MACHINE LEARNING ANALYSIS OF PHOTOCATALYTIC CO_2 REDUCTION ON PEROVSKITE MATERIALS

by

İrem Gülçin Zırhlıoğlu

B.S., Chemical Engineering, Boğaziçi University, 2021

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering

Boğaziçi University

2023

ACKNOWLEDGEMENTS

First, I would like to express my great appreciation to my supervisor Prof. Ramazan Yıldırım for his valuable guidance, support, help and endless patience during the planning and development of my thesis. He encouraged and supported me for all my opinions and never lost his faith.

I would like to thank my thesis committee members, Prof. Ahmet Erhan Aksoylu and Assoc. Prof. Mehmet Erdem Günay for their valuable time and contributions.

Very special thanks to PhD student of my laboratory, Burcu Oral for her help and endless patience. She answered all my questions at any time without hesitation. Besides, I would like to thank the other SOLCAT-Machine Learning and ChE family for their support and patience.

I would like to thank my housemate Iraz Bölükbaşı for sharing all my troubles, picking me up every time I fell and making me laugh. I would like to thank dear Ömer for his endless support and belief.

I would like to thank all my dear friends whose names I could not mention one by one for their love and support.

Finally, I would like to thank my dear family, my mother Çiğdem, my father Gürol, and my sister Kübra for their endless support and unconditional love. Without them, none of these would have been possible.

ABSTRACT

MACHINE LEARNING ANALYSIS OF PHOTOCATALYTIC CO_2 REDUCTION ON PEROVSKITE MATERIALS

The purpose of this study is to construct a database from the experimental studies about CO_2 reduction on perovskite materials from published articles, then to extract information from this dataset to predict CO_2 production yields and the bandgap of the perovskites by using machine learning methods such as decision tree (DT), random forest (RF), gradient boosting (XGBoost), association rule mining (ARM), and linear regression (LR). By using Web of Science, relevant articles were examined, and 61 articles were selected for data extraction; 309 samples with 29 features (14 numerical and 15 categorical) were collected; these features included properties of perovskites such as bandgap, elemental information, and conditions of the experiments such as reaction temperature, phase of reaction collected as the features. Before the machine learning applications, pre-processing steps were applied to the dataset for cleaning and organizing. For the missing bandgap values, linear regression was applied for prediction from the available data. The biased and the highly absent features were eliminated while the missing values of others were filled with the mod or mean of the dataset. The ML methods were applied using two separate databases which were for gas and liquid phase reactions. 133 out of 309 samples with 30 features were used for gas phase dataset while the remaining 176 samples with 29 features were for liquid phase. 17 missing band gap values were predicted using linear regression with the R-square and RMSE were found as 0.75 and 0.36 respectively for validation set. With DT, the accuracy for test set was obtained 0.76 for gas phase and 0.84 for liquid phase. In the RF predictions, R-square and RMSE were found to be 0.64 and 24.5, respectively for test set in gas phase while they were 0.49 and 221.0 in liquid phase. Bandgap was the most important feature for gas phase while the most important feature for the liquid phase was found to be the cocatalyst. Finally, in the XGBoost, R-square and RMSE for test set in gas phase were 0.65 and 14.75, respectively and for liquid phases, they were 0.79 and 145.6.

ÖZET

PEROVSKİT MATERYALLER ÜZERİNDE CO_2 FOTOKATALİTİK İNDİRGENMENİN YAPAY ÖĞRENME YÖNTEMLER İLE ANALİZİ

Bu çalışmanın amacı, yayınlanmış makalelerde bulunan deneysel çalışmalardan perovskit fotokatalizörler üzerinde gerçekleşen CO_2 indirgenmesi ile ilgili bir veri seti oluşturmak ve bu veri set yapay öğrenme yöntemleri kullanılarak bilgi çıkarmak, toplam üretim hızını ve perovskitlerin bant aralığı tahmin etmektir. Karar ağacı (DT), rassal orman (RF), gredyan arttırma (XGBoost), birliktelik kural çıkarımı (ARM) ve doğrusal regresyon (LR) gibi yapay öğrenme yöntemleri kullanılmıştır. Web of Science kullanılarak, bu konuyla ilgili tüm makaleler incelenmiş ve 61 makaleden, 29 tanımlayıcı özelliği olan 309 örnek toplanmıştır. Makalelerden, bant aralığı, elementel bilgiler gibi perovskit özellikleri ve reaksiyon sıcaklığı, reaksiyonun gerçekleştiği faz gibi reaksiyon koşulları tanımlayıcı özellik (girdi değişkeni) olarak toplanmıştır. Yapay öğrenme yöntemlerine geçmeden önce veri seti ön analiz adımından geçirilmiş, gereksiz veriler temizlenmiş, eksik bilgiler uygun yöntemlerle doldurulmuştur. Doğrusal regresyon modeli, eksik bant aralıklarını tahmin etmek için kullanılmıştır. Çok sayışa tanımlayıcı özelliği eksin olan veriler çıkartılmış, diğerleri ise ortalama ve mod kullanılarak doldurulmuştur Yapay öğrenme yöntemleri, gaz ve sıvı faz verilerini kapsamak üzere iki farklı veri seti için uygulanmıştır. 309 örnekten 30 tanımlayıcı özellik içeren 133 tanesi gaz fazı için kullanılırken geri kalan 176 veri noktası 29 tanımlayıcı özellikle birlikte sıvı veri setinde yer almıştır. Doğrusal regresyon kullanılarak veri setindeki eksik 17 bant aralığı diğer bant aralığı verilerinden yararlanılarak tahmin edilmiş, modelin kök ortalama hatanın karesi (RMSE) değerleri gaz ve sıvı fazların doğrulama setleri için sırasıyla 0.75 ve 0.36 bulunmuştur. DT ile gaz fazı için test setinde doğruluk oranı 0.75, sıvı fazında ise 0.84 olarak elde edilmiştir. RF ile gaz fazı için test setinde R^2 değeri 0.64 ve RMSE değeri 24.5 , sıvı fazı için bu değerler 0.49 ve 221.0 olarak bulunmuş; en önemli özelliğin gaz faz için bant aralığı iken, sıvı faz için ko-katalizör olduğu saptanmıştır. XGBoost ile gaz fazı için test setinde R^2 ve RMSE değerleri sırasıyla 0.65 ve 14.75 olarak bulunmuştur ve bu değerler sıvı fazında 0.79 ve 145.6'dır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

x	Input
$X,$	Predicted value
$Y,$	Actual values
$\tilde{y}$	Predicted output
$\tilde{Y}$	Mean of the true values
α	Intercept coefficient
β	Regression coefficient

LIST OF ACRONYMS/ABBREVIATIONS

APC	Adopting Polymer Complex
ARM	Association Rule Mining
CB	Conduction Band
DT	Decision Tree
ES	Electrospinning
GBMR	Gas Bubbling Assisted Membrane Reduction
HT	Hydrothermal
KNN	K-nearest Neighbor
LHS	Left Hand Side
LR	Linear Regression
ML	Machine Learning
MP	Materials Project
MPC	Modified Polymer Complex
MS	Molten Salt
MSE	Mean Squared Error
NN	Neural Network
PC	Polymer Complex Method
PTP	Pechini
RF	Random Forest
RHS	Right Hand Side
RMSE	Root Mean Squared Error
SG	Sol-gel
SS	Solid State
SVM	Support Vector Machine
UV	Ultraviolet
VB	Valence Band
XGBR	Gradient Boosting Regression Algorithm

1. INTRODUCTION

In recent years, the use of coal, oil, and fossil fuels has increased exponentially with dramatic rise in the population and economic growth; this enormous consumption rate increases carbon dioxide (CO_2) emissions, which contributes to the greenhouse effect, climate changes and becomes a threat for both humans and the environment. To avoid these problems, studies for CO_2 reduction have been continuously increased in recent years. In addition to more traditional ways of CO_2 capturing, the new trends such as photocatalytic reduction, electrocatalytic reduction, biological transformation, and solar-thermal catalytic reduction have also been extensively studied. These methods are used to convert CO_2 into C-based compounds which can be used as used as products. Among all applications, photocatalytic CO_2 reduction can be selected as the greener and more promising process (He et al., 2022).

In CO_2 reduction process, breaking the bond between C and O needs a great amount of energy because of its stability. Photocatalytic reduction is one of the challenging methods for converting CO_2; hoverer, it has clear benefits compared to its alternatives. First, mild conditions such as room temperature and pressure can be sufficient for photocatalytic reduction. Second, the unused CO_2 can be used as the starting carbon source of the process to obtain clean solar energy. By the photocatalytic reduction, short-chain hydrocarbons are produced such as CH_4, CH_3OH, C_2H_6 with solar energy to use them as carbon fuels. Finally, with all these, the carbon dioxide emission could be decreased with the use of alternative hydrocarbon fuels produced from CO_2 instead of fossil fuels (Tu et al., 2014).

During the photocatalytic CO_2 reduction process, multiple variables are considered for the reaction. The transformation step occurs on the surface named as photocatalyst (or semiconductor) which accelerates the reaction. This surface also absorbs the light and creates electrons and the holes to convert the CO_2. The types of semiconductors are used for this purpose vary from works to works while the perovskites are among the most preferred materials due to their easy to modify structures. Perovskites can be inorganic, hybrid organic-inorganic, double, layered, defect etc. relying on their features (Oku, 2020; Sun et

al., 2018; L. Wang et al., 2018). In the studies, machine learning application of photocatalytic CO_2 reduction over inorganic perovskites is studied.

As a research topic, photocatalytic CO_2 reduction on perovskites is an increasing trend. Wu and his group studied on the $SrTiO_3$, which is a common perovskite, by modifying it and promoting it with cocatalyst to improve the performance. They have been successful in enhancing the performance of an already good perovskite further by modifications (X. Wu et al., 2019a). Another titanate type perovskite $BaTiO_3$ was studied by Zengeneh and his group. They also tried to increase the performance of the perovskite by additives. With the additives, the properties and the performance of the perovskite were increased (Zangeneh et al., 2020). Shoa studied Photocatalytic CO_2 reduction on a tantalate, $KTaO_3$, which can be cubic or octahedral. With the experiments in UV light, the photocatalytic activity of octahedral crystal structure was found to be better. It was also observed that NiO as a co-catalyst increased the performance (Shao et al., 2018). The effect of the crystal structure was also studied by Li and his group; they examined the cubic and orthorhombic $NaNbO_3$ and obtained better activity with cubic $NaNbO_3$ (P. Li, Ouyang, Xi, et al., 2012). In the study of Kwak, the performance of a layered perovskite Sr_2TiO_4 was examined. The photocatalytic activity was tried to increase by the surface modifications, and they have been successful (Kwak et al., 2017).

With the increasing size of experimental and computational data, it becomes possible to analyze the results of previous studies by machine learning and extract valuable information for the future works. Remarkable progress in machine learning methods and computational tools also contributed to this trend. Indeed, various works involving the machine learning applications of photocatalysis have been published by various groups including ours. For example, Odabaşı and Yıldırım used ML tools such as decision tree classification and random forest regression to analyze the performance of perovskite solar cells from the experimental samples collected from 800 published articles (Odabaşı Özer & Yıldırım, 2019) while Can & Yıldırım published a ML work on water splitting over perovskite semiconductors (Can & Yildirim, 2019). Similarly, Oral et al. analyzed the data for photoelectrochemical water splitting (Oral et al., 2022) and Saadetnejad and Oral studied for photocatalytic CO_2 reduction (Saadetnejad et al., 2022).

In this study, a dataset was constructed by selecting 61 articles from all published articles until March 2023. 309 experiments were extracted from these articles with various descriptors. The dataset was separated into two subsets due to the reaction phase because of the distinct input variable, H_2O:CO_2, which is meaningful for gas phase only; consequently, 29 descriptors were selected for liquid phase dataset and the 30 descriptors were selected for gas phase dataset. The dataset went through pre-processing steps to make ready for ML methods while inputs and the outputs correlations were also analyzed. The decision tress, random forest, gradient boosting, and association rule mining were used to analyze the dataset and predict the total yield of the CO_2 reduction reaction. Linear regression was used to predict the missing bandgap values in the dataset. By the feature importance analysis of the DT and the RF, the effect of the inputs on the result was also analyzed.

This thesis consists of five chapters. The detailed information and the literature samples about the photocatalytic CO_2 reduction and the machine learning methods were given in Literature Survey (Chapter 2). The constructing database and the details about the modelling were given in Computational Details (Chapter 3). The results of the ML methods and the database analysis were discussed in the Result and Discussion (Chapter 4). The conclusion and the recommendation of this work were given in the Conclusion and Recommendation (Chapter 5).

2. LITERATURE SURVEY

2.1. Photocatalytic CO_2 Reduction

In recent years, with the increasing use of the fossil fuel and the industrial manufacturing, the amount of the CO_2 in the atmosphere has been increased significantly resulting various effects such as acid raining, global warming and glacier melting, that threaten the human health and environment (Sun et al., 2018; Zeng et al., 2018). To reduce these problems, various precautions such as reducing the use of fossil fuels, changing the energy resources to renewable sources, and converting CO_2 to less harmful chemicals (Kumar et al., 2020).

