at the melting point of the B position (B-Hfus)
19	Density of the A position (D-A)
20	Density of the B position (D-B)
21	Calcination temperature (CT)
22	Calcination time (AH)
23	Drying temperature (DT)
24	Tolerance factor (TF)

So in order to begin the work the first step is to import the raw-data into Weka to begin pre-processing phase:

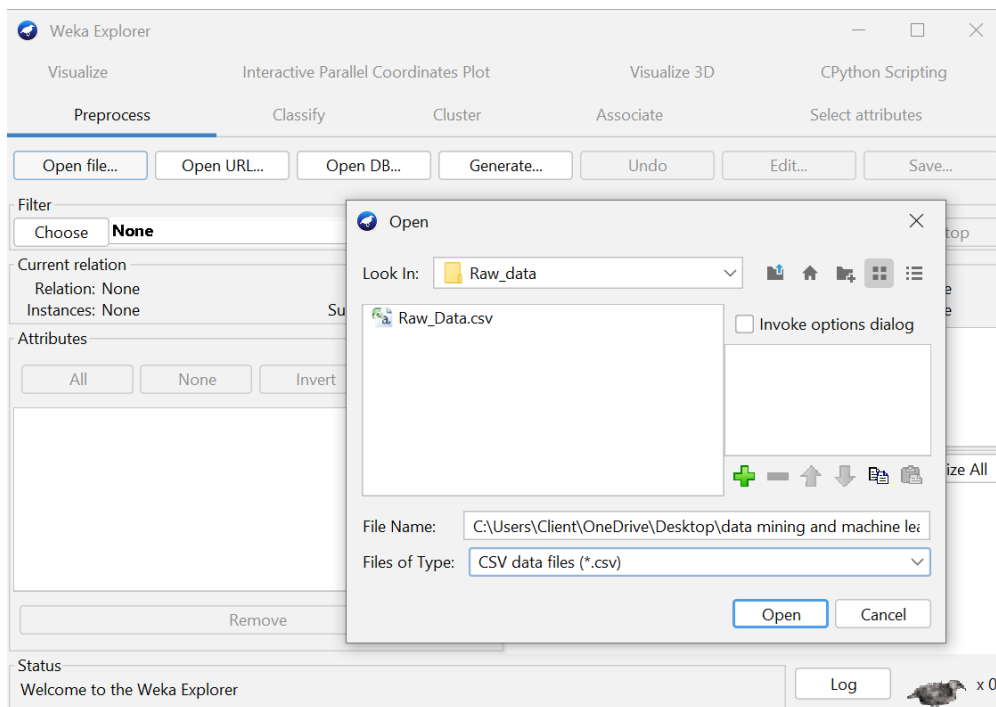
1. Opening Weka and selecting the explorer option



2. Next another window will pop up its the explorer interface



3. In order to open the data we need to clique on open file, a small window will open asking us for the data file directory we locate where we put the file then clique open; a small note will open. Weka can import different file extension we can convert to any type supported by Weka so pay intension our data file called raw\_data is a .csv file :



4. Now we can see the list of the candidate features that are relevant to the dataset appear in the list section, besides the number of samples we have. We can see also some information about each attribute (feature) like the general distribution there are no missing values, the data is in good shape, besides the type of values (numeric and nominal) we can see the minimum mean and maximum value for any of the features.

In the case of missing or incomplete value we can add values by going to the edit option to see our dataset in a single tabular form.

There are no missing values in each feature

List of candidate features

The general distribution of the dataset

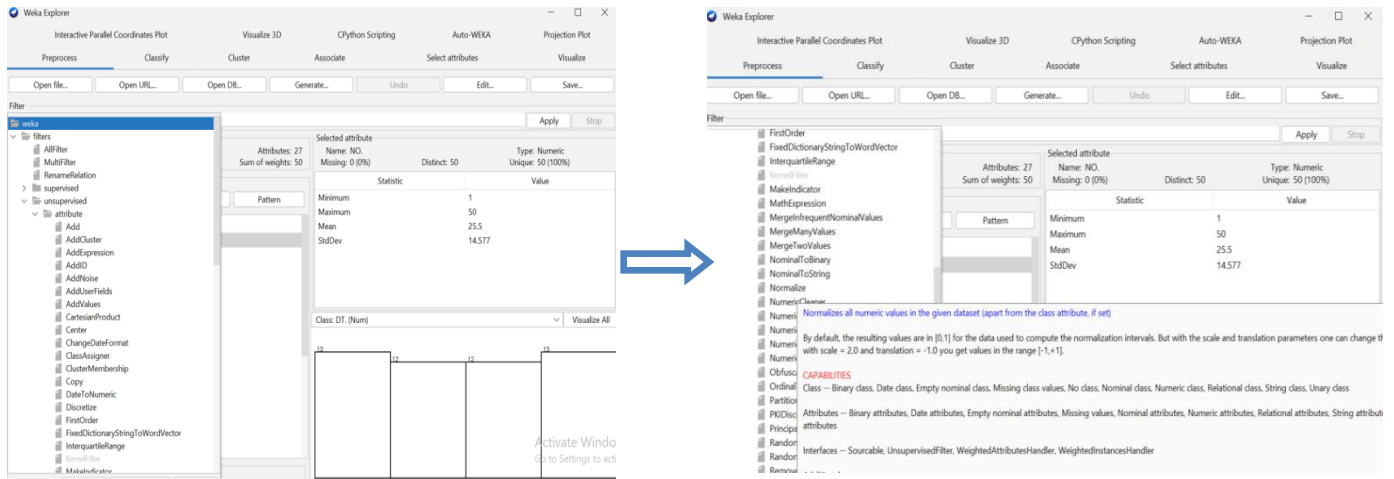
Add additional samples (instances) if necessary

No.	1: NO	2: Molecular	3: SSA	4: Ra	5: Rb	6: Ea	7: Eb	8: TF	9: aO3	10: rc	11: Za	12: Zb	13: Ra/Rb	14: Mass	15: A-aff	16: B-aff	17: A-Tm
1	1.0	ZnTiO3	1.05	74.0	67.0	1.65	1.54	0.731	160.5	-21.762	9.394	6.828	1.104	161.27	-58.0	7.6	419.53
2	2.0	LaFeO3	1.08	103.2	55.0	1.1	1.83	0.882	132.5	11.855	5.577	7.902	1.876	242.75	48.0	15.7	918.0
3	3.0	BiFeO3	0.7514	103.0	55.0	2.02	1.83	0.881	132.55	11.727	7.286	7.902	1.873	312.85	91.3	15.7	271.4
4	4.0	BiTiO15FeO...	0.9507	103.0	56.8	2.02	1.7865	0.873	136.7...	9.959	7.286	7.741	1.813	311.6545	91.3	14.485	271.4
5	5.0	LaCoO3	1.70	103.2	54.5	1.1	1.88	0.884	131.3...	12.341	5.577	7.881	1.894	245.83	48.0	6.38	918.0
6	6.0	LaCo0.94M...	1.90	103.2	55.55	1.1	1.8458	0.879	133.8...	11.318	5.577	7.867	1.858	243.7528	48.0	57.632	918.0
7	7.0	LaCo0.90M...	2.10	103.2	56.25	1.1	1.823	0.876	135.5	10.631	5.577	7.858	1.835	242.368	48.0	53.52	918.0
8	8.0	LaCo0.80M...	2.20	103.2	58.0	1.1	1.766	0.869	139.6...	8.898	5.577	7.834	1.779	238.906	48.0	43.24	918.0
9	9.0	La0.580.2B...	27.75	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28
10	10.0	La0.580.2B...	20.63	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28
11	11.0	La0.580.2B...	12.46	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28
12	12.0	La0.580.2B...	5.91	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28
13	13.0	La0.580.2B...	4.19	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28
14	14.0	LaFeO3	1.139	103.2	55.0	1.1	1.83	0.882	132.5...	11.855	5.577	7.902	1.876	242.75	48.0	15.7	918.0
15	15.0	LaMg0.2 Fe...	15.07	103.2	58.4	1.1	1.726	0.867	140.5...	8.499	5.577	7.851	1.767	236.442	48.0	4.76	918.0
16	16.0	LaMg0.4Fe...	17.63	103.2	61.8	1.1	1.622	0.852	148.5...	5.063	5.577	7.8	1.67	230.134	48.0	-6.18	918.0
17	17.0	LaMg0.6Fe...	24.41	103.2	65.2	1.1	1.518	0.838	156.6...	1.55	5.577	7.749	1.583	223.826	48.0	-17.12	918.0
18	18.0	LaMg0.8Fe...	13.32	103.2	68.6	1.1	1.414	0.824	164.6...	-2.036	5.577	7.697	1.504	217.518	48.0	-28.06	918.0
19	19.0	LaMgO3	8.65	103.2	72.0	1.1	1.31	0.811	172.6...	-5.692	5.577	7.646	1.433	211.21	48.0	-39.0	918.0
20	20.0	La0.6CoO3	3.95	103.2	61.5	1.1	1.66	0.853	147.8...	5.369	5.577	7.677	1.678	238.9	48.0	6.38	918.0

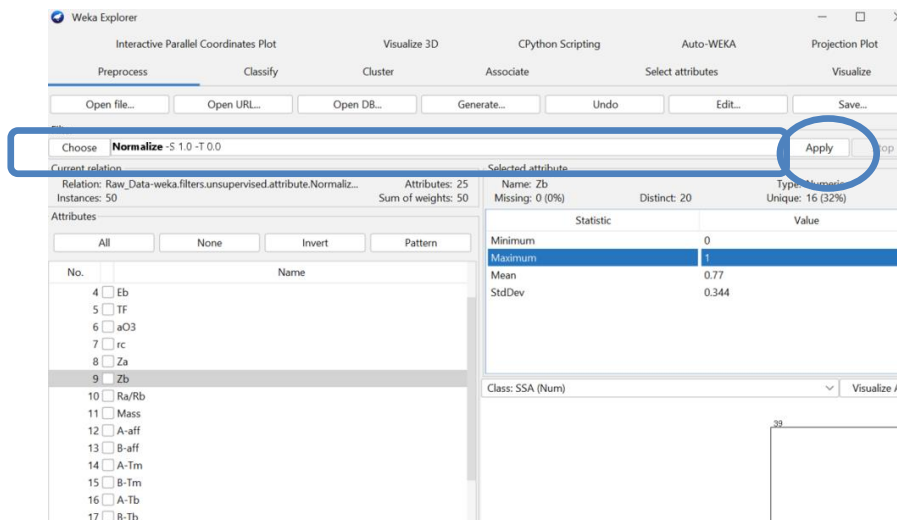
- Now we need to identify the target feature which we will build upon the predictive model the target property is the specific surface area (SSA)  $m^2.m^{-1}$  in order to identify as a target attribute we need to open our data by clicking on edit option than go to the column where SSA situated then selecting attribute as class (target). The target feature SSA will appear in the last column in bold:

No.	1: NO	2: Molecular	3: SSA	4: Ra	5: Rb	6: Ea	7: Eb	8: TF	9: aO3	10: rc	11: Za	12: Zb	13: Ra/Rb	14: Mass	15: A-aff	16: B-aff	17: A-Tm	18: B-Tm
1	1.0	ZnTiO3	1.05	74.0	67.0	1.65	1.54	0.731	160.5	-21.762	9.394	6.828	1.104	161.27	-58.0	7.6	419.53	1670.0
2	2.0	LaFeO3	1.08	103.2	55.0	1.1	1.83	0.882	132.5	11.855	5.577	7.902	1.876	242.75	48.0	15.7	918.0	1538.0
3	3.0	BiFeO3	0.7514	103.0	55.0	2.02	1.83	0.881	132.55	11.727	7.286	7.902	1.873	312.85	91.3	15.7	271.4	1538.0
4	4.0	BiTiO15FeO...	0.9507	103.0	56.8	2.02	1.7865	0.873	136.7...	9.959	7.286	7.741	1.813	311.6545	91.3	14.485	271.4	1557.8
5	5.0	LaCoO3	1.70	103.2	54.5	1.1	1.88	0.884	131.3...	12.341	5.577	7.881	1.894	245.83	48.0	6.38	918.0	1495.0
6	6.0	LaCo0.94M...	1.90	103.2	55.55	1.1	1.8458	0.879	133.8...	11.318	5.577	7.867	1.858	243.7528	48.0	57.632	918.0	1444.3
7	7.0	LaCo0.90M...	2.10	103.2	56.25	1.1	1.823	0.876	135.5	10.631	5.577	7.858	1.835	242.368	48.0	53.52	918.0	1410.5
8	8.0	LaCo0.80M...	2.20	103.2	58.0	1.1	1.766	0.869	139.6...	8.898	5.577	7.834	1.779	238.906	48.0	43.24	918.0	1326.0
9	9.0	La0.580.2B...	27.75	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28	1538.0
10	10.0	La0.580.2B...	20.63	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28	1538.0
11	11.0	La0.580.2B...	12.46	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28	1538.0
12	12.0	La0.580.2B...	5.91	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28	1538.0
13	13.0	La0.580.2B...	4.19	105.0	55.0	1.287	1.83	0.888	132.5...	13.005	6.031	7.902	1.909	248.054	36.46	15.7	783.28	1538.0
14	14.0	LaFeO3	1.139	103.2	55.0	1.1	1.83	0.882	132.5...	11.855	5.577	7.902	1.876	242.75	48.0	15.7	918.0	1538.0
15	15.0	LaMg0.2 Fe...	15.07	103.2	58.4	1.1	1.726	0.867	140.5...	8.499	5.577	7.851	1.767	236.442	48.0	4.76	918.0	1360.4
16	16.0	LaMg0.4Fe...	17.63	103.2	61.8	1.1	1.622	0.852	148.5...	5.063	5.577	7.8	1.67	230.134	48.0	-6.18	918.0	1182.8
17	17.0	LaMg0.6Fe...	24.41	103.2	65.2	1.1	1.518	0.838	156.6...	1.55	5.577	7.749	1.583	223.826	48.0	-17.12	918.0	1005.2
18	18.0	LaMg0.8Fe...	13.32	103.2	68.6	1.1	1.414	0.824	164.6...	-2.036	5.577	7.697	1.504	217.518	48.0	-28.06	918.0	827.6
19	19.0	LaMgO3	8.65	103.2	72.0	1.1	1.31	0.811	172.6...	-5.692	5.577	7.646	1.433	211.21	48.0	-39.0	918.0	650.0
20	20.0	LaCoO3	3.95	103.2	61.5	1.1	1.66	0.853	147.8...	5.369	5.577	7.677	1.678	238.9	48.0	6.38	918.0	1907.0
21	21.0	LaMg0.2Co...	8.42	103.2	63.6	1.1	1.59	0.845	152.8...	3.212	5.577	6.942	1.623	233.362	48.0	43.64	918.0	1655.6
22	22.0	LaMg0.4Co...	29.71	103.2	65.7	1.1	1.52	0.836	157.7...	1.027	5.577	7.118	1.571	227.824	48.0	22.98	918.0	1104.2
23	23.0	LaMg0.6Co...	18.41	103.2	67.8	1.1	1.45	0.828	162.7...	-1.186	5.577	7.294	1.522	222.286	48.0	2.32	918.0	1152.8
24	24.0	LaMg0.8Co...	11.15	103.2	70.0	1.1	1.35	0.819	170.8...	-3.222	5.577	7.246	1.478	215.218	48.0	-11.12	918.0	1018.0
25	25.0	LaMgO3	8.65	103.2	72.0	1.1	1.31	0.811	172.6...	-5.692	5.577	7.646	1.433	211.21	48.0	-39.0	918.0	650.0
26	26.0	LaCoO3	3.95	103.2	61.5	1.1	1.66	0.853	147.8...	5.369	5.577	7.677	1.678	238.9	48.0	6.38	918.0	1907.0
27	27.0	LaMg0.2Co...	8.42	103.2	63.6	1.1	1.59	0.845	152.8...	3.212	5.577	6.942	1.623	233.362	48.0	43.64	918.0	1655.6
28	28.0	LaMg0.4Co...	29.71	103.2	65.7	1.1	1.52	0.836	157.7...	1.027	5.577	7.118	1.571	227.824	48.0	22.98	918.0	1104.2
29	29.0	LaMg0.6Co...	18.41	103.2	67.8	1.1	1.45	0.828	162.7...	-1.186	5.577	7.294	1.522	222.286	48.0	2.32	918.0	1152.8
30	30.0	LaMg0.8Co...	11.15	103.2	70.0	1.1	1.35	0.819	170.8...	-3.222	5.577	7.246	1.478	215.218	48.0	-11.12	918.0	1018.0

- We need to normalize and scale our dataset in order to bring all features to the same range by giving them values ranging from 0 to 1 this is an important step it improve the model accuracy[6].In order to begin these steps we need to go to the filter option than choose unsupervised then attribute then selecting normalize option:



The normalize filter option will then appear in the bare next to choose option we can modify and edit the properties of this filter but we recommend leaving the default parameters, to activate this filter we need to clique on apply you will notice that the values of each features will have a minimum value of 0 and a maximum value of 1:



## II.2.Feature selection:

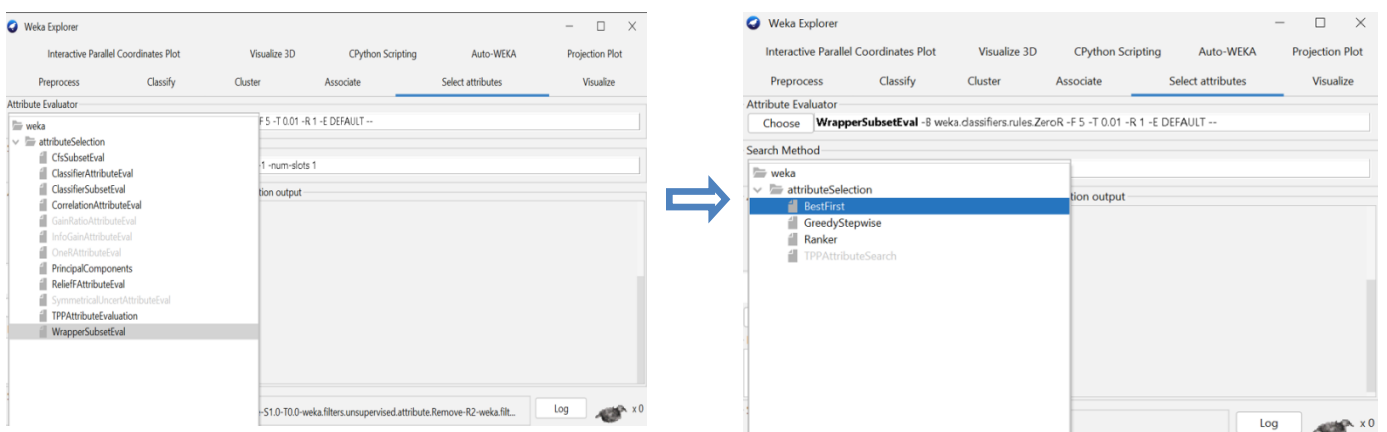
Now that the data is prepared the next important step is selecting the key features that influence heavily the target property. This will ensure an effective data training and avoid over-fitting problems by reducing the quantity of the features.

In order to use this option we need to go to the attribute selection section we will find in front of us two dialogue bars one called attribute evaluator and the other one is search method. Weka uses different feature selection methods the one we choose called wrapper subset Evaluation with a search method called best first.

Wrapper methods utilize the prediction ability of some machine learning algorithms (in our case SVR), to evaluate the feature subset. Wrapper method can assure to get a feature subset with higher accuracy by using the specified learning machine (SVR here).

Wrapper algorithm examines the feature space to qualify subsets of features according to their predictive power and optimizes the subsequent induction algorithm that uses each subset for classification. [7]

First we will begin by choosing from the attribute evaluator bar the wrapper subset Evaluation or [WrapperEvalssubset](#), Weka automatically will select the right search method in this case its [best first](#) search method:



Then we come at an important step which is selecting the algorithm the wrapper method will work on. We choose the Support vector machine for regression (SVM for regression) also known as SMOreg in Weka it's the main algorithm that we will deploy later to train our data and build the model.

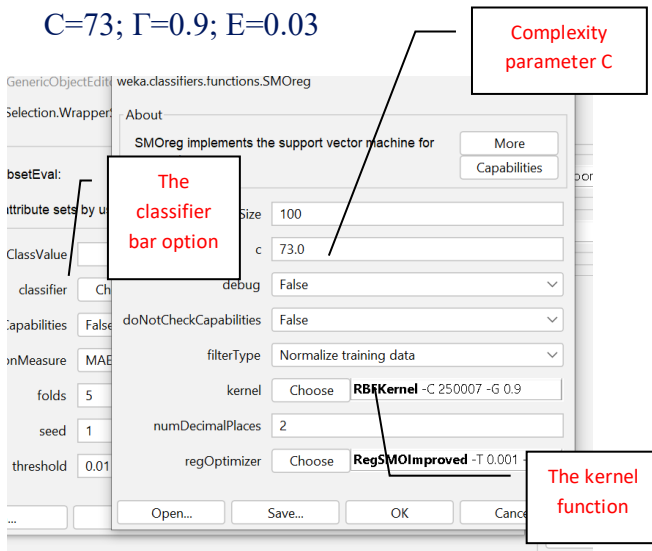
To choose SVM for regression algorithm we need to clique on WrapperSubsetEval bar an object editor will appear then go for classifiers, the functions option and finally SMOreg which represent the Support vector machine for regression algorithm:

To select the best possible subset of features we need to make sure that the SVR algorithm is well optimized by carefully choosing the right parameters which are the kernel function the complexity parameter C the gamma parameter (the kernel parameter)  $\gamma$  and the epsilon parameter  $\epsilon$ , without forgetting normalize the training data filter. Each of these parameters needs to set at these values according to [4] to avoid over fitting and under fitting problems, they achieved that by performing a grid search [8] and evaluating the LOOCV results for SVR models.

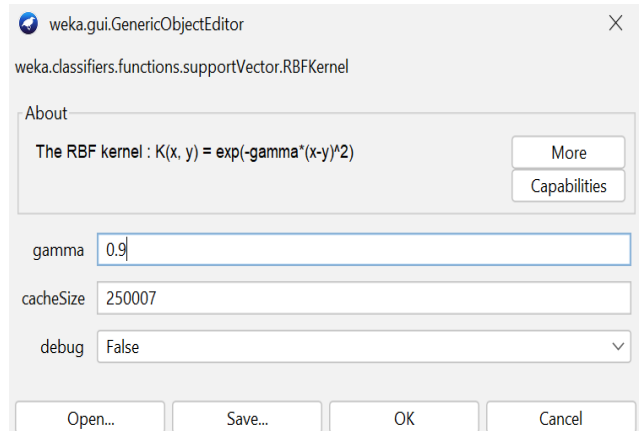
Another study shows that the value of these parameters according to the grid search varies according to the kernel function. It shows that the SVR model is well fitted and optimized when the gamma parameter was giving these values (0.001, 0.01, 0.1, 0.5, 0.8, 1, and 3) and the C parameter (5, 10, 20, 50, 100, 200, 500, and 1000). [9]

The kernel function is RBFKernel (the radial basis function kernel)

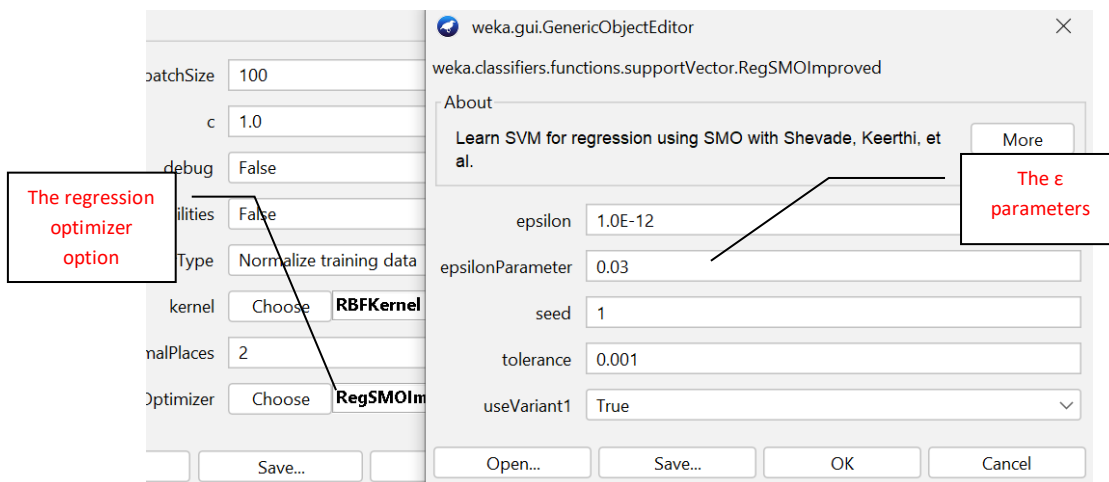
$C=73; \Gamma=0.9; E=0.03$



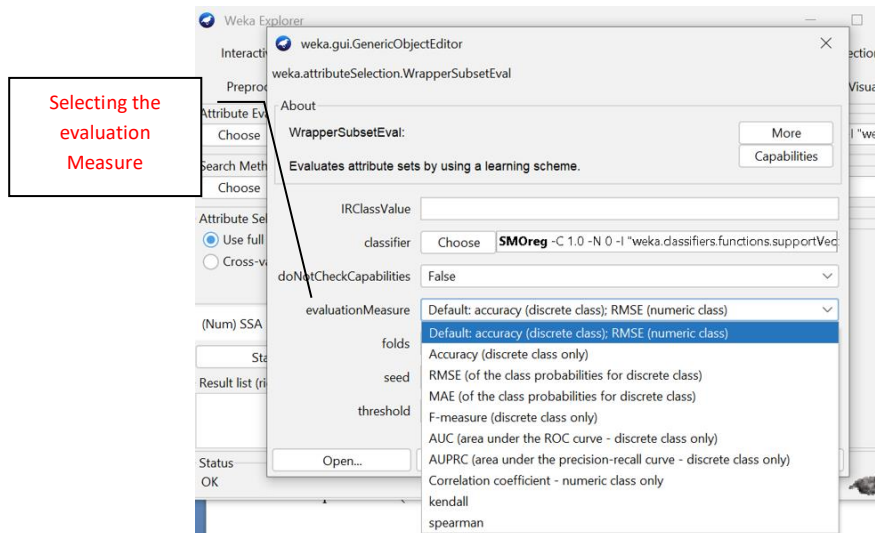
To change  $\gamma$  clique on the RBFKernel bar:



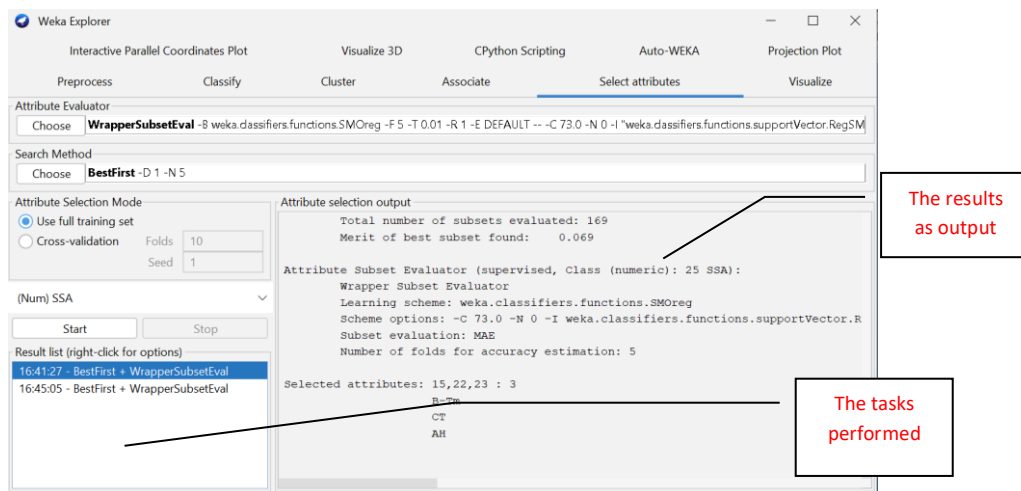
To change the  $\epsilon$  parameters we clique on the regression optimizer option:



Next is selecting the evaluation measure which will evaluate our feature selection method. There are a lot of options for the evaluation metrics, for us we prefer the correlation coefficient R and the root mean square error (RMSE).



We will use both training set data and cross-validation 10 folds options the results are shown in the result panel:



- The results as shown in the output section are:

Table.II.3.Feature selection results using Wrapper method

Number of instances	Selected attributes	RMSE	R
50	B-Tm CT AH	4.624	0.89

The results shows good evaluation by this method we have the best subset of features with R= 0.89 which is good and RMSE= 4.624 minimizing the error.

B-Tm: the melting point of the B-position, CT: The calcinations temperature, AH: the calcinations time.



To further confirm the results we proceed with cross-validation test:

The cross validation results:

Evaluation mode: 10-fold cross-validation

Attribute selection 10 fold cross-validation seed: 1

Table.II.4.Feature selection results with 10 folds cross validation results

Number of folds (%)	Attribute
10	1 Ra
0	2 Rb
20	3 Ea
0	4 Eb
0	5TF
0	6 aO3
0	7 rc
10	8 Za
0	9 Zb
0	10 Ra/Rb
10	11 Mass
10	12 A-aff
30	13 B-aff
0	14 A-Tm
80	15 B-Tm
10	16 A-Tb
0	17 B-Tb
0	18 A Hfus
20	19 B Hfus
0	20 D-A
0	21 D-B
100	22 CT
100	23 AH
30	24 DT.

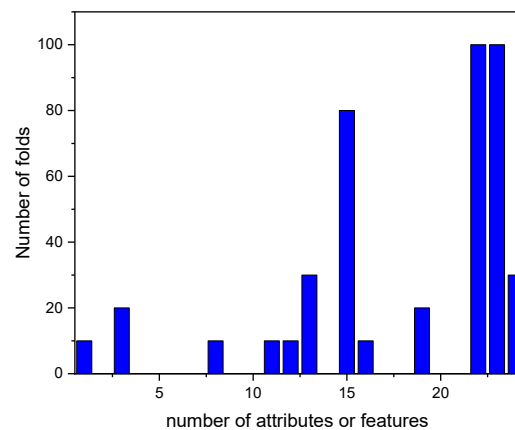


Figure.II.2.A histogram representing the selected features (attributes)

We can clearly see through the 10-folds cross-validation test that the optimal subset of feature according to wrapper method consist of The calcinations temperature (CT) with 100% dominance also the calcinations time (AH) scores 100% and the melting point of the B position (B-Tm) with 80%.this is good because as chemists we know that the calcinations temperature and the calcinations time has a significance impact on the surface area of  $ABO_3$ -type perovskite.

There other descriptors sharing the same number of folds to pick one of them that depend on the chemist's knowledge.

Now we reduced our data to just 3 set of features with the target feature being the SSA (Table.II.5).

Table.II.5.The reduced data well normalized and ready

B-Tm	CT	AH	SSA (The target property)
0.811456	1	0	1.05
0.706444	1	0.25	1.08
0.706444	1	0.25	0.7514
0.722196	1	0.25	0.9507
0.672235	0.625	0.25	17
0.631901	0.625	0.25	19
0.605012	0.625	0.25	21
0.537788	0.625	0.25	22
0.706444	0	0.25	27.75
0.706444	0.25	0.25	20.63
0.706444	0.5	0.25	12.46
0.706444	0.75	0.25	5.91
0.706444	1	0.25	4.19
0.706444	0.25	0.375	11.39
0.565155	0.25	0.375	15.07
0.423866	0.25	0.375	17.63
0.282578	0.25	0.375	24.41
0.141289	0.25	0.375	13.32
0	0.25	0.375	8.65
1	0.25	0.375	3.95
0.8	0.25	0.375	8.42
0.6	0.25	0.375	29.71
0.4	0.25	0.375	18.41
0.2	0.25	0.375	14.46
0.706444	0.5	0.375	10.88
0.703023	0.625	1	51.2
0.675656	0.625	1	42.8
0.706444	0.5	0.125	8.5
0.811456	0.375	1	16.4
0.811456	0.375	1	19.7
0.811456	0.375	1	22.3
0.811456	0.375	1	24.1
0.811456	0.375	1	23.2
0.706444	0.5	0.25	9.5
0.706444	0	0	25.8
0.706444	0.25	0	22.55
0.706444	0.5	0	20.04
0.706444	0.75	0	8.5
0.706444	1	0	5.8
0.640414	0.25	0	14.1
0.640414	0.5	0	12.7
0.640414	0.75	0	11.8
0.640414	1	0	6.5
0.640414	0.25	0.25	15.1
0.640414	0.25	0.5	12.2
0.706444	0	0.25	21.9
0.706444	0.25	0.25	15.4
0.706444	0.5	0.25	10.1
0.706444	0.75	0.25	5.2
0.706444	1	0.25	1.1

## Data n°02:

For this time we will be selecting a different set of features in the bases of Doc.Kenoushe's method which he dividing the raw data into 10000 subdivisions to a get a better generalization of the data and get the best possible set of features besides the selection of features performed 10000 times all this work was realized using a software called math lab. To summarize it here below:

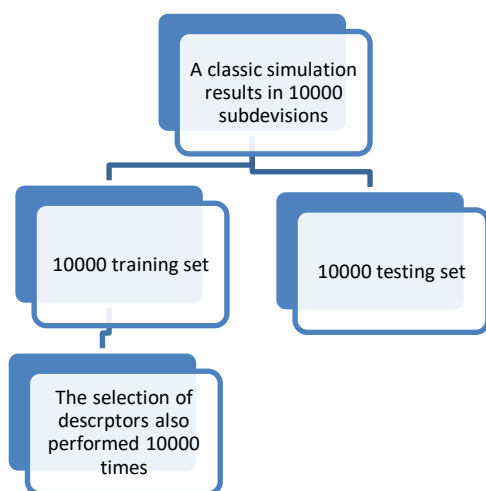


Figure.II.3.A simple figure representing the method

For this method we were only interested in the selection of features which according to this method was like:

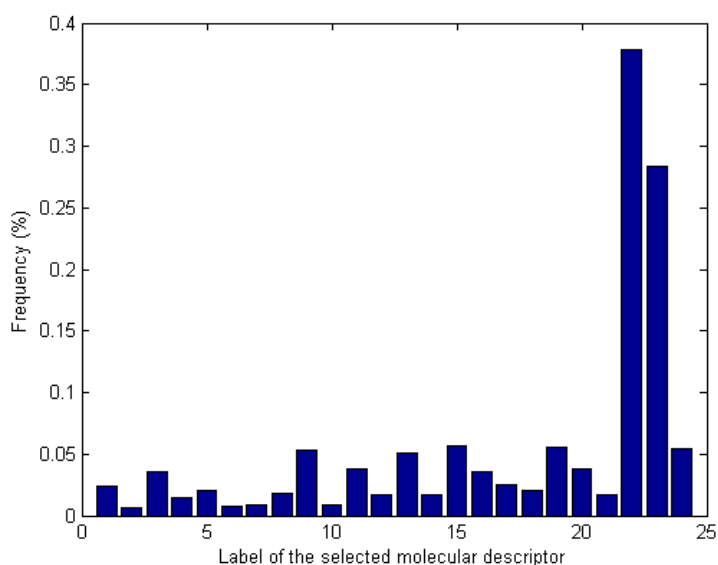


Figure.II.4. A histogram representing the best performed features

We can clearly see according to (Figure.II.3) the best features are the numbers 22 23 which are respectively calcinations temperature and the calcinations time to select the other descriptors I have the number 19, 15, and 10.

19 is the Enthalpy of fusion at the melting point of the B position and 15 is the melting point of the position B the last number 10 which is the Ratio of the atomic radii of the A and B positions.

We selected the same B-Tm, CT, AH but this time adding the Enthalpy of fusion at the melting point of the B position B-Hfus.

So the data n°2 has a features set consisted of B-Tm, CT, AH, and B-Hfus.

Further we normalized the data with the new set features this time not with Weka but by calculating the standard deviation and divide each values of the features on its standard deviation.