CO_2 reduction is used to convert CO_2 to hydrocarbons through processes such as biological conversion, electrochemical conversion, thermochemical conversion, and photocatalytic conversion. Despite the photocatalytic CO_2 reduction is harder than other methods, it can help the carbon cycle by converting the CO_2 into C-based hydrocarbon fuels which can be an alternative for fossil fuels and reducing air pollution (Tu et al., 2014).

CO_2 is a highly stable molecule; sufficient energy should be applied to break the bonds. Like photosynthesis, photocatalytic CO_2 uses the power of light; a semiconductor is also needed to absorb solar energy to provide sites for catalytic reactions. The CO_2 molecules converted to the C-based hydrocarbons such as CO, CH_3OH, CH_4, CH_2O etc. like in Figure 2.1 (Kumar et al., 2020; Tu et al., 2014). CO_2 is adsorbed to the surface while the light is absorbed generating electron and hole pairs. The electrons, with the increased energy, are transferred from valence band (VB) to conduction band (CB) while holes remain in valance band. Then the electron is used to reduce CO2 while the hole oxidize water generating products, which are desorbed form the surface (Ran et al., 2018; Sun et al., 2018). Various semiconductors like metal oxide, perovskites, metal sulfide, metal-free, and metal nitride are used for this purpose (L. Wang et al., 2018).

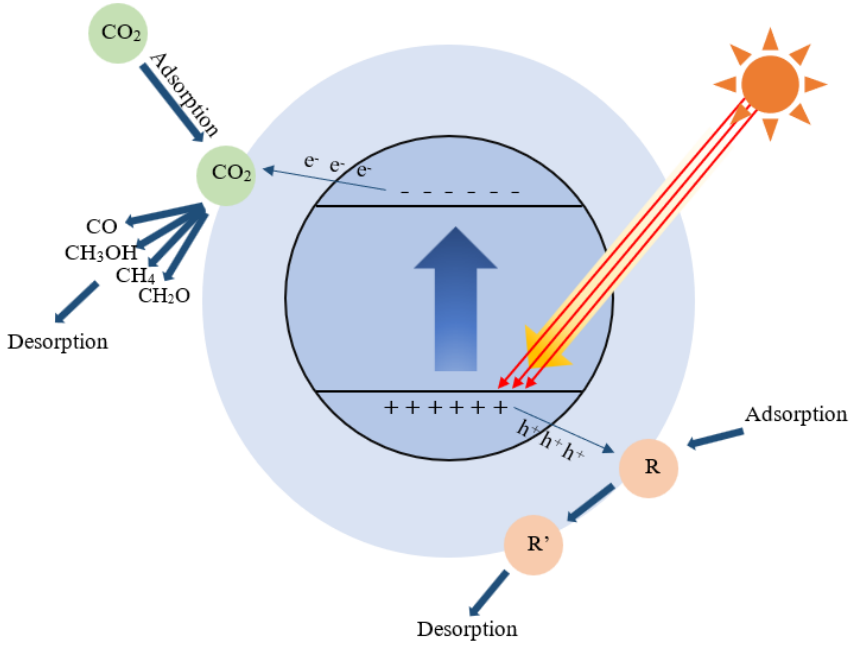

Figure 2.1. The scheme of photocatalytic CO_2 reduction. Adopted from (Sun et al., 2018; Xin et al., 2014)

The semiconductor choice is critical for CO_2 reduction efficiency. The bandgap which is the difference between CB and VB is one of the most important features of semiconductors. Bandgap is significant for the occurrence of the electrons and hole pairs. Besides, the reactions of the system take part on the semiconductors while transportation of the charges also occurs through band gap structure benefits to these reactions. Perovskites usually have suitable properties for CO_2 reduction exist in perovskites (C. Huang et al., 2016).

By the discovery of the $CaTiO_3$ mineral, the origin of the perovskite was found, and the general formula was generated as ABX_3 (Reshmi Varma, 2018). The cations are A and B, and the anions are X. A represents organic or inorganic metals and the X represents halogens, oxygen, or sulfur but generally oxygen. Due to the common features, perovskites are grouped as inorganic, hybrid organic-inorganic, double, layered, defect etc. (Can & Yildirim, 2019; Oku, 2020).

The inorganic perovskites are usually classified according to the B-site; Each element in B-site represents a group of perovskites like Ce; cerate , Fe; ferrite, Nb; niobate, Ta; tantalate, Ti; titanate, Zr; zirconate (Kumar et al., 2020).

The synthesis method of perovskite is critical to have the desired properties of the perovskite; It affects efficiency and other performance measures including the cost. Hydrothermal, solvothermal, sol-gel, polymer complex, solid-state, molten salt, sonochemical methods can be examples for perovskite synthesis. With aqueous solutions in high pressure and temperature, hydrothermal method is used (Xu et al., 2014) while the one used with other solvents (non-aqueous solutions) is called solvothermal methods, and utilized in perovskite synthesis (Caruntu et al., 2015). Sol-gel method involves the use of colloidal solutions as precursor and gelation of the solution (Parida et al., 2010). By polymer complex precursor is formed from various metal is used for synthesis in polymer complex method (Phokha et al., 2014). From the reaction of solid materials (solid-state method) in high temperature, perovskites are also obtained (Yin et al., 2014).

In the study of Nakanishi and his group, $NaTaO_3$ was used as the semiconductor to reduce CO_2 to CO, H_2 and O_2 by using H_2O as the electron donor. The experiments were done under UV light and with Ag cocatalyst in liquid phase. However, NiO, Ni, Cu, Ru, Rh, Pd and Au were also used as the cocatalyst. The perovskite was synthesized with solid-state method. Photodeposition, impregnation and liquid-phase reduction were used as the loading method of cocatalysts. With the cocatalyst, doping effects were analyzed with different combinations. Mg, Ca, Sr, Ba, and La were used as dopant. The selectivity was affected by every parameter such as dopant, cocatalyst, cocatalyst amount, loading method etc. in the experiment. The highest efficiency was obtained with Ag cocatalyst which was loaded as 3% by liquid-phase method and Ba dopant as 95% selectivity of CO (Nakanishi, Lizuka, et al., 2016).

$BaTiO_3$ was another perovskite which has high potential for CO_2 reduction. The hydrothermal method is used for the synthesis. The reaction occurred under UV light in gas phase. CO_2 was reduced to CH_3OH, CO, and CH_4. The effect of the dopant, Eu, was measured in different amounts. The dopant amounts affect both the bandgap and the

selectivity; the bandgap decreased, and the selectivity increased with the increasing amount of dopant (Hwang et al., 2023).

Chen and coworkers worked on barium zirconate as a new photocatalyst. The Pechini method was used to obtain $BaZrO_3$, and the reaction occurred in liquid phase at room temperature and pressure under UV light. The effects of cocatalyst and the calcination temperature were analyzed. 1000 °C, 1100 °C, and 1200 °C were used as the calcination time while Ru, Cu, Au, Pt, and Ag were tested as the cocatalyst. The highest efficiency was obtained over 0.3% of Ag photocatalyst calcined at 1000 °C (X. Chen et al., 2015a).

In another work, $BaCeO_3$ perovskite was used as the photocatalyst, which was synthesized by Pechini method. The experiments were done under UV light. The effects of the cocatalyst, Ag, Au, Pt, CuO, and RuO_2 were measured in liquid phase in the room conditions. The highest efficiency was obtained over Ag cocatalyst. Hence, different amounts of Ag (0.2%, 0.3%, and 0.5%) were also tested, and the perovskite with 0.3% cocatalyst had higher efficiency (J. Wang et al., 2015).

Fresho and coworkers studied $NaNbO_3$, $NaTaO_3$, and the intermediate form $NaNb_{0.5}Ta_{0.5}O_3$ in their article. The perovskites were synthesized with hydrothermal method. The experiments were done under UV light in gas phase. A dramatic difference in selectivity was observed between $NaNbO_3$ and $NaTaO_3$; $NaNbO_3$ performance was significantly higher. The combined perovskite also showed better performance than tantalate (Fresno et al., 2021).

In another study, $LaFeO_3$ was used as semiconductor because of its low price and efficiency. It synthesized by sol-gel method. The experiments were done under visible light in liquid phase. Different modifications, like doping, were performed on the perovskite. The efficiency was calculated for carbon templated porous $LaFeO_3$, amino-functionalized carbon templated porous $LaFeO_3$, N-doped $LaFeO_3$, and TiO_2 coupled N-doped $LaFeO_3$. The highest efficiency obtained from the last form of the semiconductor. Above semiconductors, the production yield increased while the bandgap decreased (Humayun et al., 2016).

As an example of layered perovskite, an article reporting CO_2 reduction over $Bi_4Ti_3O_{12}$ can be given. In this study, the perovskites were synthesized by molten salt and hydrothermal method. The experiments were done under visible light and room conditions in liquid phase. The perovskite which was obtained by molten-salt method showed better performance because it had higher photocarrier separation efficiency, larger surface area, better photocurrent response and lower resistance for charge transfer (Jia et al., 2022).

2.2. Machine Learning Methods

In recent years, the big data concept has evolved with the increase of the data generation in every area of life; from social disciplines to quantitative studies, large amounts of data already accumulated, and it continues to increase exponentially. Big data concept is relevant for many areas from biomolecular research, society, security, energy studies etc. (L'Heureux et al., 2017; Qiu et al., 2016). The knowledge extraction from big data is possible by latest computer science technologies as well; this is the main reason for the popularity of the concept. Machine learning is one of the computer science technologies, which has been improved in recent years, to extract useful information from big data and make important contributions to various domains (L. Zhou et al., 2017).

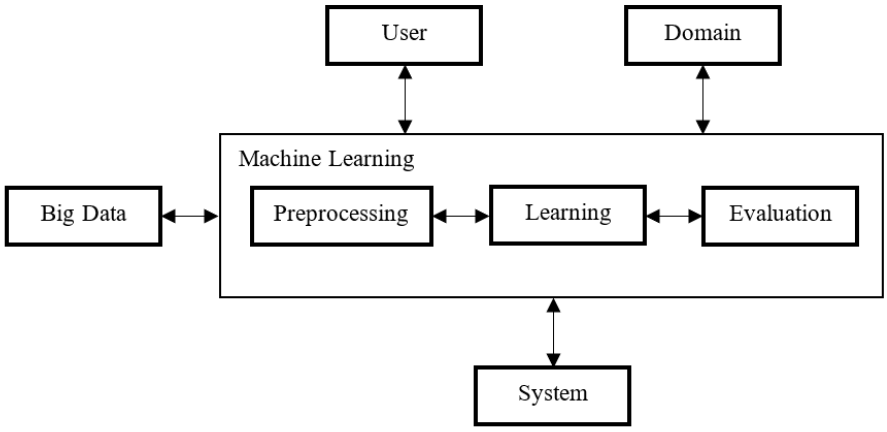

Figure 2.2. Machine learning frame on big data adopted from (L. Zhou et al., 2017)

In Figure 2.2., the relation of the big data, machine learning, user, domain, and system can be seen. Big data gets into machine learning to provide information for system, user, and

domain. The double arrows show that, by the feedback, the ML system is improved, and, in return, improves the data. Preprocessing is used to organize the big data before learning step; preprocessing tasks and methods may be different for different ML method. After learning from the prepared dataset, evaluation step is executed, and the results are obtained (L. Zhou et al., 2017).

ML replaces some parts of experiments in science gradually because in the past, there is huge amount of almost experimental data to extract information and trained in an ML model. With the knowledge from past experiments, new results can be obtained to guide the new experimental works better.

The ML techniques are quite diverse; new techniques have been developed continuously based on the needs. These techniques are divided into three main titles as supervised, unsupervised and reinforcement techniques. Recently, with the increasing in the data amount, the complexity of the models also rises. Hence, different main titles occurred as classical machine learning and deep learning. Classical machine learning techniques are always effective and useful for statistical analysis and information extraction, even deep learning is improved day by day. Linear regression, logistic regression, decision tree, support vector machine algorithm, naïve Bayes algorithm, KNN algorithm, random forest algorithm, gradient boosting algorithm, etc. are the examples for classical ML techniques (Bonaccorso, 2017).

As in Figure 2.2., before all ML techniques, preprocessing step should be applied to raw dataset. The dataset usually has missing values especially when it is collected from the experimental results. In such cases, the blanks should be filled or removed. If there is not enough information that results are removed from the dataset. In some cases, the unknown values of variables can be filled with the mean or the median of the other samples or another predictive model can be used to predict them. The dataset can be constructed by both categorical and numerical data. Depending on the machine learning techniques used, the categorical data stays, or changes to numeric data (1,0) by one-hot encoded algorithm. Normalization and standardization are another technique for preprocessing. They are applied to remove the effect of numerical magnitude, fasten the convergence, and remove the negativity of the outliers. However, it is not necessary for tree-based models. Another

method is featuring selection. The features (input variables) need to be chosen reasonably. Hence, after collecting all the samples, the unnecessary features are removed. Another crucial preprocessing step is shuffling. It is necessary to prevent learning from the sorted similar samples (Bonaccorso, 2017; J. Huang et al., 2015).

2.2.1. Linear Regression

Linear regression is a common and the simplest data analysis tool. It is used to find the relationship between dependent and independent variables. The dependent variable is output, and the independent variables are inputs.

$$\tilde{y} = \alpha + \beta x \tag{2.1}$$

Equation (2.1) is the general formula. $\tilde{y}$ is the predicted output variable and x is the input variable. α and β are the coefficients represents the intercept and regression coefficient respectively. The aim of the linear regression is predicting output values from the independent values. Hence, the difference between the predicted $\tilde{y}$ value and the observed y value need to be minimized (Bazdaric et al., 2021; Schneider et al., 2010).

In the basic statistical application, linear regression is used frequently. In their study, Pino-Mejias and his group compare linear regression with artificial neural network models on energy use in office buildings to predict heating and cooling demand, energy consumption and CO_2 emissions. Their dataset consists of 77000 cases which are evaluated over eight features. They obtained better results with linear regression to predict energy consumption and CO_2 emissions cases when the transformed predictive variables are used. On the other hand, without being transformed, multilayer perceptron shows much better accuracy for all topics (Pino-Mejías et al., 2016).

2.2.2. Decision Tree

Basically, a decision tree is an algorithm represented by a graph with tree form showing the choices to make and their possible outcome. In detail, it is a classification model (or can be regression model) that represents an iterative division of the analyzed instances. A decision tree composed of nodes and branches. There is root in each tree which is the first

node and has no previous node (Age in Figure 2.3.). The nodes which are used to decide to another node are called the internal nodes (leaves or terminal node) and they split into two or more branches due to the condition of the inputs. The conditions of the division can also be seen in Figure 2.3. (Mahesh, 2018; Rokach & Maimon, 2005).

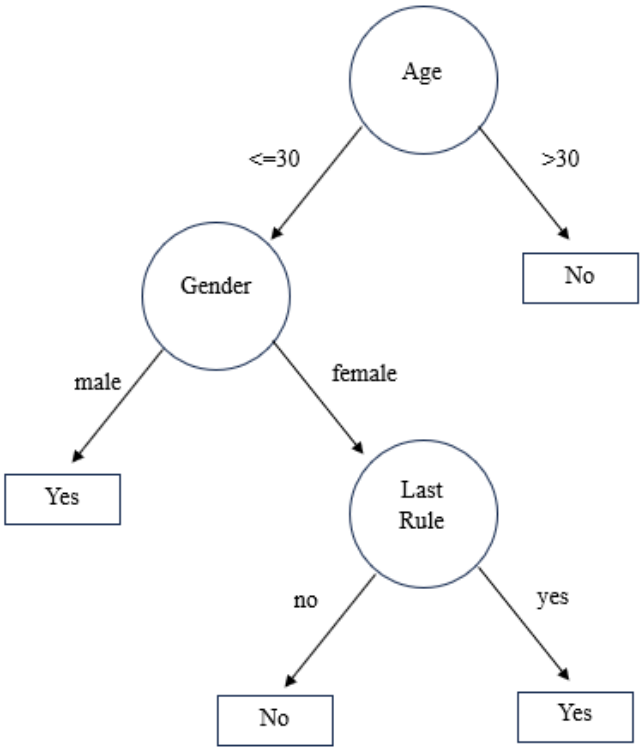

Figure 2.3. Decision tree sample adopted from (Rokach & Maimon, 2005)

Decision tree algorithms are Iterative Dichotomser (ID3), C4.5, C5.0 and Classification and Regression Trees (CART). In classification problems, ID3 and C4.5 algorithms which were found and developed by Ross Quinlan, respectively concentrate on classification. (Domor Mienye et al., 2019). In the study of Priyam and his friends, all the decision tree algorithms are compared over the educational database. With the small database, C4.5 shows a better performance than ID3 and CART. On the other hand, with the large dataset, serial algorithms cannot show the same performance and SPRINT and SLIQ decision tree algorithms are used. Due to such requirements, the decision tree algorithms are developed (Priyam et al., 2013).

During the hyperparameter tuning for decision tree algorithms, it is important to balance the error rate, complexity, and computational power. In a while the error rate decreases with the increase in model complexity; however, after a point the model starts to memorize, and the error rate starts to increase. At this point the hyperparameters are chosen by taking computational power into account. If the performance is not affected by a large amount, the complexity should be decreased. Tuning the complexity parameter and the number of nodes is an example for hyperparameter tuning (Mantovani et al., 2018).

In the study of Can et al. the prediction of water solubility in ionic liquids was done by machine learning algorithms such as decision tree and deep learning (multilayer fully connected NN). Their dataset, which was computational dataset produced by COSMO-RS consists of 16137 ionic liquids. The heuristic rules about the ionic liquids were extracted by the decision tree. By classifying the output into three groups as low, medium, and high solubility, the effect of the properties on the solubility can be obtained by the classificational decision tree (Can et al., 2021).

2.2.3. Random Forest

In 2001, random forest (RF) was proposed by L. Breiman as a good machine learning algorithm which could do both classification and regression. RF can readily adjust to the non-linear pattern in the data while it can handle the large datasets. Unlike logistic regression, or linear regression, RF works even when the number of observations is less than independent variables. RF is a tree-based algorithm where the dataset is divided into two by using specific criterion until the pre-defined condition is satisfied. Depending on the predefined condition, the decision tree can do regression or classification. The limitation of the decision tree is overfitting, which means the model over learns the train dataset and causes poor performance (i.e., low accuracy) in the test set. Constructing multiple individual trees by considering only a subset from the observations, like in random forest, is a good strategy to increase the accuracy of model; RF is one of such ensemble tree-based learning algorithms. RF works by averaging the predictions from multiple individual trees, each constructed using bootstrap samples, instead of the original sample (Bonaccorso, 2017; Schonlau & Zou, 2020).

For RF, the hyperparameters, which can be changed by the users, must be optimized; these can be *mtry, sample size, replacement, node size, number of trees* and *splitting rule.* The quantity of candidate variables chosen in each split is the *mtry*. *Sample size* represents the observations number for each tree while *replacement* is chosen as true if the draw observation is done with replacement. In the terminal node, the minimum number of observations is the *node size*. The *number of trees* is the quantity of the trees in the forest. The *splitting rule* is used for deciding how to split the data at each node during the construction of the tree for the most effective separation. Optimum model with random forest is obtained by tuning these hyperparameters (Probst et al., 2019).

In the study of Oral et a. three machine learning algorithms including random forest were used to analyze dataset with 10,560 samples from 584 experiments in 180 articles. The experiments were about photoelectrochemical water splitting. Before the machine learning application, pre-analysis was done with simple descriptive statistics. Association rule mining (ARM), decision tree (DT), and random forest (RF) were used as ML techniques to find out the patterns between the samples. The aim was to predict photocurrent density by using 33 features as the inputs with decision tree and to predict the band gap with RF and DT. With DT model for band gap the training and test accuracy were obtained as 78% and 72% respectively. These results were 61% and 54% for photocurrent density prediction. With RF, the root mean square error were obtained as 0.24 and 0.27 for validation and test set respectively for band gap prediction (Oral et al., 2022).

2.2.4. Gradient Boosting

Ensemble methods such as random forest are highly impactful for improving the performance of the models. It uses randomization techniques which handle the problems with different solutions to find out the optimum. With recent technology, gradient boosting algorithms such as eXtreme Gradient Bosting (XGBoost) (T. Chen & Guestrin, 2016), Light Gradient Boosting Machine (LightGBM) (Ke et al., 2017), and Categorical Boosting (CatBoost) (Prokhorenkova et al., 2017) were also developed. These algorithms enhance the training speed and generalization capacity (Bentéjac et al., 2021).

Boosting algorithms work by integrating the weak learners into a powerful learner through an iterative process. It is a regression algorithm that follows the boosting principles. When the iterative process is not appropriately regulated, the boosting algorithms can have drawbacks such as overfitting. In such cases, preprocessing of the dataset becomes crucial. Gradient boosting algorithms can handle numerical data in harmony. The users can tune the hyperparameters such as learning rate, maximum depth of the tree, the subsampling rate, number of features for best split and the minimum number of the samples for internal node splitting (Bentéjac et al., 2021).

In the study of Chen and his group, ML method was used to analyze the electroreduction of CO_2. They used extreme gradient boosting regression algorithm (XGBR) to enable effective exploration of electrocatalysts for CO_2 reduction. Their model predicted the Gibbs free energy change of CO adsorption efficiently and quickly. The dataset was obtained from the DFT computations with VASP. They studied K-nearest neighbor regression (KNR), RF, support vector machine (SVM), gradient boosting regression and XGBR with various amount of training and test data. The best ratio with the data was obtained by 80% of train and 20% of test set with 0.902 r-square and 0.1652 root mean squared error (A. Chen et al., 2020).

2.2.5. Association Rule Mining

Data mining is classified as descriptive and prescriptive. Descriptive mining involves summary or description of the features of the dataset while prescriptive mining involves analyzing current data to make predictions relying on historical data. Association rules, classifications and clustering are examples of data mining. Association rule mining (ARM), which is introduced by Agrawal in 1993 is a widely studied data mining technique; it focuses on extracting meaningful correlations, frequent patterns, associations within the sets of items in the databases (Zhao & Bhowmick, 2003).

In ARM method, the rules are described through basic measures such as support, confidence, and lift. During the association rule mining, two sides are decided as the rules. These are right-hand-side (RHS) and left-hand-side (LHS). Assume that X is LHS, and Y is the RHS. The aim is to find the association between LHS and RHS with support, confidence,

and lift. Support shows the ratio of the number of X in the Y to the whole dataset (# of the data have both X and Y/# of all data) . Confidence shows the ratio of the number X in Y to the number of X in the dataset (# of data have both X and Y/# of X in data). Finally, the lift shows the ratio of the confidence to fraction of Y in the whole dataset (confidence/fraction of Y to all data) (Zhao & Bhowmick, 2003).

2.2.6. Error Metrics

After training and testing the model, a certain measure is needed to evaluate the performance of the model; error metrics as R-squared, root mean square error (RMSE), or mean square error (MSE) are used for this purpose. If the X_i assumed as the predicted value of i^{th} value, the Y_i is the actual value and $\tilde{Y}$ is the mean of the true values, the formulas are shown as follows (Chicco et al., 2021):

$$\tilde{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i, \tag{2.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}\left(\tilde{Y} - Y_i\right)^2}, \tag{2.3}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2, \tag{2.4}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2}\,. \tag{2.5}$$

3. COMPUTATIONAL DETAILS

3.1. Constructing Database

In this study, the dataset was constructed from experimental articles from Web of Science, by using keywords, of "CO_2 reduction, perovskite" and "Photocatalytic CO_2 reduction, perovskite". All articles (until March 2023) were examined, and 61 articles were selected. 309 data were collected in 29 different aspects which contain both input variables and the output variable. 14 numerical and 15 categorical data were taken from the articles as in Table 3.1.

Table 3.1. The categorical and the numerical features in the dataset

Categorical Features	**Numerical Features**
Perovskite	x of A
A	x of A1/B1
A1/B1	x of B
B	x of X
X	Dope percentage (%)
Additional material	Cocatalyst percentage (%)
Synthesis method of perovskite	2^{nd} cocatalyst percentage (%)
Dope	BET surface area (m^2g^{-1})
Synthesis method of cocatalyst	Calcination time (h)
Cocatalyst	Calcination temperature (C°)
Synthesis method of 2^{nd} cocatalyst	Reaction temperature (C°)
2^{nd} cocatalyst	Reaction pressure (bar)
Light type	Bandgap (eV)
Phase of the reaction	H_2O:CO_2
Crystal structure	

The dataset contains 41 different perovskites. The metal oxide ABX_3 perovskite consists of three sites, A, B and X; they were categorized according to the B-site, and these categories were used as one of the features of the dataset (as the type of the perovskite). In Table 3.2., the perovskites are tabulated for each category.

Table 3.2. Perovskites and the category of the perovskite

Type of Perovskite	**Perovskite**
Ferrite	$BiFeO_3$, $LaFeO_3$
Niobate	$KNbO_3$, $NaNbO_3$
Tantalate	$KTaO_3$, $LiTaO_3$, $NaTaO_3$,
Titanate	$BaTiO_3$, $BiTiO_3$, $CaTiO_3$, $NiTiO_3$, $SrTiO_3$, $Ca_{0.75}Ti_{1.12}O_3$, $Ca_{1.28}Ti_{0.85}O_3$, $Ca_{1.5}Ti_{0.75}O_3$
Layered	Ba_2NbFeO_6, $BaLa_4Ti_3O_{12}$, Bi_2WO_6, $Bi_4Ti_3O_{12}$, $CaLa_4Ti_4O_{15}$, $CsCa_2Ta_3O_{10}$, $H_2SrTa_2O_7$, $HCa_2Ta_3O_{10}$, HNb_3O_8, KNb_3O_8, $La_2Ti_2O_7$, $PbBi_2Nb_2O_9$, Sr_2TiO_4, $Sr_3Ti_2O_7$, Sr_3Ti_2ON, Sr_3TiFeO_7, $Sr_3TiFeON$, Sr_3TiFeS_2ON, $Sr_3TiFeSON$, Sr_3TiS_2ON, $Sr_4Ti_3O_{10}$, $SrLa_4Ti_4O_{15}$, $SrTiFeSON$
Other	$BaCeO_3$, $BaZrO_3$, $NaNb_{0.5}Ta_{0.5}O_3$

In the dataset, 12 A-sites, seven B-sites, five A or B additional sites and one X sites exist. In Table 3.3., the elements of the sites are given.