$$S_{(B-Tm)} = 222.487709763749$$

$$S_{(B-Hfus)} = 42.2815399718063$$

$$S_{(CT)} = 117.477744309989$$

$$S_{(AH)} = 2.40958968477614$$

Table.II.6.The normalized dataset for the new selected feature

NO.	B-Tm	B_Hfus	CT	AH	SSA
1	7.5060325883767200	6.9912306930426000	7.6610255439120700	0.8300168334202530	1.05
2	6.9127413897744900	5.8488881948221700	7.6610255439120700	1.6600336668405100	1.08
3	6.9127413897744900	5.8488881948221700	7.6610255439120700	1.6600336668405100	0.7514
4	7.0017350695648200	6.0202395695552400	7.6610255439120700	1.6600336668405100	0.9507
5	6.7194722871995200	6.4448929765023900	6.3841879532600600	1.6600336668405100	17
6	6.4915945313727500	6.5533090844080400	6.3841879532600600	1.6600336668405100	19
7	6.3396760274882400	6.6255864896784600	6.3841879532600600	1.6600336668405100	21
8	5.9598797677769600	6.8062800028545300	6.3841879532600600	1.6600336668405100	22
9	6.9127413897744900	5.8488881948221700	4.2561253021733700	1.6600336668405100	27.75
10	6.9127413897744900	5.8488881948221700	5.1073503626080500	1.6600336668405100	20.63
11	6.9127413897744900	5.8488881948221700	5.9585754230427200	1.6600336668405100	12.46
12	6.9127413897744900	5.8488881948221700	6.8098004834773900	1.6600336668405100	5.91
13	6.9127413897744900	5.8488881948221700	7.6610255439120700	1.6600336668405100	4.19
14	6.9127413897744900	5.8488881948221700	5.1073503626080500	2.0750420835506300	11.39
15	6.1144950498369400	6.3294761775103500	5.1073503626080500	2.0750420835506300	15.07
16	5.3162487098993900	6.8100641601985300	5.1073503626080500	2.0750420835506300	17.63
17	4.5180023699618400	7.2906521428867100	5.1073503626080500	2.0750420835506300	24.41
18	3.7197560300242900	7.7712401255748900	5.1073503626080500	2.0750420835506300	13.32
19	2.9215096900867500	8.2518281082630700	5.1073503626080500	2.0750420835506300	8.65
20	8.5712599676852700	9.5549972936035500	5.1073503626080500	2.0750420835506300	3.95
21	7.4413099121655600	9.2943634565354600	5.1073503626080500	2.0750420835506300	8.42
22	6.3113598566458600	9.0337296194673600	5.1073503626080500	2.0750420835506300	29.71
23	5.1814098011261600	8.7730957823992600	5.1073503626080500	2.0750420835506300	18.41
24	4.0514597456064500	8.5124619453311600	5.1073503626080500	2.0750420835506300	14.46
25	6.9127413897744900	5.8488881948221700	5.9585754230427200	2.0750420835506300	10.88
26	6.8934144795169900	5.9084886729902000	6.3841879532600600	4.1500841671012700	51.2
27	6.7387991974570100	6.3852924983343700	6.3841879532600600	4.1500841671012700	42.8
28	6.9127413897744900	5.8488881948221700	5.9585754230427200	1.2450252501303800	8.5
29	7.5060325883767200	6.9912306930426000	5.5329628928253800	4.1500841671012700	16.4
30	7.5060325883767200	6.9912306930426000	5.5329628928253800	4.1500841671012700	19.7
31	7.5060325883767200	6.9912306930426000	5.5329628928253800	4.1500841671012700	22.3
32	7.5060325883767200	6.9912306930426000	5.5329628928253800	4.1500841671012700	24.1
33	7.5060325883767200	6.9912306930426000	5.5329628928253800	4.1500841671012700	23.2
34	6.9127413897744900	5.8488881948221700	5.9585754230427200	1.6600336668405100	9.5
35	6.9127413897744900	5.8488881948221700	4.2561253021733700	0.8300168334202530	25.8
36	6.9127413897744900	5.8488881948221700	5.1073503626080500	0.8300168334202530	22.55
37	6.9127413897744900	5.8488881948221700	5.9585754230427200	0.8300168334202530	20.04
38	6.9127413897744900	5.8488881948221700	6.8098004834773900	0.8300168334202530	8.5
39	6.9127413897744900	5.8488881948221700	7.6610255439120700	0.8300168334202530	5.8
40	6.5396870755018700	6.8658804810225500	5.1073503626080500	0.8300168334202530	14.1
41	6.5396870755018700	6.8658804810225500	5.9585754230427200	0.8300168334202530	12.7
42	6.5396870755018700	6.8658804810225500	6.8098004834773900	0.8300168334202530	11.8
43	6.5396870755018700	6.8658804810225500	7.6610255439120700	0.8300168334202530	6.5
44	6.5396870755018700	6.8658804810225500	5.1073503626080500	1.6600336668405100	15.1
45	6.5396870755018700	6.8658804810225500	5.1073503626080500	2.4900505002607600	12.2
46	6.9127413897744900	5.8488881948221700	4.2561253021733700	1.6600336668405100	21.9
47	6.9127413897744900	5.8488881948221700	5.1073503626080500	1.6600336668405100	15.37
48	6.9127413897744900	5.8488881948221700	5.9585754230427200	1.6600336668405100	10.07
49	6.9127413897744900	5.8488881948221700	6.8098004834773900	1.6600336668405100	5.24
50	6.9127413897744900	5.8488881948221700	7.6610255439120700	1.6600336668405100	1.09

So to summarize:

Table.II.7.the selected features in the two datasets

	Data n°1	Data n°2
The number of samples	50	50
The set of features	B-Tm, CT, AH.	B-Tm, CT, AH, B Hfus

### II.3.Model selection:

#### Data n°1:

##### ✓ Splitting the dataset:

After we reduced the data and made sure it meets and qualifies for a good generalization and effect on the SSA we need to execute one more important step which is splitting the dataset into a training set and a test set generally the splitting is random but we prefer splitting our data into 80% for training and the rest 20% for testing that makes 40 perovskite samples up for training and the rest 10 for testing.

The importance of splitting the data is ensuring the accuracy of the created model you have 80% dataset for building and training the model and the 20% for testing the model.

In Weka splitting the data is made easy let us see how in the following steps:

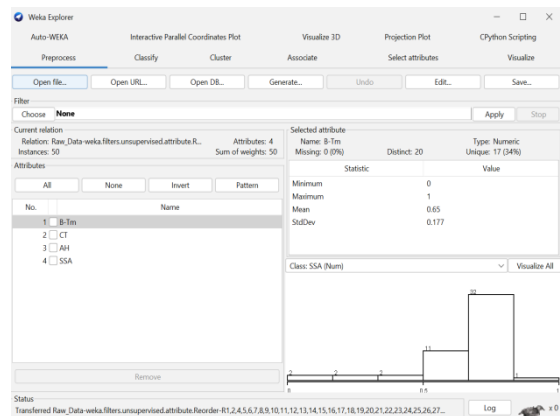


Figure.II.5.the classify section in Weka where we will perform all the tasks

First we need to go to the filter option and choose filters unsupervised then instance option after that select the Remove Percentage filter:

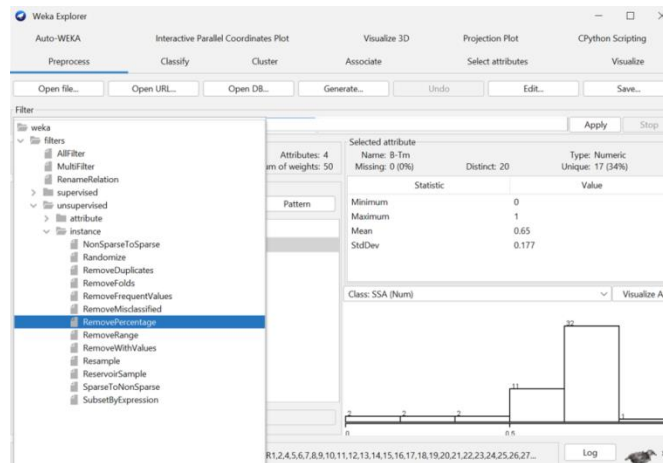


Figure.II.6.choosing Remove percentage Filter

Next we need to configure the filter so that it can remove 80% from the data for training and 20% for testing for that happen we need to clique on the Remove Percentage filter bar an object editor will appear we go for the percentage option and selecting 20:

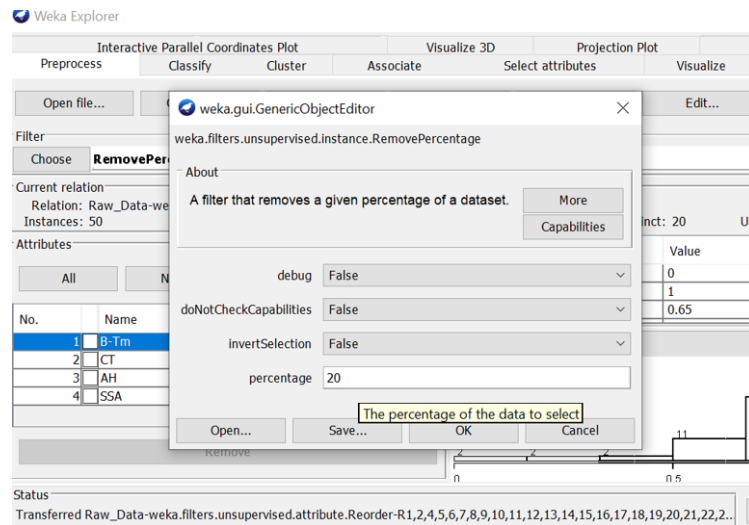


Figure.II.7.Configuring remove percentage filter

This will give the training dataset with 40 samples ready to train. We need to save the training data set by clicking on save option:



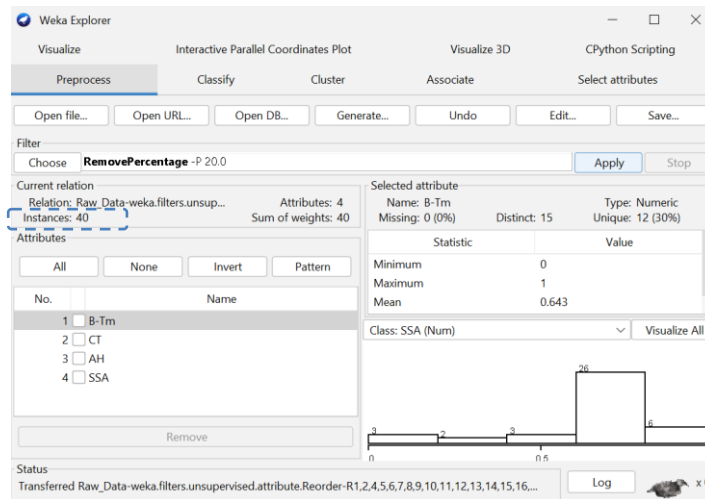


Figure.II.8.The training dataset with 40 samples (instances)

Next is our testing set for that we need to undo what we did early about the training dataset by selecting the undo option after that we go for the same filter option cliquing on Remove Percentage bar we will see the object editor in front of us this time we will go for the invert Selection option and select true this will give us our testing set with 10 samples ready to test on.

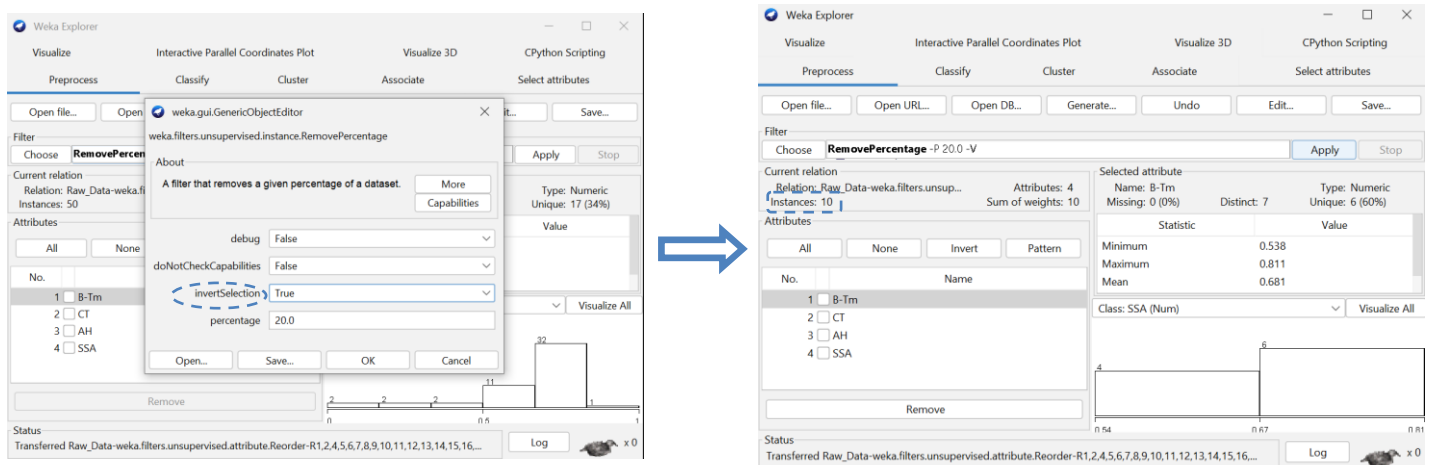


Figure.II.9.Creating the testing set with 10 samples (instances)

Now that we have our training and testing data we are ready to build our predictive model to achieve that we need to fellow these easy steps:

- Opening the training dataset file (the same when we import our raw data)
- We will find in front of us the list of feature that we selected early threw the feature selection method(wrapper method) that is 4 features (attributes) with the target

property SSA (the class), the number of samples which is 40 perovskite type materials that data is well normalized and scaled (Figure.II.7).

- To build our model we need to select the algorithm which will do the work for us the algorithm selected called SMoreg which is the support vector machine for regression (SVR), we are building a regression model using this algorithm to make this happen we need to switch the section and go for the classifier section (Figure.I.4).
- To choose the algorithms we go to the classifier option bar and clique on choose then select function then **SMoreg** algorithm option.