Table 3.3. A, B, A1/B1 and X sites of the perovskites

Sites of perovskite	**Elements**
A	Ba, Bi, Ca, Cs, H, K, La, Li, Na, Ni, Pb, Sr
B	Ce, Fe, Nb, Ta, Ti, W, Zr
A1/B1	Bi, Ca, La, Nb, Sr
X	O

In addition to the perovskite type, there were also other categorical features as tabulated in Table 3.4.

Table 3.4. Variable range of the features

Features/Number of features	**Variable range**
Additional material (2)	Yes, No
Synthesis method of perovskite (21)	(APC) Adopting polymer complex, (COP) Chemically oxidative polymerization, (ES) Electrospinning, FAPO, Flux, GBMR, Generic, (HT) Hydrothermal, (MPC) Modified polymer complex, (MS) Molten salt, (PTP) Pechini, (PC) Polymer complex, (SG) Sol-gel, (SS) Solid-state, Solution combustion, Solvo-combustion, Solvothermal, Sonication, Sonochemistry, Spin-coating, (UT) Ultrasonic treatment
Dope (16)	Al, B, Fe, Ba, C, Ca, Cr, Cu, Eu, N, S, La, Mg, N, Sr, None
Synthesis method of cocatalyst (6)	Chemical reduction, Impregnation, Liquid-phase reaction, Photodeposition, Precipitation, None
Cocatalyst (13)	Ag, Au, Cu, Cu_2O, CuO, Ni, NiO, Pd, Pt, Rh, Ru, RuO_2, None
Synthesis method of 2^{nd} cocatalyst (3)	Impregnation, Photodeposition, None
2^{nd} cocatalyst (3)	Cu, Rh, None
Light type (2)	UV, Visible
Phase of the reaction (2)	Liquid, Gas
Crystal Structure (10)	Cubic, Cubic-orthorhombic, Hexagonal, Monoclinic, Octahedron, Orthorhombic, Plane, Rhombohedral, Tetragonal, Trigonal

The range of numerical features varies with phase of reaction. In Table 3.5, the minimum and the maximum values of the numerical features are given for gas and liquid phase. The dataset was split into two parts due to the reaction phase. The reaction phase affects the input variables as well as the product distribution. The ratio of the H_2O to CO_2 is important for the gas phase to obtain a precise output while it is not relevant for the liquid phase. The same machine learning tools were implemented on both data subsets.

Table 3.5. Minimum and maximum values of the features for liquid and gas phase

	Liquid Phase		**Gas Phase**	
Features	**Min**	**Max**	**Min**	**Max**
x of A	1	4	0.75	2
x of A1/B1	1	4	0.50	2
x of B	1	4	0.50	3
x of X	3	15	3	10
Dope percentage (%)	2	16	0.1	0.5
Cocatalyst percentage (%)	0.1	5	0.1	2.05
2^{nd} cocatalyst percentage (%)	-	-	0.5	0.97
BET surface area (m^2g^{-1})	0.78	358.54	0.1	207
Calcination time (h)	1	24	1	24
Calcination temperature (C°)	120	1200	120	1100
Reaction temperature (C°)	25	35	4	200
Reaction pressure (bar)	0.138	2	1	2
Bandgap (eV)	1.74	4.8	2.06	5.1
H_2O:CO_2	-	-	0.08	13.8

In the articles, the experimental conditions and materials were different hence the products of the reactions varied. As the target value of the dataset, the yield of each product was added. CH_4, O_2, CO, H_2, CH_2O, CH_3OH, C_2H_6O, and CHO_2 were the products of the reactions in the articles. The output value of the dataset was obtained from the sum of yield of the products. The input of the datasets is in Table 3.1. as the features and the output of the datasets is the total yield. The range of the output is 0-8184 μmol.g^{-1}.h^{-1} for liquid data and 0-654.2 μmol.g^{-1}.h^{-1} for gas data.

3.2. Modeling

3.2.1. Preprocessing of dataset

The dataset was preprocessed before machine learning analysis. After constructing the dataset, it was noisy and deficient because the data, which was taken from the various articles, was not homogenous. It is needed for cleaning and organizing, which is called preprocessing. Some preprocessing methods are common while some are distinct for each ML application because different ML methods require different property of the dataset.

At the beginning, the features such as reduction time, reaction medium etc., which were absent for the majority of the articles were eliminated. Additionally, the biased variables which existed in another feature were also removed. For a model, the output variable must be unique; hence, the different type of outputs was eliminated; only total yield (μmol.g^{-1}.h^{-1}) remained. Since the perovskite was used as semi-conductor in the experiments in this study, the perovskites with different functions were also removed to prevent complications. After all these eliminations, 309 data points remain as samples of the database.

Then the empty cells (missing variables) were filled with the mean, mod, or a more meaningful variable. For example, the missing values of reaction temperature and reaction pressure were filled with mods of the variables. For the calcination temperature, on the other hand, it was assumed that if the data is not provided, it usually means that calcination was performed, which is equivalent to say calcination is done in room temperature". In such cases missing value was assumed to be 25 °C. The average of the given values of the calcination time were used to fill the blank for this feature. For the BET surface area, the dataset is filtered by crystal structure and the perovskite, and the average from the remaining data was used to fill the empty cells. For the crystal structure, the Materials Project (MP) (Jain et al., 2013) was used for this purpose; if a unique structure was given in MP for the perovskite, it was used; if there are more than one structure and the correct structure is not known, that data point was also eliminated from the dataset. The ratio of H_2O to CO_2 is crucial for output value, so the blanks were filled with the average of the values which were obtained from the filter due to the light type and catalyst weight for only gas phase dataset.

The absent band gap values were determined using linear regression and the available data as explained in Section 3.2.2. For categorical features, the blanks were filled with “none” because the feature was not used in the articles as part of the experiment. It means, “none” was a meaningful property for the ML models.

As mentioned in Section 3.1 a new variable as the type of the perovskite was added dataset as a feature. Eventually, 29 variables and 30 variables are obtained for liquid and gas data, respectively. As for the data size, there are 133 samples for gas dataset and 176 samples for liquid dataset.

The bandgap prediction was done by linear regression. For this, python (Rossum et al., 1995) was used as the programming language. During the bandgap prediction more features were extracted from the dataset, and it was not separated for different phases as the bandgap is the property of the material itself. The type of perovskite, perovskite, A, x of A, A1/B1, x of A1/B1, B, x of B, X, x of X, additional material, synthesis method of perovskite, doping used, BET surface area, calcination time, calcination temperature, and crystal structure were used as the features to predict bandgap. Because python was used, the categorical data was converted to numerical data by one-hot encoding method. The type of the data was changed as float. The band gap prediction was done to predict the missing bandgap values in the dataset. After separation of blank part of the data as the test set, the shuffling was done to separate the similar data to prevent the bias. Shuffling is crucial for the ML models to prevent learning from similar data and for this dataset similar data were in serial because they were taken from the same articles group by group.

The decision tree (DT) was used to deduce rules for high total yield in R language. In this model, data was separated into gas data and liquid data. The same pre-processing steps were applied separately. In DT, categorical data and numeric data can be used without changing. 75% of the data is separated as train set. The numeric output of the train set was sorted from smallest to largest and split equally as in Table 3.6. The output of the test set also grouped due to the range from the train set.

Table 3.6. The range of the groups for output value in liquid and gas set

Group name	Liquid set	Gas set
Low: C	(0.00,8.69]	(0.00,1.42]
Medium: B	(8.69,143]	(1.42,6.70]
High: A	(143,2352]	(6.70,655]

After that, the numerical values of output were removed, and the categorical values were added. The dataset was shuffled randomly, 60% of data was separated as train set while 15% was separated as validation data each time to prevent the learning from the same part of the data. For each time, one of five (15% of the data) became validation data and the rest was used for training in a loop (five-fold cross validation) to provide learning from all data. The testing was done with the optimum model among five models.

For random forest, the dataset was used to predict total yield value in R language. The categorical data was converted into factor and the numeric data was taken as numeric. The data was split into 75% of the train set and 25% of test set. The shuffling was done.

For gradient boosting, the categorical data was converted to numeric data by one-hot encoding. The shuffling was done and 80% of the data was taken as train set and 20% of data was test set.

In the association rule mining (ARM), the categorical data was taken as factor and the numeric data was grouped by sorting from smallest to largest and splitting equally into categorical values. The x of A, x of A1/B1, x of B, x of X, dope percentage, cocatalyst percentage, 2^{nd} cocatalyst percentage, reaction temperature, and reaction pressure were separated into two groups. BET surface area, calcination time, and bandgap were separated into 10 groups. Calcination temperature and H_2O: CO2 were separated into nine groups. The number of groups differ due to the range and the distribution of the numbers for each feature. All values were taken as factors. The output variable was split into three groups as in decision tree and taken as factor.

3.2.2. Linear Regression

The linear regression model was coded in python language in Google Colab. (Bisong, 2019) The library, *LinearRegression*, from *sklearn* was used. It was used to predict missing bandgap values. The dataset was read, and the pre-processing was applied. After one-hot encoding, the number of samples was 309 and the number of features was 134. The train data was obtained by removing the rows of empty bandgaps which were used to generate test set. The number of samples in the train set was 292 while it was 17 for the test set. After separating the output value from the train set, validation set was obtained from it by 30%. Because of the simplicity of linear regression, no parameters were changed. R-square, mean squared error and root mean squared error were obtained.

3.2.3. Decision Tree

The decision tree model was coded in the R language in R Studio. The libraries, *rpart* and *rattle*, were used. The decision tree was used to predict total yield. In decision tree, the categorical and numeric values were used together. The results were taken for liquid and gas set separately. 25% of the dataset was split as test set. 60% of the dataset was train set and 15% of the dataset was validation set. Five-fold cross validation was applied, and the optimum model was used. The optimum *minsplit* and *cp* values were determined by testing all combinations from five to 30 with five differences for *minsplit* and for *cp* value, from 0.01 to 0.1 with 0.005. The *minsplit* and *cp* values, which gave the highest accuracy were selected. The test set was also grouped with values of train set. With the best hyperparameters, the model prediction was done. After prediction, tree and confusion matrix were obtained. The recall and the precision values for each group of the output were used as the performance measures.

3.2.4. Random Forest

The random forest model was coded in R language. The library, *randomForest*, was used. The random forest was used to predict total yield by regression. In random forest, categorical and numeric data were used together. 75% of the data was train set while 25% of it was test set. Because of the small dataset, the *ntree* was decided as 500. *mtry* (number

of variables tried at each split) was five to avoid overfitting. The importance of the features was calculated, and the graph was obtained. Root mean squared error and R-square were also calculated as performance measures.

3.2.5. Gradient Boosting

The gradient boosting was coded in R language. The library, *xgboost*, was used. The total yield was predicted by regression. The numeric data was used, so one-hot encoding was done. 80% of data was train set and the rest was test set. To prevent overfitting, *nrounds* was decided as 50. The *max.depth* was decided as two and *learning rate* was decided as default value 0.1. Root mean squared error and R-square were calculated as the performance indicators.

3.2.6. Association Rule Mining

The association rule mining was done in R language. The library, *arules*, was used. The relation between the features and the output was obtained. The all features and the output were grouped. The values, support, confidence, and lift were calculated. The values for *minlen* and *maxlen* were decided as one and two, respectively which provides a rule that includes only one feature information. 50 rules by left-hand side (LHS) and right-hand side (RHS) for group A (higher yield) and group C (lower value) were obtained. The ranges and the name of classes are given in Appendix A.

4. RESULTS AND DISCUSSION

4.1. Pre-analysis of Database

The dataset should be analyzed to understand its structure and the properties before the machine learning application is done. The different experimental data were involved in the dataset, hence the relation between total yield, and the features were examined. Most of the analysis was done for UV light and visible light separately. For the bubble charts, the y-axis represents the average yield of the represented feature while the size of the bubbles represents the number of data that belongs to analyzed feature.

The perovskites were analyzed by their sites. B- site was used to decide the type of perovskite. In the dataset, there are six categories of perovskite. In Figure 4.1., the average yield was given depending on the type of perovskite by separating the dataset for the visible and UV light. For ferrite, the experiments were done only with visible light whereas, for tantalate and other, only UV light was used. A distinct difference is not observed between UV and visible light for each category except tantalate. Layered and niobate perovskites show similar performance due to the light types in their categories. For titanate, the difference increases slightly.