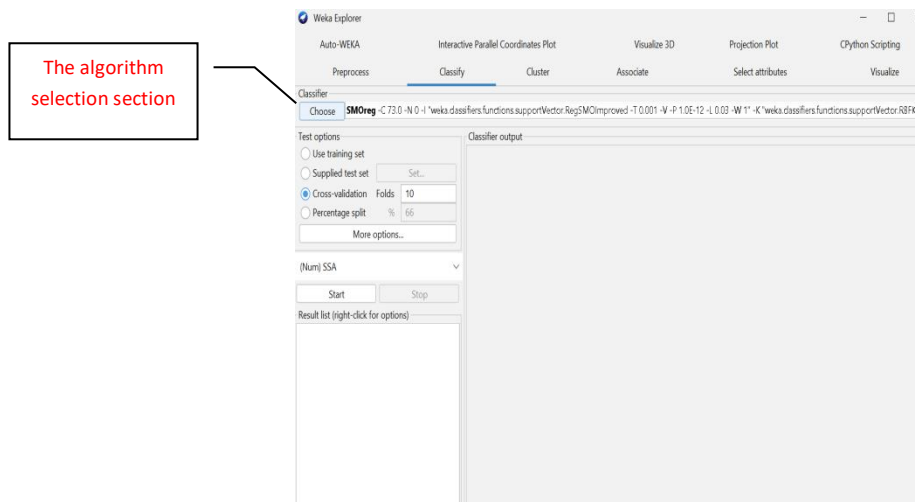


Figure.II.10.selecting the algorithm

- To choose between different metrics to evaluate our model we need to go to more options than on the bottom there are a list of metrics that can help. For us we chose correlation coefficient, Root Mean absolute error, and Root mean squared error.

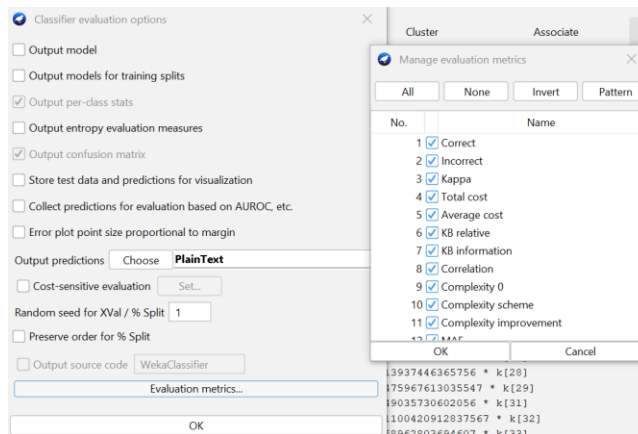


Figure.II.11.Different metrics evaluation for testing the model performance

The building of the model is done

The results appear as output in the classifier output section

We achieved these results:

==== Run information ====

Test mode: evaluate on training data

==== Classifier model (full training set) ====

Number of kernel evaluations: 820 (99.851% cached)

Time taken to build model: 0.01 seconds

==== Evaluation on training set ====

==== Summary ====

- **Correlation coefficient** 0.9451
- Mean absolute error 2.1146
- Root mean squared error 3.2839
- Total Number of Instances 40

➤ To validate our model we need to test it further on the testing set built earlier. We begin our test by clicking on the supplied test set option then set it by locating the file we created earlier for the testing set it will detect automatically the target attribute

(SSA class) then clique on start to start the test:

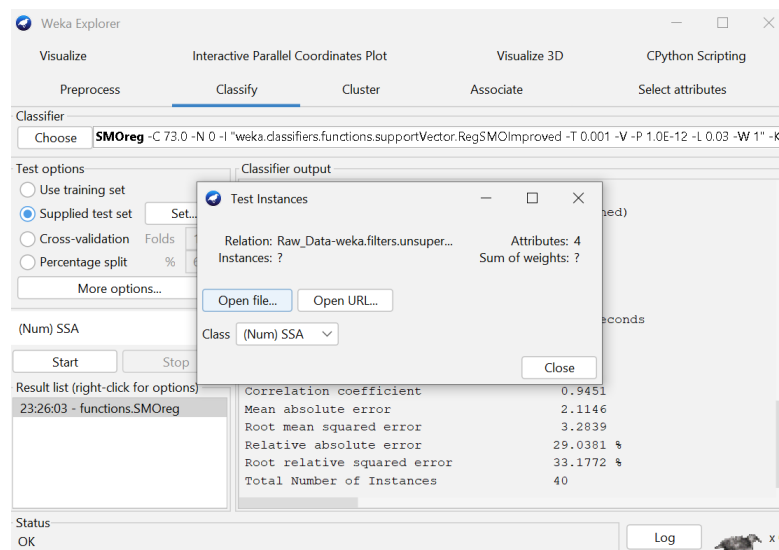


Figure.II.12. Testing the model on the testing set

The testing of the model is done.

The results appear as output in the classifier output section

We achieved these results:

==== Run information ====

Test mode: evaluate on testing data

==== Classifier model (full training set) ====

Number of kernel evaluations: 820 (99.851% cached)

Time taken to build model: 0.01 seconds

==== Evaluation on testing set ====

==== Summary ====

- Correlation coefficient 0.9571
- Mean absolute error 5.4547
- Root mean squared error 6.4127

Total Number of Instances 10

We can see clearly how our model performed well in the training data and in the testing data scoring 0.95 for the training set and 0.96 in the testing set avoiding under fitting and over fitting problems.

- We can visualize our classifier errors and see the error between the predicted and the actual SSA.

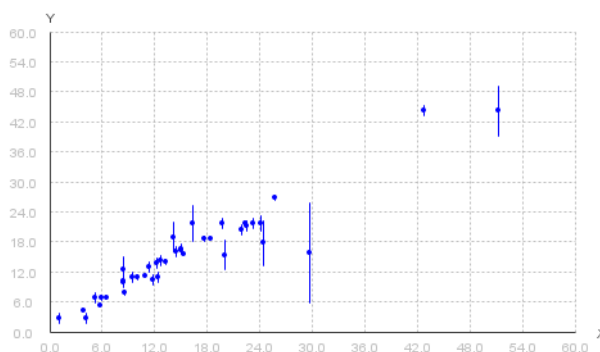


Figure.II.13.a plot visualizing classifiers error on the training set for the y (predicted SSA) and x (actual SSA) using the visualize tool JMathtools.

Besides that we can see the predicted values and the actual values of the SSA of the samples all in this table here:

Table.II.8. the predicted values and the actual values of the SSA of the samples

inst#	actual	predicted	error
1	12.46	10.907	-1.553
2	5.91	6.753	0.843
3	4.19	2.643	-1.547
4	11.39	12.896	1.506
5	15.07	16.613	1.543
6	17.63	18.499	0.869
7	24.41	17.74	-6.67
8	13.32	14.07	0.75
9	8.65	7.873	-0.777
10	3.95	4.16	0.21
11	8.42	9.968	1.548
12	29.71	15.827	-13.883
13	18.41	18.571	0.161
14	14.46	15.939	1.479
15	10.88	11.036	0.156
16	51.2	44.13	-7.07
17	42.8	44.269	1.469
18	8.5	12.44	3.94
19	16.4	21.696	5.296
20	19.7	21.696	1.996
21	22.3	21.696	-0.604
22	24.1	21.696	-2.404
23	23.2	21.696	-1.504
24	9.5	10.907	1.407
25	25.8	26.836	1.036
26	22.55	21.095	-1.455
27	20.04	15.338	-4.702
28	8.5	10.046	1.546
29	5.8	5.281	-0.519
30	14.1	18.954	4.854
31	12.7	14.226	1.526
32	11.8	10.286	-1.514
33	6.5	6.768	0.268
34	15.1	16.178	1.078
35	12.2	13.654	1.454
36	21.9	20.373	-1.527
37	15.4	15.417	0.017
38	10.1	10.907	0.807
39	5.2	6.753	1.553
40	1.1	2.643	1.543

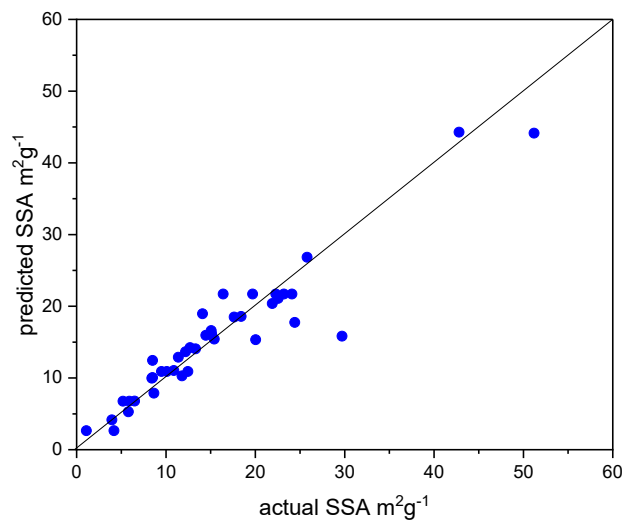


Figure.II.14. a Scatter plot representing Actual SSA vs. predicted SSA for perovskite samples by SVR model using Origin software.

The difference is very small besides the good correlation coefficient and the little difference in the experimental (actual) and predicted values of the samples representing by the RMSE factor .but there are some errors resulting in noticeable differences between the Predicted and actual SSA.

#### Data n°02:

For building the SVR model using the other set of features on the same computational software weka so all the previous steps are the same. The results are:

Number of kernel evaluations: 820 (99.402% cached)

Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

- Correlation coefficient 0.99
- Mean absolute error 0.5812
- Root mean squared error 1.4236

Total Number of Instances 40

Number of kernel evaluations: 820 (99.95% cached)

Time taken to build model: 0.15 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.01 seconds

=== Summary ===

- Correlation coefficient 0.9449
- Mean absolute error 5.7012
- Root mean squared error 6.5545

Total Number of Instances 10

We can see that our second model performed very well on the training set scoring 0.99 and a very good result on the testing set almost 0.95.

We can also visualise the classifier error on the training set and see that it's a very good performance:

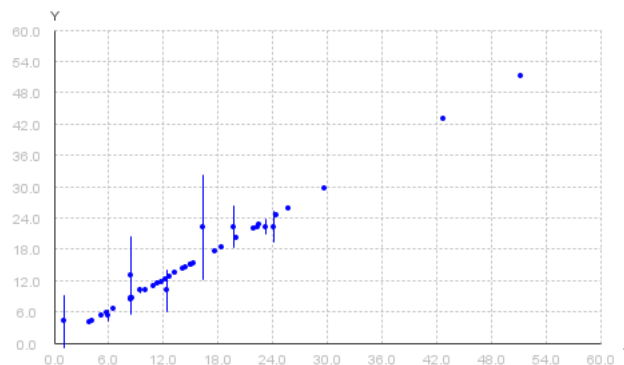


Figure.II.15. a plot visualizing classifiers error on the training set for the y (predicted SSA) and x (actual SSA) for the second model using the visualize tool JMathtools

We can see the predicted values and the actual values of the SSA of the samples all in this table here:

Table.II.9. the predicted values and the actual values of the SSA of the samples for the second model

inst#	actual	predicted	error
1	12.46	10.04	-2.42
2	5.91	5.269	-0.641
3	4.19	4.16	-0.03
4	11.39	11.42	0.03
5	15.07	15.04	-0.03
6	17.63	17.66	0.03
7	24.41	24.38	-0.03
8	13.32	13.35	0.03
9	8.65	8.68	0.03
10	3.95	3.98	0.03
11	8.42	8.45	0.03
12	29.71	29.68	-0.03
13	18.41	18.441	0.031
14	14.46	14.43	-0.03
15	10.88	10.909	0.029
16	51.2	51.17	-0.03
17	42.8	42.83	0.03
18	8.5	12.948	4.448
19	16.4	22.27	5.87
20	19.7	22.27	2.57
21	22.3	22.27	-0.03
22	24.1	22.27	-1.83
23	23.2	22.27	-0.93
24	9.5	10.04	0.54
25	25.8	25.77	-0.03
26	22.55	22.58	0.03
27	20.04	20.009	-0.031
28	8.5	8.53	0.03
29	5.8	5.769	-0.031
30	14.1	14.131	0.031
31	12.7	12.73	0.03
32	11.8	11.77	-0.03
33	6.5	6.53	0.03
34	15.1	15.07	-0.03
35	12.2	12.23	0.03
36	21.9	21.87	-0.03
37	15.37	15.339	-0.031
38	10.07	10.04	-0.03
39	5.24	5.269	0.029
40	1.09	4.16	3.07

The error between the predicted and the actual SSA in the majority of samples is barely noticeable and the difference is very small resulting in a good prediction by this model.



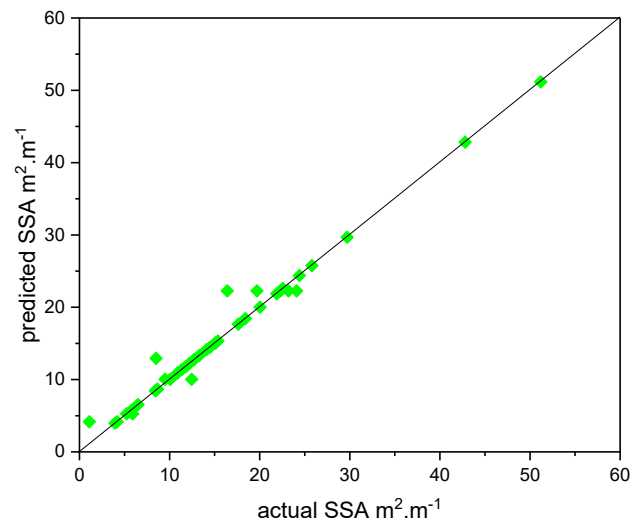


Figure.II.16. a Scatter plot representing Actual SSA vs. predicted SSA for perovskite samples by SVR second model using Origin software

Now we evaluate both models by testing them on 5 fold-cross validation (5 folds-CV) and leave on out cross validation (LOOCV).

The tests were carried out using the same computational software Weka through the classifier section and selecting the cross validation test.

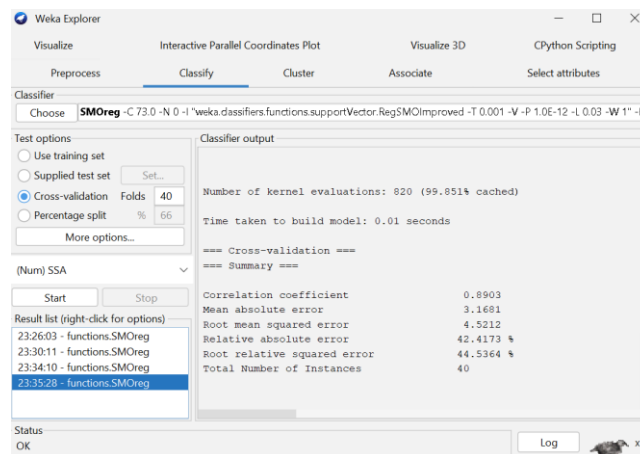


Figure.II.17.5-folds CV and LOOCV tests on Weka

To summarize the results:

Table.II.10.A simple table representing all the results of the evaluation of the first and the second model.

The test	The first model			The second model		
	R	RMSE	MAE	R	RMSE	MAE
Training data	0.9451	3.2839	2.1146	0.99	0.5812	1.4236
Testing data	0.9571	6.4127	5.4547	0.9449	6.5545	5.7012
5 fold-Cross validation	0.9025	4.2744	3.0148	0.8807	5.0033	2.9526
LOOCV	0.8903	4.5212	3.1681	0.90	4.7415	2.7954

We can see that the first model scores good for the correlation coefficient on the training and testing data with not so big difference plus a good evaluation by the cross validation that give us 0.9025 but a less score when it comes to the cross validation. The RMSE factor on the training data is small which means that our model is fitting pretty well with the data.

In the other hand the second model gave us an excellent results scoring 0.99 on training set and 0.9449 on testing set we can see there is a noticeable difference in RMSE and MAE factors between the training and testing set same with the first model.

The model did well too in the 5-fold cross validation test and LOOCV tests but with a slit difference the first model perform well in the 5 fold-CV but a less result on the LOOCV which is odd (the LOOCV test tend to give more accurate results then K-fold CV) however the second model did good on the 5 fold-CV and better on the LOOCV test.

Both models have ups and downs regarding the evaluation the first model built with the first data with 3 set of features being B-Tm, CT, AH, with the target property SSA. The second model with 4 set of features being B-Tm, B-Hfus, CT, AH, also with the target property SSA.

We will be selecting the second model for the reason that it performed very well on the training set and gave an acceptable score regarding the LOOCV test.

## II.4. Model application:

Once we selected the desired model we will be putting it to prediction by creating a new data (removing some of SSA values mainly from the testing set which the model practically never seen) and run it through Weka and see if the model will be capable of predicting the removed data.

We have removed 5 SSA values being:

Sample 20 representing  $\text{LaCrO}_3$  with  $3.95\text{m}^2.\text{g}^{-1}$

Sample 26 representing  $\text{LaFe}_{0.9}\text{Co}_{0.1}\text{O}_3$  with  $51.2\text{m}^2.\text{g}^{-1}$

Sample 27 representing  $\text{LaFe}_{0.1}\text{Co}_{0.9}\text{O}_3$  with  $42.8\text{m}^2.\text{g}^{-1}$

Sample 32 representing  $\text{La}_{0.01}\text{Sr}_{0.995}\text{TiO}_3$  with  $24.1\text{m}^2.\text{g}^{-1}$

Sample 36 representing  $\text{La}_{0.5}\text{Bi}_{0.2}\text{Ba}_{0.2}\text{Mn}_{0.1}\text{FeO}_3$  22.55 with  $\text{m}^2.\text{g}^{-1}$

Viewer

Relation: WekaExcel-weka.filters.unsupervised.attribute.Remove-R1

No.	1: B-Tm Numeric	2: B_Hfus Numeric	3: CT Numeric	4: AH Numeric	5: SSA Numeric
16	5.316...	6.810064	5.10735	2.075...	17.63
17	4.518...	7.290652	5.10735	2.075...	24.41
18	3.719...	7.77124	5.10735	2.075...	13.32
19	2.92151	8.251828	5.10735	2.075...	8.65
20	8.57126	9.554997	5.10735	2.075...	
21	7.44131	9.294363	5.10735	2.075...	8.42
22	6.31136	9.03373	5.10735	2.075...	29.71
23	5.18141	8.773096	5.10735	2.075...	18.41
24	4.05146	8.512462	5.10735	2.075...	14.46
25	6.912...	5.848888	5.958...	2.075...	10.88
26	6.893...	5.908489	6.384...	4.150...	
27	6.738...	6.385292	6.384...	4.150...	
28	6.912...	5.848888	5.958...	1.245...	8.5
29	7.506...	6.991231	5.532...	4.150...	16.4
30	7.506...	6.991231	5.532...	4.150...	19.7
31	7.506...	6.991231	5.532...	4.150...	22.3
32	7.506...	6.991231	5.532...	4.150...	
33	7.506...	6.991231	5.532...	4.150...	23.2
34	6.912...	5.848888	5.958...	1.660...	9.5
35	6.912...	5.848888	4.256...	0.830...	25.8
36	6.912...	5.848888	5.10735	0.830...	
37	6.912...	5.848888	5.958...	0.830...	20.04
38	6.912...	5.848888	6.8098	0.830...	8.5
39	6.912...	5.848888	7.661...	0.830...	5.8
40	6.539...	6.86588	5.10735	0.830...	14.1
41	6.539...	6.86588	5.958...	0.830...	12.7
42	6.539...	6.86588	6.8098	0.830...	11.8
43	6.539...	6.86588	7.661...	0.830...	6.5
44	6.539...	6.86588	5.10735	1.660...	15.1
45	6.539...	6.86588	5.10735	2.490...	12.2

Figure.II.18. the removed values of SSA in Weka

- To put the model on prediction on Weka we need to open our normalized data again in Weka and go to the classifier section and then load our model into the result list with the same hyperparameters.
- Then selecting the model and clicking on Re-evaluating model on current test set not forgetting to go to more options and uncheck all the information that we don't need plus we need to choose our Output predictions to be as a Plaintext:

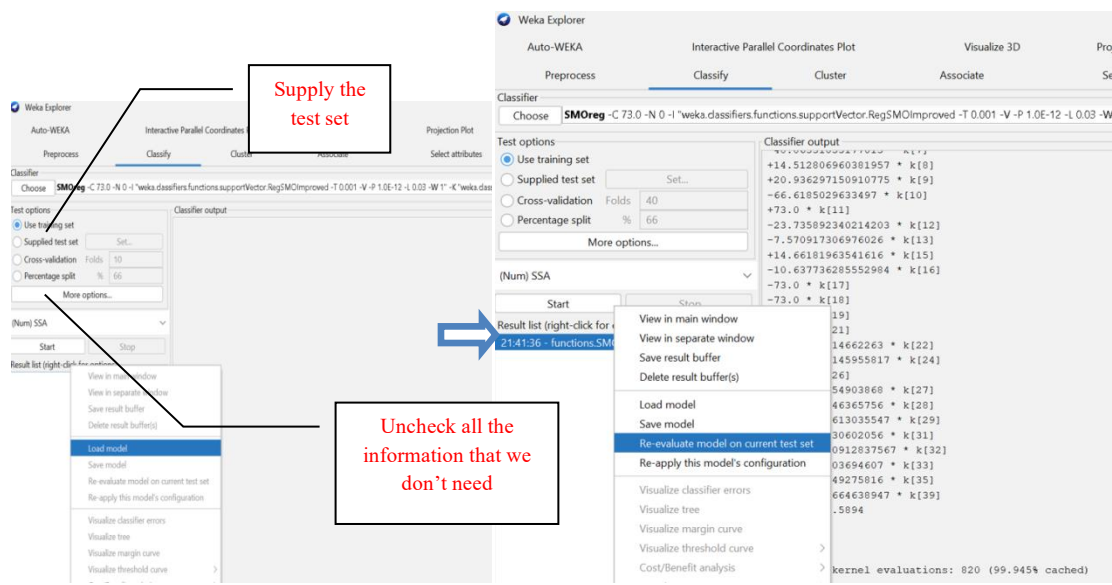


Figure.II.19. Preparing the model for predictions

The prediction result:

Table.II.11.The results of the prediction of the second SVR model

The sample	Actual SSA (m2.g-1)	Predicted SSA (m2.g-1)
LaCrO <sub>3</sub>	3.95	3.98
LaFe <sub>0.9</sub> Co <sub>0.1</sub> O <sub>3</sub>	51.2	51.17
LaFe <sub>0.1</sub> Co <sub>0.9</sub> O <sub>3</sub>	42.8	42.83
La <sub>0.01</sub> Sr <sub>0.995</sub> TiO <sub>3</sub>	24.1	22.27
La <sub>0.5</sub> Bi <sub>0.2</sub> Ba <sub>0.2</sub> Mn <sub>0.1</sub> FeO <sub>3</sub>	22.55	22.58

We can see that the predicted values are nearly the same as the actual values which tell us how good and well fitted our SVR model.

## Bibliography

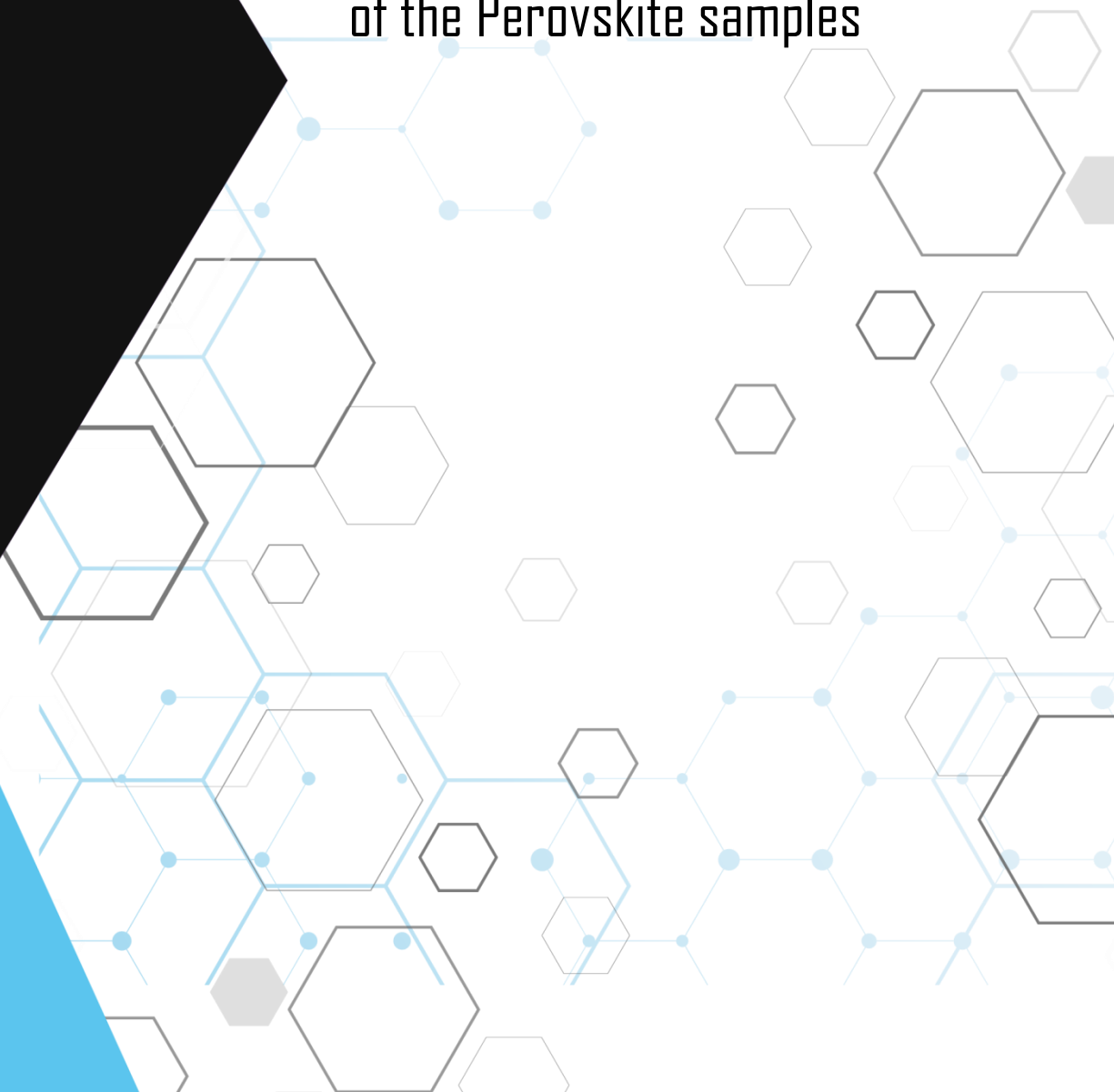
### References:

- 1 Chen, H., Yu, H., Peng, F., Yang, G., Wang, H., Yang, J., & Tang, Y. (2010). Autothermal reforming of ethanol for hydrogen production over perovskite LaNiO<sub>3</sub>. *Chemical engineering journal*, 160(1), 333-339.
- 2 Liu, H. X., Zhang, R. S., Yao, X. J., Liu, M. C., Hu, Z. D., & Fan, B. T. (2004). Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. *Journal of chemical information and computer sciences*, 44(1), 161-167.
- 3 Rossel, R. V., & Behrens, T. (2010). Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma*, 158(1-2), 46-54.
- 4 Shi, L., Chang, D., Ji, X., & Lu, W. (2018). Using data mining to search for perovskite materials with higher specific surface area. *Journal of chemical information and modeling*, 58(12), 2420-2427.
- 5 Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.
- 6 Ali, P. J. M., Faraj, R. H., Koya, E., Ali, P. J. M., & Faraj, R. H. (2014). Data normalization and standardization: a technical report. *Mach Learn Tech Rep*, 1(1), 1-6.
- 7 Chen, N. (2004). *Support vector machine in chemistry*. World Scientific.
- 8 Belete, D. M., & Huchaiah, M. D. (2021). Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *International Journal of Computers and Applications*, 1-12.

- 9 Bahrami, H., Homayouni, S., Safari, A., Mirzaei, S., Mahdianpari, M., & Reisi-Gahrouei, O. (2021). Deep learning-based estimation of crop biophysical parameters using multi-source and multi-temporal remote sensing observations. *Agronomy*, 11(7), 1363.

# Chapter 04

Synthesis and  
characterization of some  
of the Perovskite samples



## I. Introduction

Perovskite oxides are usually synthesized by a variety of methods including reaction to the solid state, co-precipitation, and sol-gel method as I mention in chapter 02 and 01 with their importance and many properties and application. In this part of work we decide to synthesize via sol-gel method some of the samples from the collected data ( $\text{LaFeO}_3$ ,  $\text{LaMgO}_3$ , and  $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ ) and two samples ( $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$  and  $\text{LaFe}_{0.7}\text{Mg}_{0.3}\text{O}_3$ ) screened out using Li Shi, Dongping Chang, Xiaobo Ji, and Wencong Lu. Journal model and characterize them using X-ray diffraction (XRD), Thermogravimetric analysis (TG), Fourier transforms infrared spectroscopy analysis (FTIR), and Brunauer, Emmett and Teller method for Specific area measurement (BET). For The  $\text{LaFe}_{0.7}\text{Mg}_{0.3}\text{O}_3$  sample unfortunately we just did the FTIR analysis.



## II. $\text{LaFeO}_3$ , $\text{LaMgO}_3$ , $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ , $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$ , and $\text{LaFe}_{0.7}\text{Mg}_{0.3}\text{O}_3$ preparation by sol-gel method

Each of the perovskite samples was prepared via the sol-gel method in same conditions (calcinations temperature, calcinations time) for the samples from the collected data.

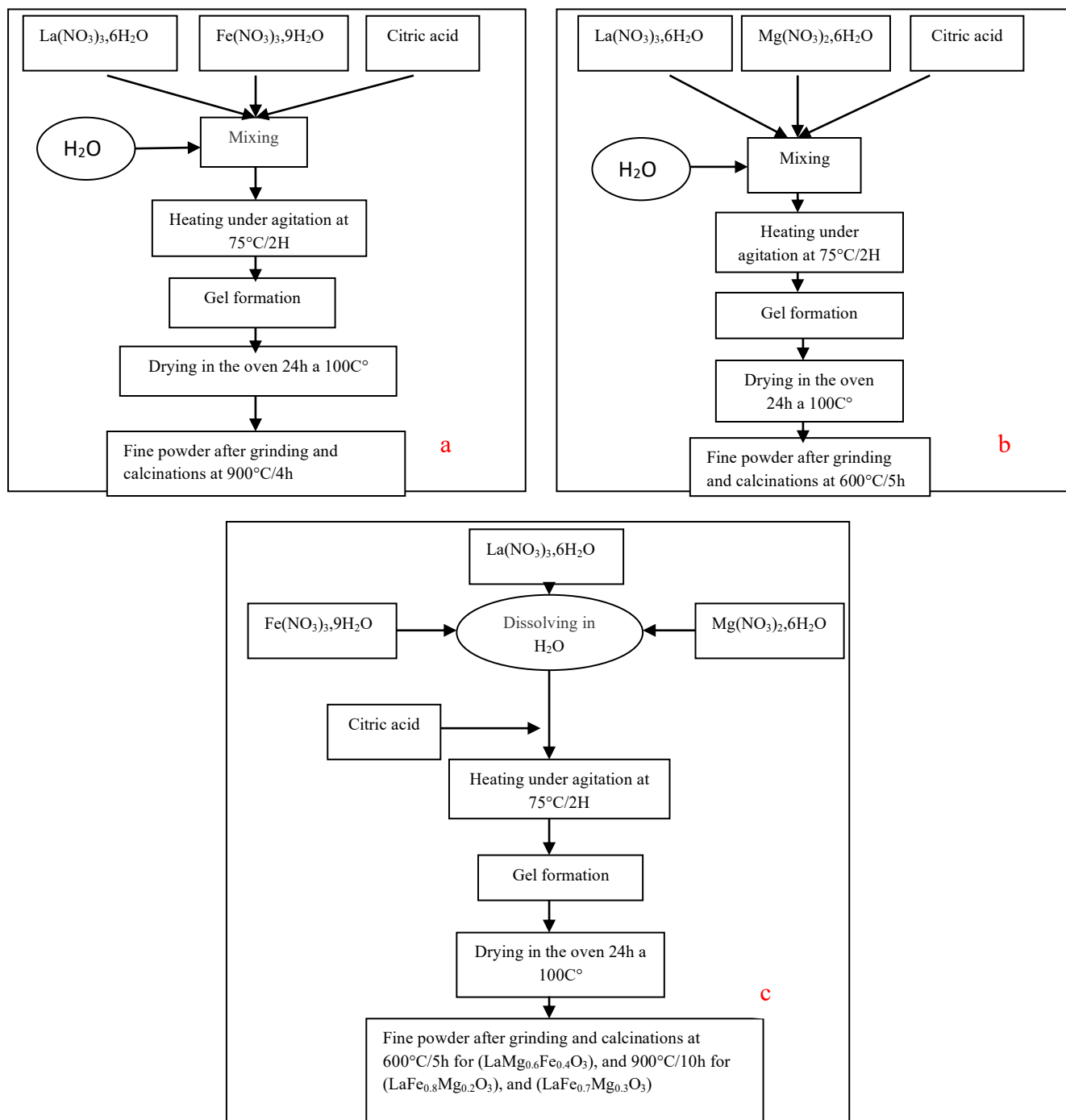


Figure.II.1. Synthesis Flowchart of (a).  $\text{LaFeO}_3$ , (b).  $\text{LaMgO}_3$ , (c).  $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ ,  $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$ , and  $\text{LaFe}_{0.7}\text{Mg}_{0.3}\text{O}_3$

### III. Characterization of the prepared samples

#### III.1. X-ray Powder diffraction (XRD)

The x-ray powder diffraction analysis was carried out for the four calcinated samples using diffractometer Brucker-D8 Advance the XRD spectrums obtained via XRD software match!

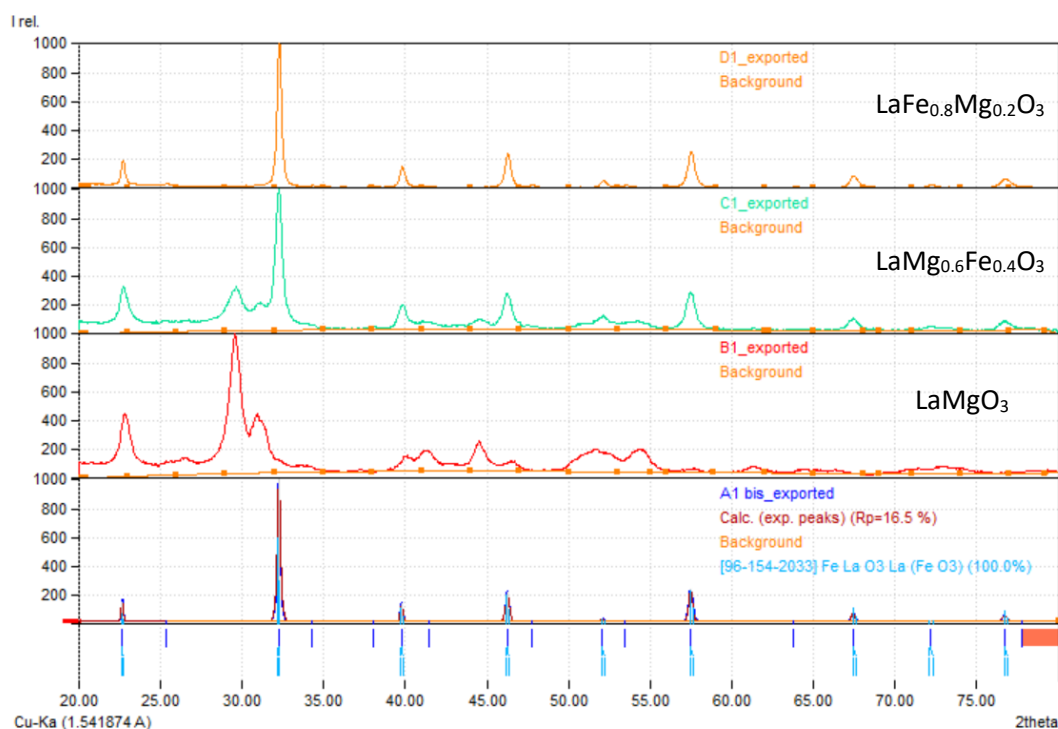


Figure.III.1. XRD pattern of the four samples

It's clearly observed in diffractograms of synthesized samples that there is just one pure phase that appeared successfully, belongs to  $\text{LaFeO}_3$  for the undoped sample comparing with  $\text{LaFeO}_3$  phase of the ICDD data file n: 98-154-2033 while we noticed that desired phases did not formed for doped materials.

III.2. TGA analyse for  $\text{LaFeO}_3$ ,  $\text{LaMgO}_3$ ,  $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ , and  $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$ 

In order to estimate beforehand the calcinations temperature to obtain a well crystallised oxide the five samples were analyzed by thermogravimetry analyse (T.G.A) on a type device TGA 1600 °C, Mettler Toledo.