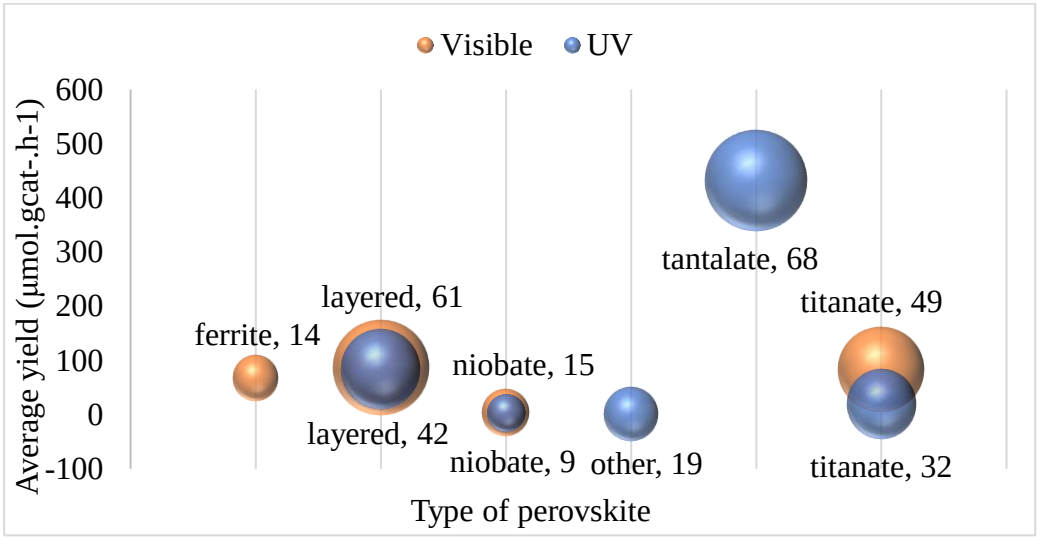

Figure 4.1. Average yield based on the type of the perovskite and the light type

To understand the effect of the reaction phase, the same category was analyzed separately for liquid and gas phase as in Figure 4.2. The superiority of liquid phase against the gas phase is quite evident. Tantalate has the highest average performance in liquid phase. The results belong to the gas phase usually under 100 μmol.gcat^{-1}.h^{-1}.

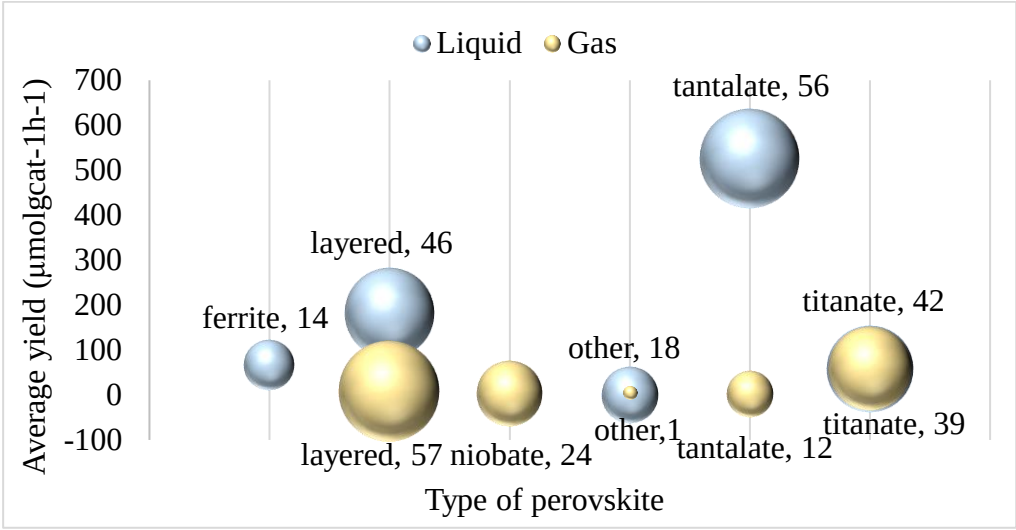

Figure 4.2. Average yield based on the type of the perovskite and the reaction phase

In Figure 4.3., the analysis was done for A-sites of the perovskites. The number of instances is given near the bubble with the name of the A-site to make the figure easier to follow. The dashed bubbles represent results under UV light while the others represent visible light. It is important that the scale of the y-axis is different. The highest performance is shown by the K and Na (K is higher) in tantalates under UV light. The niobates and the others show the worst performance while ferrite and the majority of titanate and layered show an average performance. Sr and Pb in layered, and Ni in titanate have an efficient performance under visible light.

Additional material is one of the features of the dataset; it has two categories “yes” or “no”. If an application such as coating, doping, etc. on perovskite was performed, the feature category was accepted as yes. In Figure 4.4., the additional material increases the performance in UV light significantly. For the visible light, difference is minor. Hence, additional materials do not have an important effect on the yield. In the category “no”, the performance in the visible light is higher than the performance in UV light.

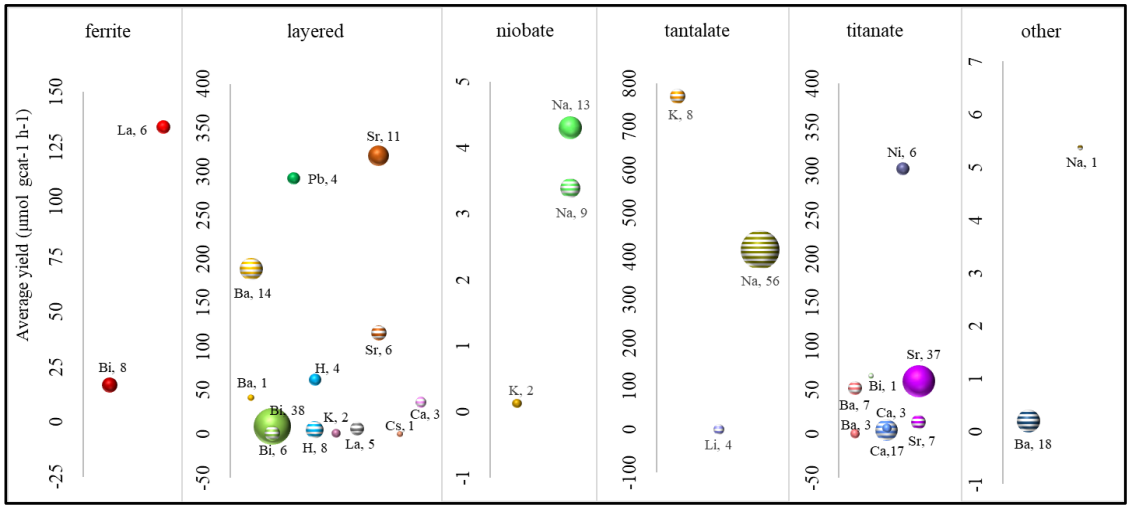

Figure 4.3. Average yield based on A-site due to the category of perovskite (dashed-UV)

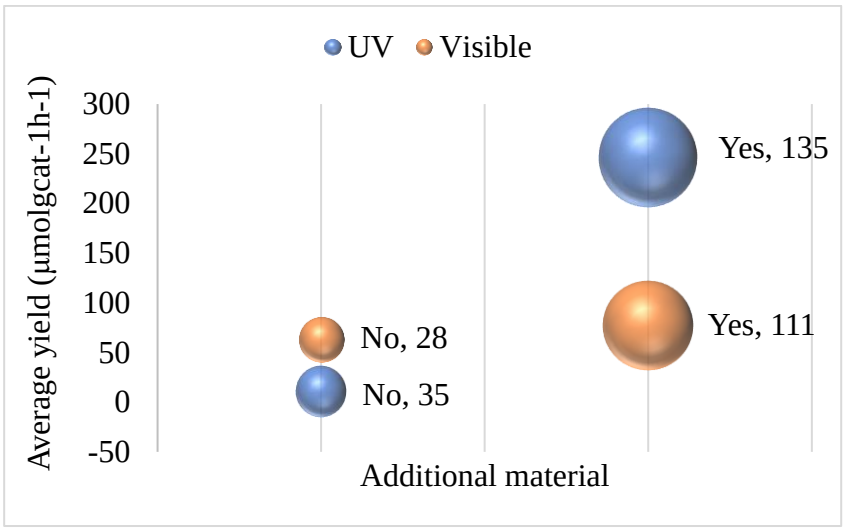

Figure 4.4. The effect of the additional material to the yield

The amount of doped perovskite in the dataset is small (62 out of 309). In UV light, Ba (803 µmol.gcat^{-1}.h^{-1}), La (452 µmol.gcat^{-1}.h^{-1}), Sr (418 µmol.gcat^{-1}.h^{-1}) and Ca (297.5 µmol.gcat^{-1}.h^{-1}) have efficient average yield. In visible light, on the other hand, the combination usage of Fe, S and N doped perovskite shows efficiency with 630.5 µmol.gcat^{-1}.h^{-1} with four data and with a dopant, N, (194.8 µmol.gcat^{-1}.h^{-1}) is also efficient. The rest are under 100 µmol.gcat^{-1}.h^{-1}. The average yield is smaller in visible light for the rest.

The synthesis method of perovskite is also important for the yield. 21 different methods exist in the dataset. Adopting polymer complex (APC) has the highest yield in visible light. In 4 methods, UV light is more effective: hydrothermal (HT), modified polymer complex (MPC), sol-gel (SG) and solid-state (SS) while the difference is more distinctive for the solid-state. Solution combustion, solvo-combustion, solvothermal, sonication, sonochemistry, spin-coating and ultrasonic treatment are named as *others* because of lack of efficiency. The number of instances for HT, SS and MPC is higher as they are more common methods. The HT and SS also have good impact for yield. Pechini (PTP) method is the most inefficient method for yield under UV light. For visible light, HT seems to be the most inefficient method. The methods named as others are not used in this evaluation. APC, electrospinning (ES), and gas bubbling assisted membrane reduction (GBMR) are used only in visible light for this study.

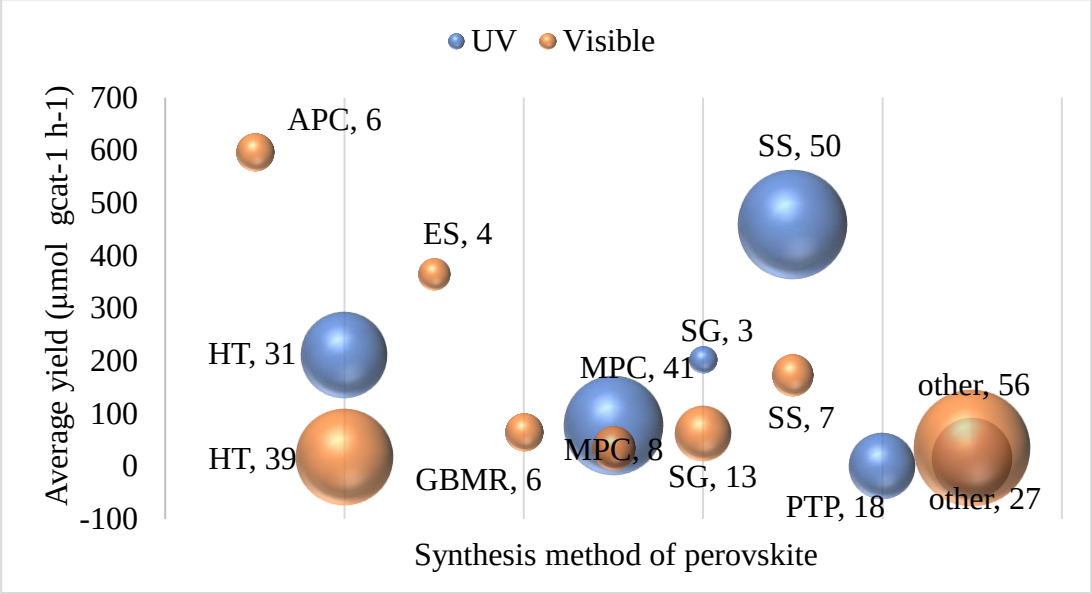

Figure 4.5. The average yield due to the synthesis method of perovskite

The co-catalyst is another feature of the dataset. 148 of 309 samples have co-catalyst. NiO affects the result positively in UV light (1832.8 µmol.gcat^{-1}.h^{-1} yield on average). In UV light, co-catalysts are more efficient such as Ru (705.7 µmol.gcat^{-1}.h^{-1}), Ni (688 µmol.gcat^{-1}.h^{-1}), Au (295.7 µmol.gcat^{-1}.h^{-1}) and Cu (282 µmol.gcat^{-1}.h^{-1}). For visible light, Au (114.2 µmol.gcat^{-1}.h^{-1}) and Ru (119.8 µmol.gcat^{-1}.h^{-1}) give good results. In both types of lights, Au and Ru are efficient. For UV light, using co-catalyst gives better results than

having none. However, in visible light, the co-catalysts, except Au and Ru, do not seem to improve their efficiency.

In Table 4.6., the effect of the co-catalysts on the most used perovskites is given. The red ones represent the visible light, and the blue dashed ones represent the UV light. $NaTaO_3$, $BaLa_4Ti_4O_{15}$, and $SrTiO_3$ are chosen as the most used perovskites in dataset. All co-catalysts have better effect on $NaTaO_3$, and $BaLa_4Ti_4O_{15}$ than $SrTiO_3$. NiO has the highest effect on $NaTaO_3$ as the other co-catalysts.