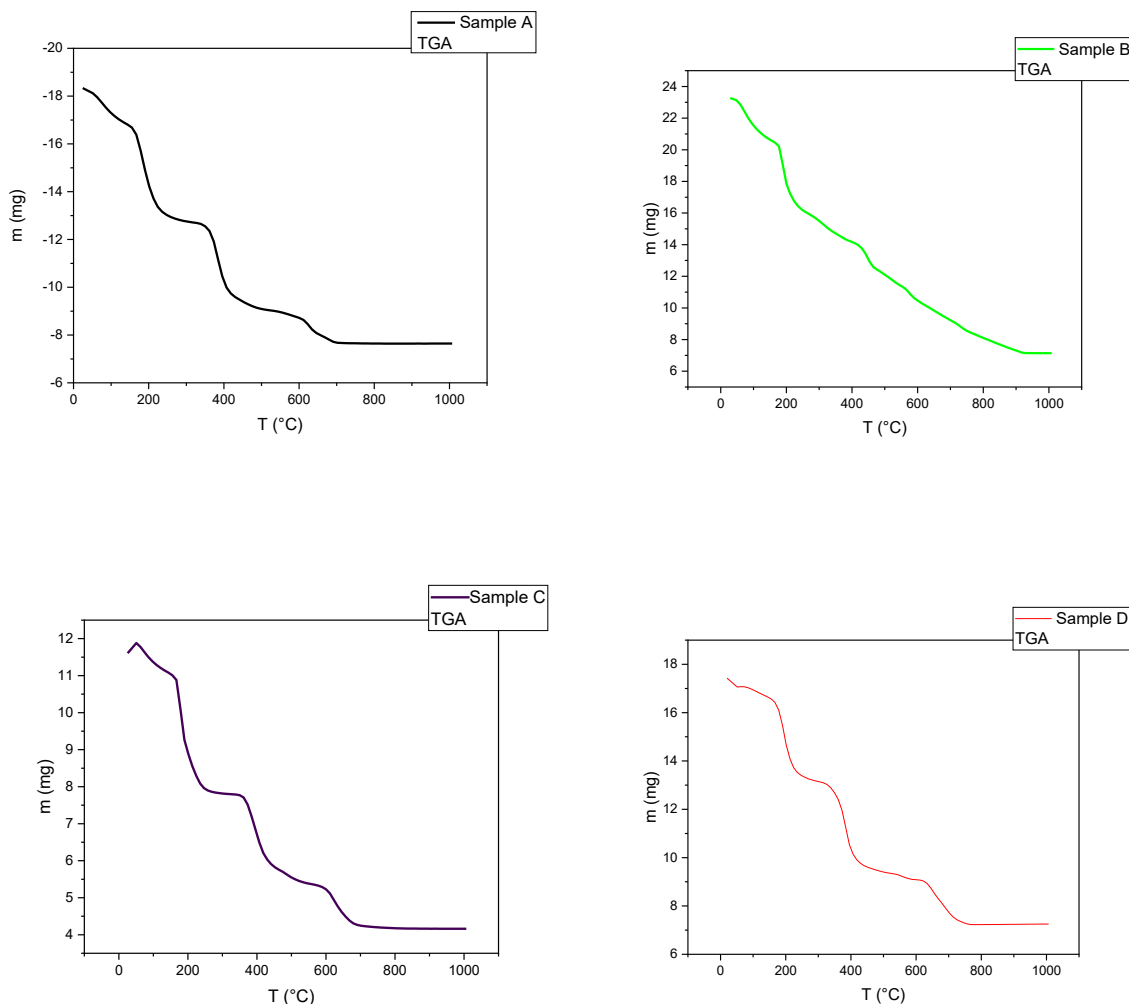


Figure.III.2.TGA analysis curve for A ( $\text{LaFeO}_3$ ), B ( $\text{LaMgO}_3$ ), C ( $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ ), D ( $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$ )

TGA analysis shows sample A ( $\text{LaFeO}_3$ ) powder behaviour with a starting mass 18,3330 mg heated up to 1000°C at a speed of 5°C per second. Sample B ( $\text{LaMgO}_3$ ) starting mass is 23,2660 mg heated up to 1000°C at a speed of 5°C per second. Sample C ( $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ ) starting mass is 11,6050 mg heated up to 1000°C at a speed of 5°C per second. Sample D ( $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$ ) starting mass is 17,4300 mg heated up to 1000°C at a speed of 5°C per second.

### III.3. FTIR analyse for $\text{LaFeO}_3$ , $\text{LaMgO}_3$ , $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ , $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$ , and $\text{LaFe}_{0.7}\text{Mg}_{0.3}\text{O}_3$

The FTIR analysis was done through FTIR-ATR Infrared Spectrometer: Bruker. The wavelengths studied are between  $4000 - 400 \text{ cm}^{-1}$ , for medium infrared. The calcinations temperatures for the four different calcinated samples are (900,600,600,900,900) respectively. The vibrational mode of the metal-oxygen (M-O) bond at around  $400-600 \text{ cm}^{-1}$  indicates the formation of a typical perovskite ( $\text{ABO}_3$ ) structure.

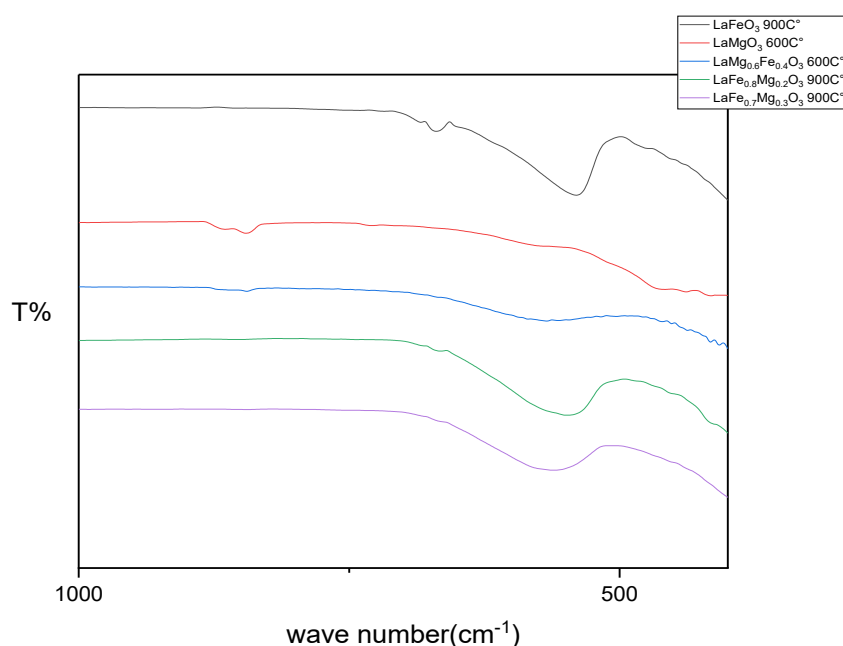


Figure.III.3. FTIR analyse for the five samples

The IR spectrum of  $\text{LaFeO}_3$  shows absorption band at  $540 \text{ cm}^{-1}$  attributed to Fe–O stretching vibration. For the second sample  $\text{LaMgO}_3$  an absorption band located at  $565 \text{ cm}^{-1}$  attributed to Mg–O stretching vibration. The third sample  $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$  the IR spectrum shows that there is an absorption band found at  $567 \text{ cm}^{-1}$ . the fourth calcinated sample  $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$  an absorption band located at  $548 \text{ cm}^{-1}$  attributed to Fe–O stretching vibration. The last sample  $\text{LaFe}_{0.7}\text{Mg}_{0.3}\text{O}_3$  Fe–O absorption bond can be found at  $560 \text{ cm}^{-1}$  attributed to Fe–O stretching vibration. There are a slight shift in the last three spectrums and the first spectrum concerning  $\text{LaFeO}_3$ ,  $\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$ ,  $\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$ , and  $\text{LaFe}_{0.7}\text{Mg}_{0.3}\text{O}_3$  with  $x=0$ ,  $x=0.6$ ,  $x=0.3$ ,  $x=0.2$  respectively.

### III.4. Specific area measurement by the BET method


The values of the specific surface area (SSA  $\text{m}^2/\text{g}$ ) for the 4 perovskite samples were calculated using BET Surface Area Analyzer quantachrome version 5.21. In order to obtain accurate measurements of the surface, the temperature and pressure of an inert gas are adjusted so that a single layer of gas molecules is adsorbed on the entire surface of the solid sample.

Table.III.1 shows the calculated values of the SSA for the synthesized perovskite samples besides the values found in the collected data set.

Table.III.1. Specific surface area of the synthesized samples and the collected samples

The samples	BET surface area ( $\text{m}^2 \text{g}^{-1}$ ) for the synthesized samples	BET surface area ( $\text{m}^2 \text{g}^{-1}$ ) for the collected samples
$\text{LaFeO}_3$	0.687	1.08
$\text{LaMgO}_3$	19.705 $\text{m}^2/\text{g}$	8.65
$\text{LaMg}_{0.6}\text{Fe}_{0.4}\text{O}_3$	122.269 $\text{m}^2/\text{g}$	24.41
$\text{LaFe}_{0.8}\text{Mg}_{0.2}\text{O}_3$	14.304 $\text{m}^2/\text{g}$	57.70

There is a big difference between in the measured surface area and the one from the collected data this probably due to errors in the experiment and some of the phases did not appear while performing the XRD analysis.

The background features a large, abstract geometric design on the left side, consisting of a black triangle pointing right and a blue triangle pointing left, meeting at a diagonal line. The rest of the page is white with a pattern of light blue and grey hexagons and lines, some containing small blue dots, resembling a molecular or network structure.

# General Conclusion

### Conclusion:

In this work implementing machine learning and data mining methods to find perovskite materials with higher specific surface area by building a predictive model capable of predicting specific surface area of  $ABO_3$  perovskite-type materials using Support vector regression algorithm SVR gave the following conclusions:

In chapter 03 by using the data mining software Weka two predictive models were built with two different features set from the same data set using Support vector regression algorithm to predict the specific surface area of 50 samples of perovskite-type materials.

For the first model the feature set consist of the melting point of the B-position (B-Tm), the calcinations temperature (CT), and the calcinations time (AH). The feature selection was done using wrappers methods implemented by Weka. The performance of the model was acceptable the correlation coefficient between the predicted and the actual specific surface area was 0.94 for the training data and 0.89 for the leave-one-out cross validation test.

The second model the feature set consist of the Enthalpy of fusion at the melting point of the B position (B-Hfus), the melting point of the B-position (B-Tm), the calcinations temperature (CT), and the calcinations time (AH). The feature selection was done by dividing the data set into 10000 different subsets. The performance of this model was very good scoring 0.99 for the training set and 0.90 for the leave-one-out cross validation test.

By selecting this model and putting it for application by removing some of the SSA values and adding one new sample. The prediction results were very good and the error between the predicted and the actual SSA value were very small.

In Chapter 4 some of the perovskite samples been synthesis via the sol-gel method and characterize with XRD analysis, TGA, FTIR analysis, and Specific area measurement by the BET method. through diffractograms of synthesized samples there is just one pure phase that appeared successfully, belongs to  $LaFeO_3$ . For the Specific area measurement the results shows a difference between the measured surface area and the one from the collected data this probably due to errors in the experiment and some of the phases did not appear while performing the XRD analysis.

In the end these simple results demonstrated how data mining and machine learning methods can improve how scientists develop, discover, and design materials.

## ملخص

مساحة السطح المحددة هي خاصية مهمة للغاية مرتبطة بالقدرة التحفيزية الضوئية للبيروفسكايت- $ABO_3$  في هذا العمل طبقنا بعض التعلم الآلي (ML) وطرق تعدين البيانات (DM) للبحث والعثور على البيروفسكايت- $ABO_3$  بمساحة سطح محددة أعلى (SSA) تتراوح من 1 إلى  $60 \text{ g}^2.\text{m}^{-1}$  من قاعدة بيانات منشأة مسبقاً مليئة بـ 50 عينة و 24 ميزة (التركيبات الكيميائية والمعايير التقنية) عن طريق بناء نموذج تنبؤي باستخدام خوارزمية تراجع ناقلات الدعم (SVR) كل هذا بمساعدة برنامج تعدين البيانات المسمى Weka. يبلغ معامل الارتباط بين القيمة المتوقعة والقيمة الفعلية لـ SSA بـ 0.99 لمجموعة بيانات التدريب و 0.90 للمصادقة المتبادلة (LOOCV).

الكلمات المفتاحية: استخراج البيانات ، التعلم الآلي، البيروفسكايت، مساحة سطح محددة.

## Abstract

The specific surface area is a very important property associated with photocatalytic ability of  $ABO_3$ -type perovskite. In this work we applied some of the machine learning (ML) and data mining (DM) methods to search and find  $ABO_3$ -type perovskite with higher specific surface area (SSA) ranging from 1 to  $60 \text{ g}^2.\text{m}^{-1}$  from a pre-established database filled with 50 samples and 24 features (chemical compositions and technical parameters) by building a predictive model using Support vector regression algorithm (SVR) all of this with a help of a data mining software called Weka. The correlation coefficient between the predicted and the actual value of SSA is 0.99 for the training data set and 0.90 for leave-one-out cross-validation (LOOCV).

**Keywords: data mining, machine learning, perovskite, specific surface area.**

## Résumé

La surface spécifique est une propriété très importante associée à la capacité photocatalytique de perovskite de type  $ABO_3$ . Dans ce travail, nous avons appliqué certaines des méthodes d'apprentissage automatique (ML) et d'exploration de données (DM) pour rechercher et trouver la perovskite de type  $ABO_3$  avec une surface spécifique (SSA) plus élevée allant de 1 à  $60 \text{ g}^2.\text{m}^{-1}$  à partir d'une base de données pré-établie remplie de 50 échantillons et 24 caractéristiques (compositions chimiques et paramètres techniques) en construisant un modèle prédictif utilisant l'algorithme de régression vectoriel de soutien (SVR) tout cela à l'aide d'un logiciel d'exploration de données appelé Weka. Le coefficient de corrélation entre la valeur prédite et la valeur réelle de l'SSA est de 0,99 pour l'ensemble de données de formation et de 0,90 pour la validation croisée sans autorisation (LOOCV).

**Mots-clés : data mining, machine learning, perovskite, surface spécifique.**