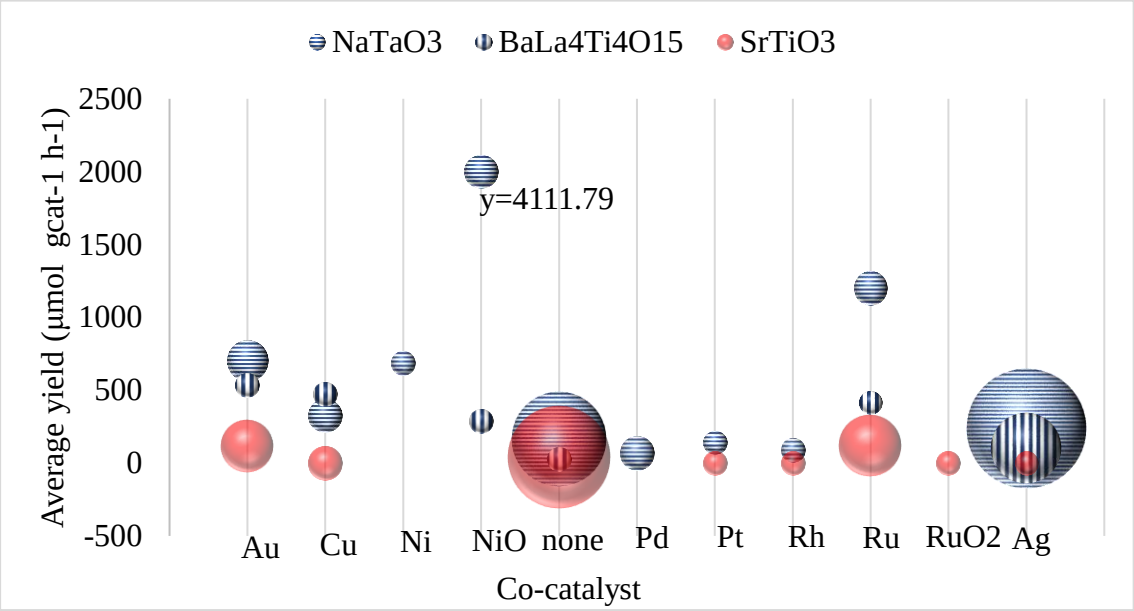

Figure 4.6. The effect of co-catalyst on the most used perovskite (dashed blue represents UV, solid represents visible light)

The phase of the reaction is one of the most important features of the dataset, which was divided into two subsets for the ML models. As shown in Figure 4.7., the results in liquid phase are higher than those in gas phase. In liquid phase, UV light provides more efficiency while visible light seems to be better in gas phase.

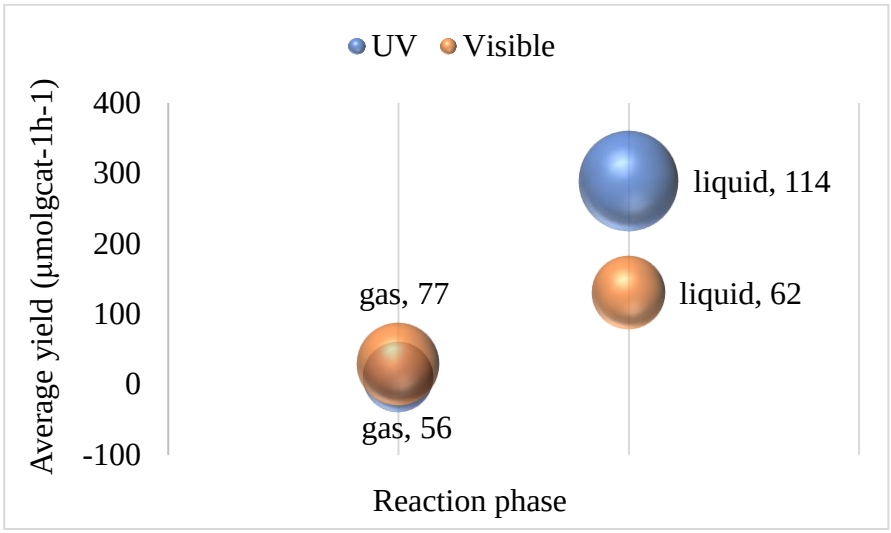

Figure 4.7. The average yield due to the reaction phase in UV and visible lights

In Figure 4.8., the number of instances for synthesis method of perovskite due to the band gap values is given in three ranges. Highest bandgap values belong to the solid-state method. For lower band gap values, on the other hand, the ultrasonic treatment method is better while the hydrothermal method is dominant in the middle range. HT is not the biggest value for the lower and higher ranges but the number instances in these ranges are also high. The polymer complex method is also dominant for higher range.

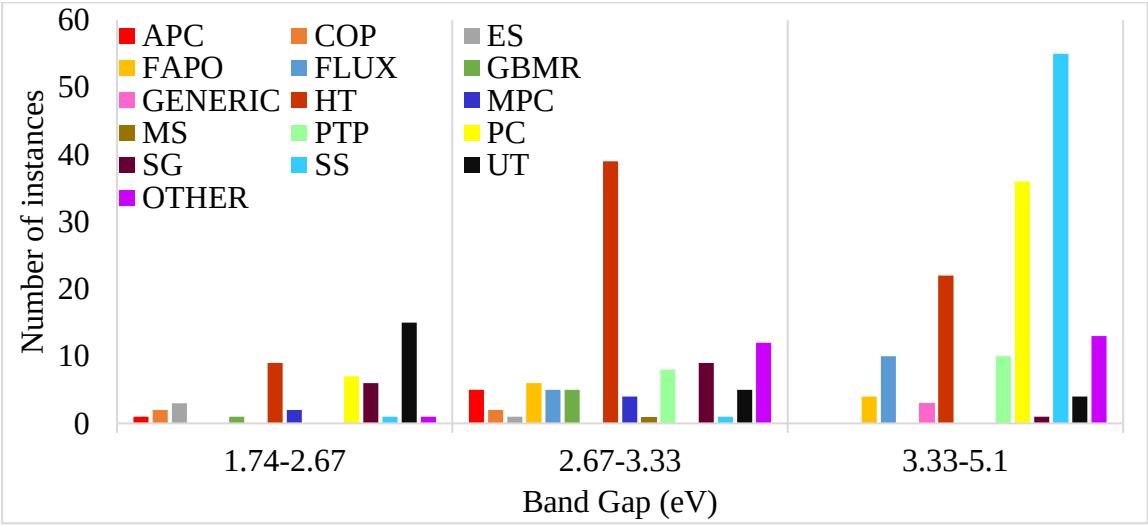

Figure 4.8. The number of instances of the synthesis method of perovskite in bandgap

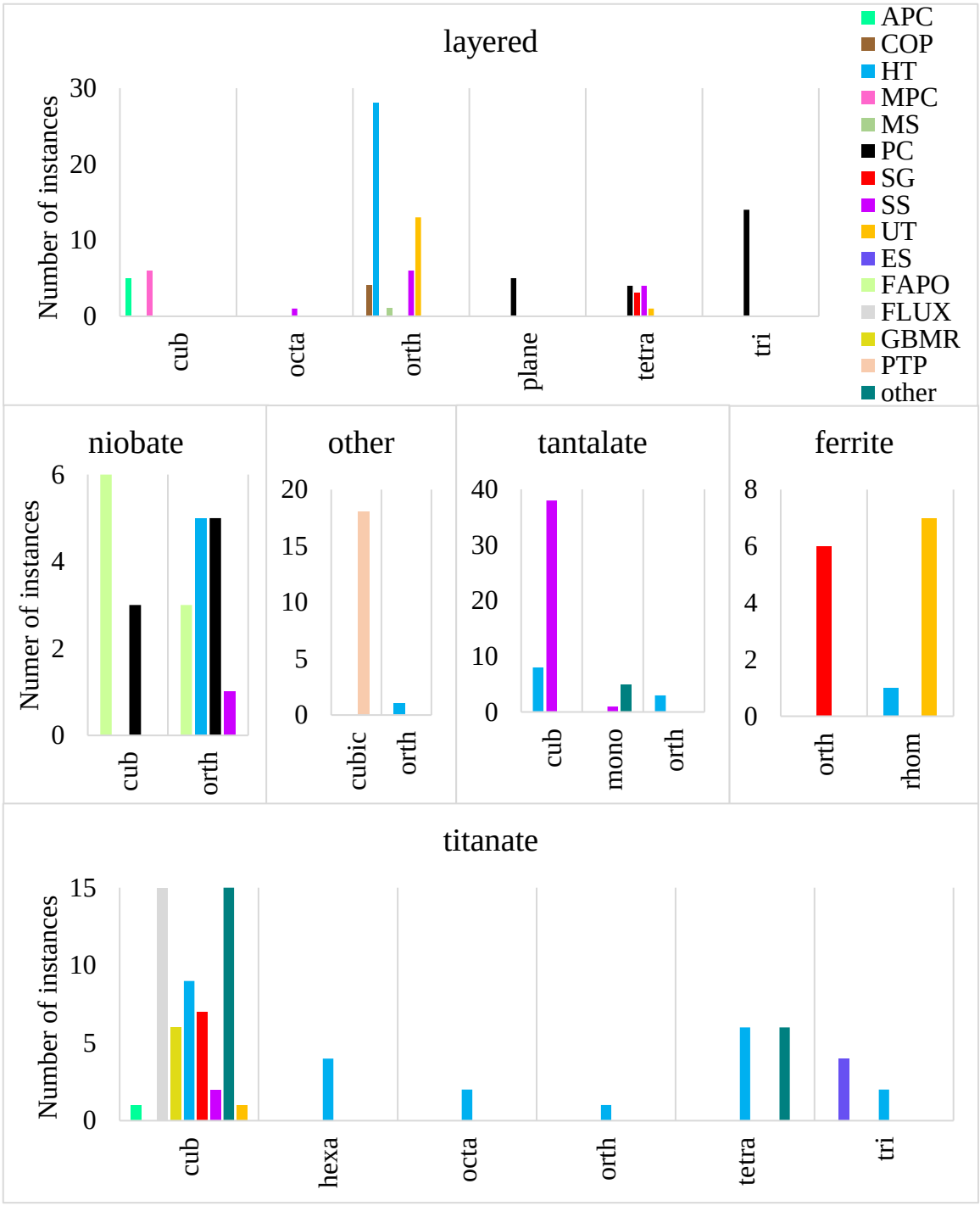

Figure 4.9. The number of instances for crystal structure in synthesis method of perovskite for each perovskite type

In Figure 4.9., the number of the instance for each crystal structure in synthesis method for each perovskite type can be seen. The largest number of synthesis methods of perovskites can form cubic structured perovskites. Besides the diversity in crystal structures are high for

layered type and titanate type perovskites. Orthorhombic crystal structure can be seen in all types of perovskites while it is most frequently observed in layered perovskites. Hydrothermal method is also used for all types of perovskites. This method especially results in orthorhombic structure.

4.2. Feature Effects by Association Rule Mining

The association rule mining was used to determine the rules that explain the relations between the features and the output variable, which was split into three classes as in Table 3.6. for both gas and liquid phase. In Table 4.1., RHS is class A which represents values bigger than 6.7 $\mu mol.gcat^{-1}.h^{-1}$. At LHS, the features of gas phase data are given with relative classes (see Appendix A).

The parameters, support, confidence, lift, and count express the relations. For the 1st rule, the number of data that have LHS and RHS is represented by "count" as seven. The support (# of data have LHS and RHS/# of all data) shows that the fraction of the data which have LHS and RHS to all gas data is 0.05. The confidence (# of data have LHS and RHS/# of LHS in data) shows that the fraction of the data which have LHS and RHS to number of LHS in all data as one which means all the data in LHS has the RHS. The lift (confidence/fraction of RHS to all data) represents the fraction of the confidence variable to fraction of RHS to all data. Lift value is higher than one and it means that the probability is high to having LHS and RHS at the same time. Lift value is important to decide to use LHS and RHS together. For the 1st rule, it can be said that a sample with sonication as synthesis method of perovskite has high probability to has high yield because the lift value is 2.96. For the rest of Table 4.1., the lift value is higher than one however the efficiency decreases in each step.

Table 4.2. show the association rules for the liquid phase; RHS is class A which represents values bigger than 143 $\mu mol.gcat^{-1}.h^{-1}$. At LHS, the features of liquid phase data are given with relative classes (see Appendix A). For the 1st rule in Table 4.2., the count is 34. The support is 0.19. The confidence is 0.79. For the 1st rule, it can be indicated that a sample with one h calcination time has a high probability to have high yield because the lift

value is 2.36. For the rest of Table 4.2., the lift value is higher than one, however the effectivity decreases gradually.

Table 4.1. Association rules with RHS is class A (>6.7 μmol.gcat^{-1}.h^{-1}) in gas phase

LHS	RHS	Support	Confidence	Lift	Count
{Synthesis.Method.of.Perovskite= sonication}	Class A	0.05	1	2.96	7
{cocatalyst=Ru}		0.05	1	2.96	7
{cocatalyst.method=impregnation}		0.08	0.83	2.46	10
{Bet.Surface.Area..m2g.1.=8}		0.1	0.76	2.26	13
{H2O.CO2=2}		0.05	0.7	2.07	7
{Calcination.Temperature..C..=7}		0.1	0.68	2.02	13
{Calcination.Temperature..C..=3}		0.06	0.67	1.97	8
{B=Ti}		0.22	0.66	1.95	29
{Perovskite=SrTiO3}		0.13	0.65	1.93	17
{A=Sr}		0.13	0.65	1.93	17

Table 4.2. Association rules with RHS is class A (>143 μmol.gcat^{-1}.h^{-1}) in liquid phase

LHS	RHS	Support	Confidence	Lift	Count
{Calcination.Time..h.=1}	Class A	0.19	0.79	2.36	34
{Band.Gap=8}		0.16	0.78	2.34	29
{Dope=Ba}		0.1	0.78	2.33	18
{Synthesis.Method.of.Perovskite=solid-state}		0.19	0.75	2.24	33
{deope.percentage....=1}		0.18	0.67	2.01	31
{Calcination.Temperature..C..=4}		0.20	0.61	1.83	35
{cocatalyst.method=liq phase reaction}		0.08	0.61	1.82	14
{Type.of.perovskite=tantalate}		0.19	0.61	1.81	34
{B=Ta}		0.19	0.61	1.81	34
{Perovskite=NaTaO3}		0.16	0.58	1.73	29

4.3. Band Gap Prediction by Linear Regression

Linear regression was used to predict the band gap values, which were not given in the reference articles. In Figure 4.10.a, the original bandgap values versus predicted bandgap values are given for the validation set. R-square, and RMSE are 0.75, and 0.36 for validation set, respectively. In Figure 4.10.b, the original data versus predicted data is given for train set in the same model. R-square, and RMSE are 0.97, and 0.11, respectively for the train set. The learning in this model is good because the range of the result is balanced and there is no outlier.

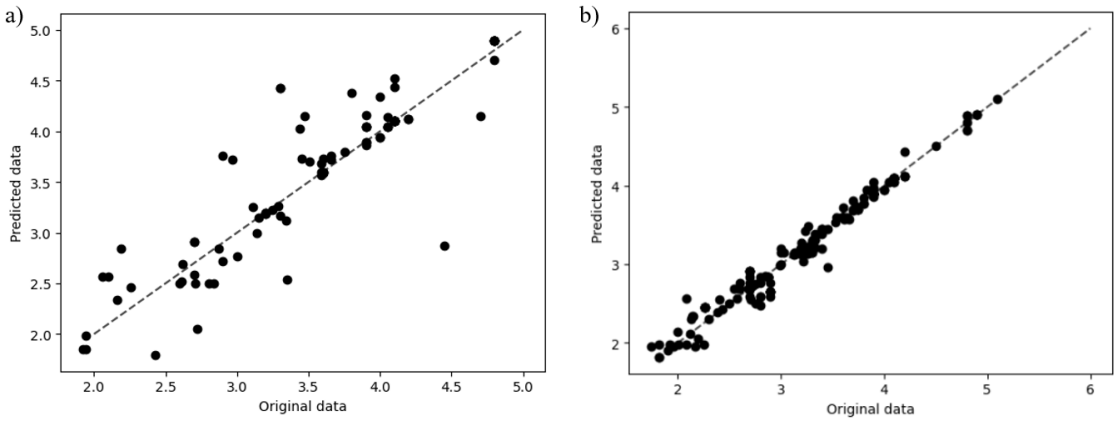

Figure 4.10. The original data versus predicted data for (a) validation set (b) train set

4.4. Decision Tree

The optimal values for *minsplit* and *cp* for both gas and liquid test were chosen as five and 0.015, respectively to obtain the optimal decision tree. In the trees, A, B and C represent the high, medium, and low yield values, respectively.

For gas phase, the decision tree is given in Figure 4.11. The first division in the tree was done according to the synthesis method of perovskites. The high values of the output were decided directly after this separation with the “yes” branch and 16% of the high results can be predicted with 0.94 accuracy. The second division was done by perovskite. With the answer “no”, the branch goes to reaction temperature. In case of the reaction temperature is smaller than 28°C, 23% of the low results are predicted with 0.96 accuracy. In medium

results, four branching from the top occur and 13% of the medium results are predicted with 0.92 accuracy.

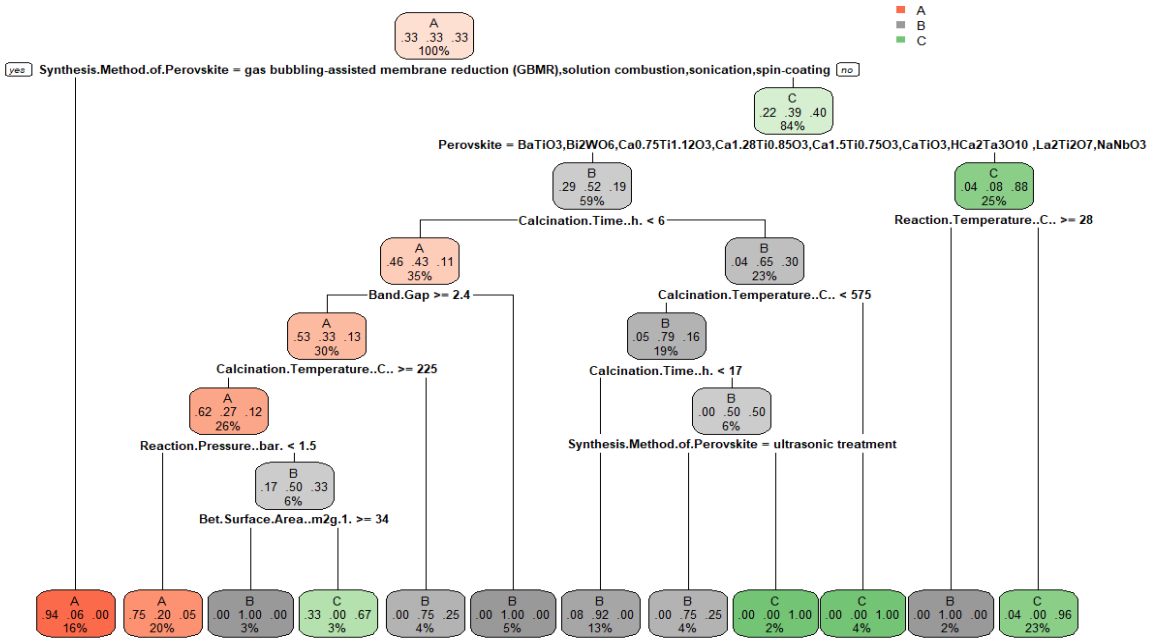

Figure 4.11. Decision tree for gas phase data while minsplit is five and cp is 0.015

The accuracy, precision and recall are tabulated in Table 4.3. for test and train set in gas phase and the confusion matrix for the entire data as well as the classes are given in Table 4.4. and Table 4.5. for testing and training set, respectively.

Table 4.3. Accuracy, precision and recall values of test set and train set for gas phase

	Accuracy	**Precision**	**Recall**
Test set	0.7647	Class A: 0.7000	Class A: 0.7778
		Class B: 0.8889	Class B: 0.5714
		Class C: 0.7333	Class C: 1.000
Train set	0.8889	Class A: 0.8333	Class A: 0.9091
		Class B: 0.9032	Class B: 0.8485
		Class C: 0.9375	Class C: 0.9091

For the test set, seven of nine A data is predicted as A, one of nine is predicted as B and one of nine is predicted as C. This shows that the recall is 0.7778. 11 of 11 C data is predicted as C so the recall is 1.000. At prediction data, three B values are predicted as A so from 10 A prediction, seven prediction is true. This means the precision is 0.7 as in Table 4.3. For experimental works, precision of A is more important than the recall because predicting the high value as low causes not to experiment. However, predicting the low value as high causes experiments with bad results. The accuracy of the test set and train set are 0.7647 and 0.8889, respectively.

Table 4.4. Confusion matrix of the test data for gas phase

Ref / Pred	A	B	C	Total
A	**7**	3	0	10
B	1	**8**	0	9
C	1	3	**11**	15
Total	9	14	11	

Table 4.5. Confusion matrix of the train data for gas phase

Ref / Pred	A	B	C	Total
A	**30**	5	1	36
B	1	**28**	2	31
C	2	0	**30**	32
Total	33	33	33	

The results of feature importance analysis, which show the relative significance of the features, are given in Figure 4.12. The synthesis method of perovskite seems to be the most important feature for the model followed by the perovskite. The reaction conditions are also important for the model. The crystal structure has a minor effect among these features. In the middle of the figure, the features have similar relative importance. Figure 4.12. maybe important to decide the features of the future experiments.

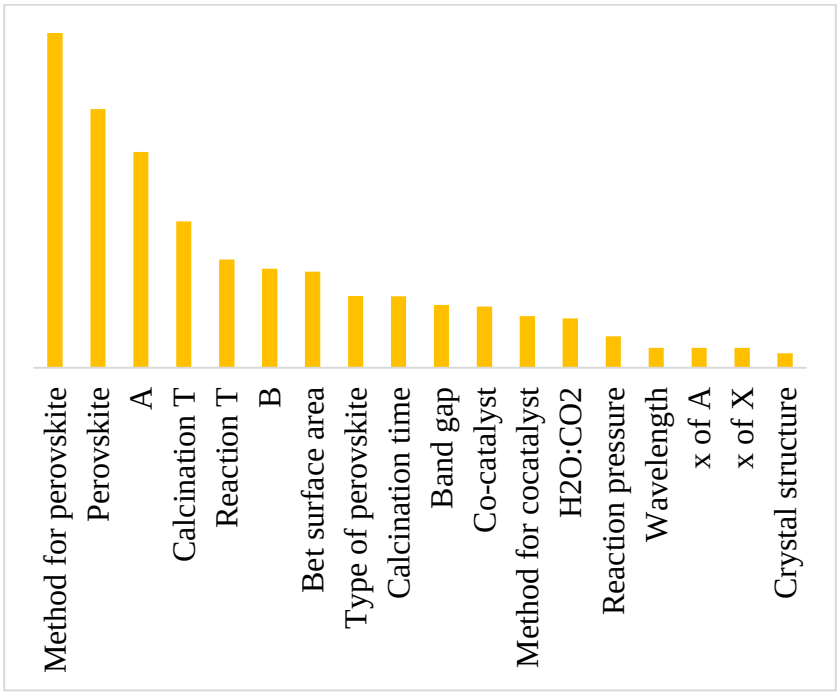

Figure 4.12. The feature importance for gas phase

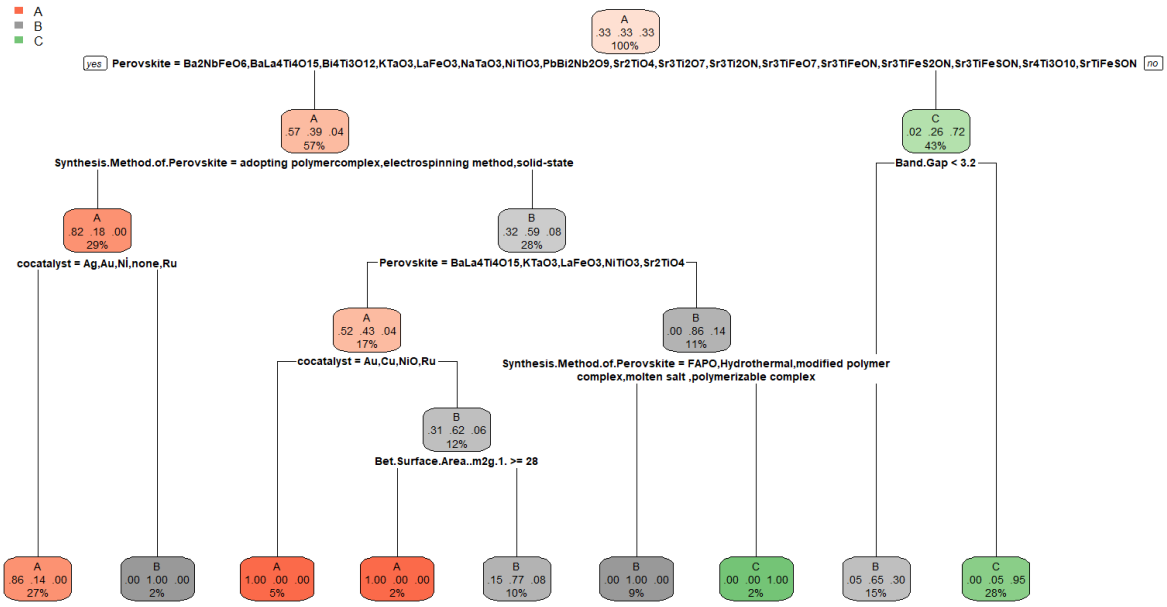

Figure 4.13. Decision tree for liquid phase data while minsplit is five and cp is 0.015

For liquid phase, the decision tree is given in Figure 4.13. The first division in the tree was done based on the perovskite. The branches are synthesis method perovskite for the answer “yes” and bandgap for the answer “no”. Third step is cocatalyst and with the answer

“yes”, 27% of high data is predicted with 0.86 accuracy. The 28% of low data is predicted with 0.95 accuracy when band gap values are higher than 3.2 eV. After four branches, 9% of B data is predicted with 1.00 accuracy.

The accuracy, precision and recall are tabulated in Table 4.6. for test and train set in liquid phase and the confusion matrices are given in Table 4.7. and Table 4.8. for testing and training set, respectively.

Table 4.6. Accuracy, precision and recall values of test set and train set for liquid phase

	Accuracy	**Precision**	**Recall**
Test set	0.8409	Class A: 0.8750	Class A: 1.000
		Class B: 1.000	Class B: 0.6111
		Class C: 0.7059	Class C: 1.000
Train set	0.8712	Class A: 0.8913	Class A: 0.9318
		Class B: 0.7872	Class B: 0.8409
		Class C: 0.9487	Class C: 0.8409

For the testing set, 14 of 14 A data is predicted as A. This shows that the recall is 1.000. 12 of 12 C data is predicted as C so the recall is 1.000. At prediction data, two B values are predicted as A so from 16 A prediction, 14 predictions is true. This means the precision is 0.8750 as in Table 4.6. At prediction data, 11 predictions are done for B with 1.000 precision. The accuracy values of test data and train data are 0.8409 and 0.8712, respectively.

Table 4.7. Confusion matrix of the test set for liquid phase

Ref / **Pred**	**A**	**B**	**C**	**Total**
A	**14**	2	0	16
B	0	**11**	0	11
C	0	5	**12**	17
Total	14	18	12	

Table 4.8. Confusion matrix of the train set for liquid phase

Ref / Pred	A	B	C	Total
A	**41**	5	0	46
B	3	**37**	7	47
C	0	2	**37**	39
Total	44	44	44	

For liquid phase, the perovskite is the most important feature. The property of the perovskite is the first five values for the decision tree as in Figure 4.14. The atom number of X in a perovskite compound (x of X) has a minor effect among these features. In the middle of the figure, the features have similar importance factors group by group.

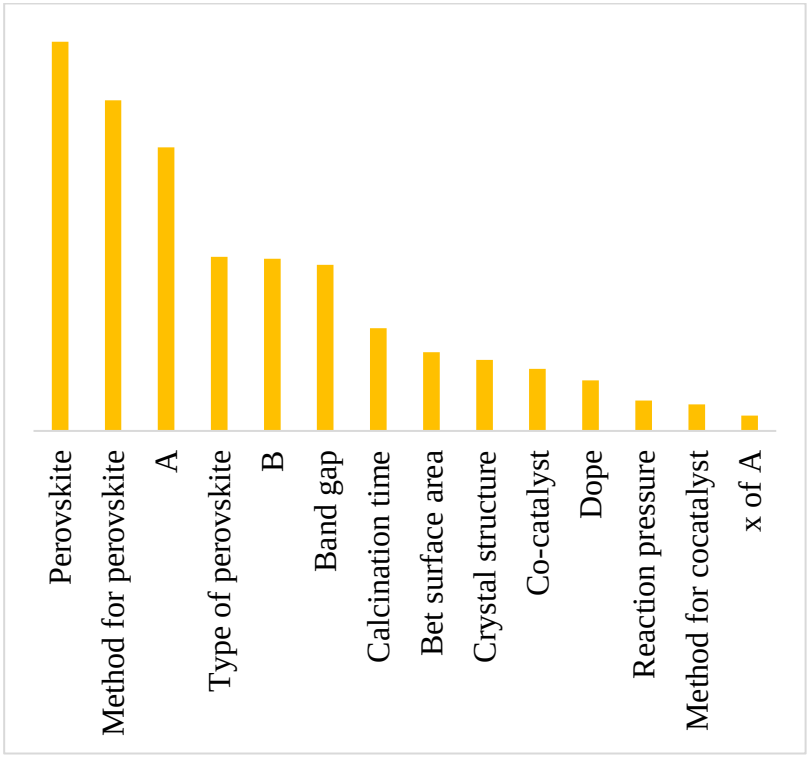

Figure 4.14. The feature importance for liquid phase

4.5. Regression Analysis to Predict the Total Yield

4.5.1. Random Forest

Random forest regression was used to predict the total production yield; separate models were developed for gas and liquid phase dataset. The feature importance was calculated for both cases while the original data versus predicted data for test and train set was plotted.

In Figure 4.15., the original data is represented in x-axis and the predicted data in y-axis in test set. The test data comprises 25% of the gas phase dataset (34 of 133 samples) while the remaining 99 datapoints were used for training. R-square and RMSE are 0.64 and 24.5, respectively. The data values are accumulated under 50 $\mu mol.gcat^{-1}.h^{-1}$ because of the small range of the gas phase data. In Figure 4.16., the experimental and the predicted data are given for the training set. The data is accumulated under 200 $\mu mol.gcat^{-1}.h^{-1}$ and there are two outlier points; the prediction for these points were between 200 $\mu mol.gcat^{-1}.h^{-1}$ and 300 $\mu mol.gcat^{-1}.h^{-1}$ while the original values are between 400-500 $\mu mol.gcat^{-1}.h^{-1}$ and 600-700 $\mu mol.gcat^{-1}.h^{-1}$. Consequently, we can say that the model was not trained to predict the higher values of outputs because of lack of samples. R-square and RMSE are 0.77 and 47.0, respectively. The error values are also affected by the outliers negatively.

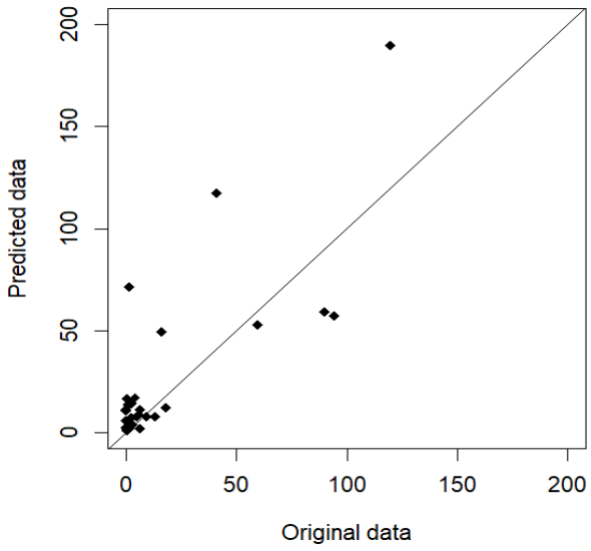

Figure 4.15. The original data versus the predicted data in test set for gas phase

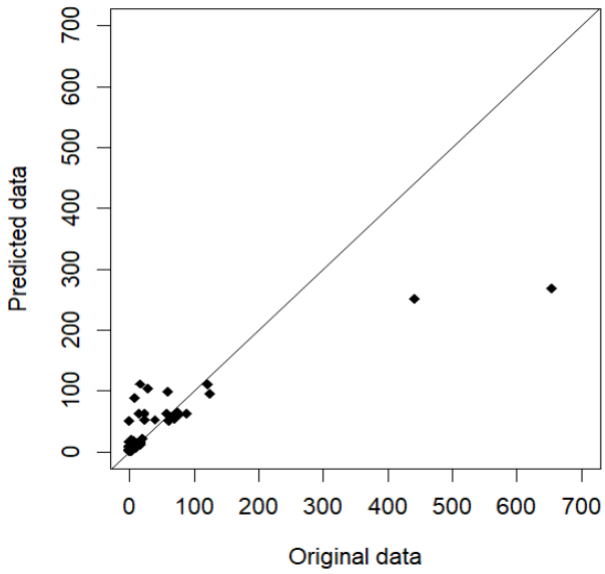

Figure 4.16. The original data versus the predicted data in train set for gas phase

For gas phase dataset, the feature importance figure can be seen in Figure 4.17. In random forest, the band dap is the most important feature fallowed by the synthesis method of perovskite; it should be remembered that the synthesis method of perovskite was the most important feature for decision tree as well; in fact, as can be seen from Figure 4.12. and Figure 4.17., similar features are important for both methods.

In Figure 4.18., the experimental and the predicted data are given for the test set. R-square and RMSE are 0.49 and 221, respectively. The test data comprises 25% of the liquid phase dataset corresponding 44 of 176 samples as the remaining 132 samples were used for training. The data values are accumulated under 1000 $\mu mol.gcat^{-1}.h^{-1}$. However, the range of samples is larger, and the distribution is more balanced here compared to the gas phase models. In Figure 4.19., the experimental and the predicted data are given for the training set as well. This time, the data is accumulated under 2000 $\mu mol.gcat^{-1}.h^{-1}$ and there is one outlier points around 8000 $\mu mol.gcat^{-1}.h^{-1}$ while the prediction for outlier is between 2000 $\mu mol.gcat^{-1}.h^{-1}$ and 4000 $\mu mol.gcat^{-1}.h^{-1}$. R-square and RMSE are 0.77 and 472.5, respectively. The error values are also affected by the outliers negatively. The random forest has better results in test set for gas phase than liquid phase. Despite the R-square values of the train set are the same, in the test set, there is a significant difference.

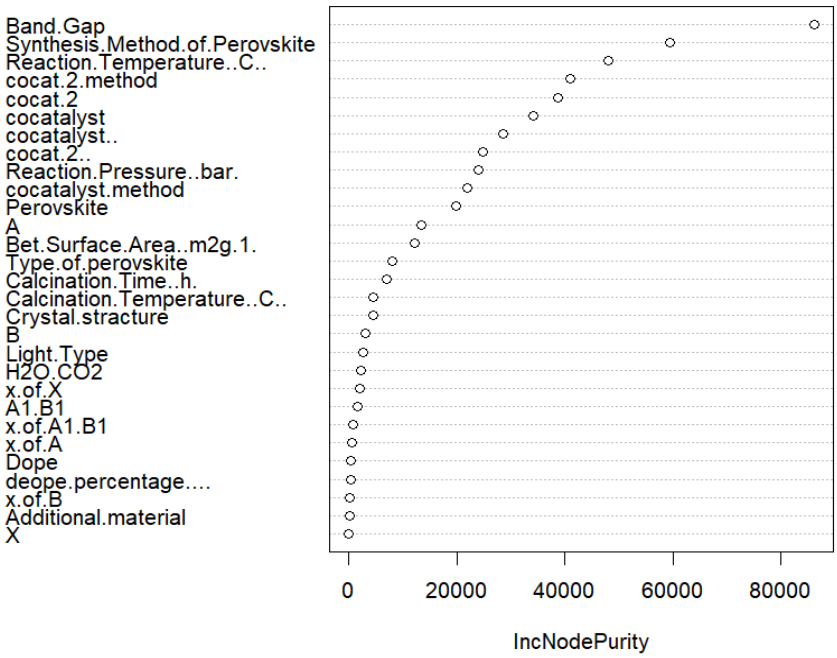

Figure 4.17. Feature importance for gas phase

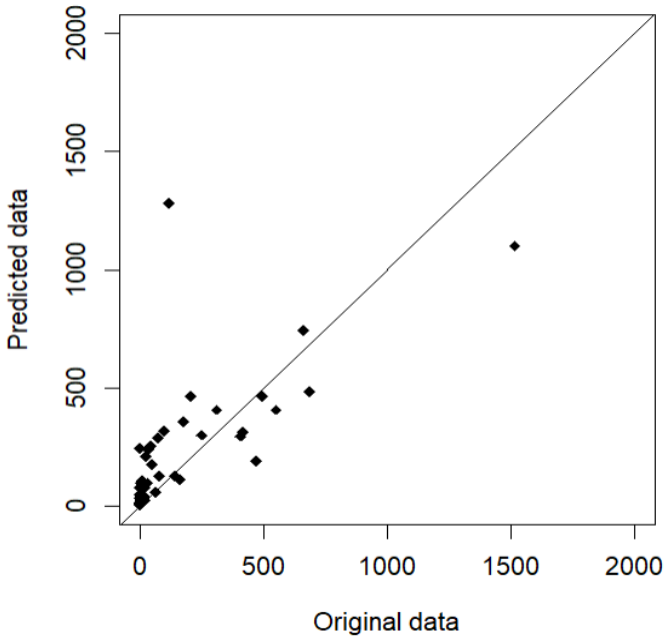

Figure 4.18. The original data versus the predicted data in test set for liquid phase

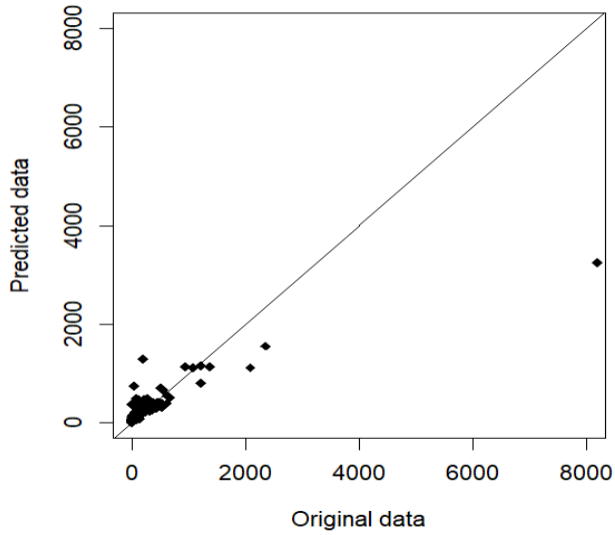

Figure 4.19. The original data versus the predicted data in train set for liquid phase

For liquid phase dataset, the feature importance figure can be seen in Figure 4.20. In random forest, cocatalyst is the most important feature to learning. After that, the synthesis method of cocatalyst is second in the list. As can be remembered, the perovskite was the most important feature for decision tree as in Figure 4.14. Although similar features are important for both methods in liquid phase dataset, the order is not as close as in the case of gas phase dataset.

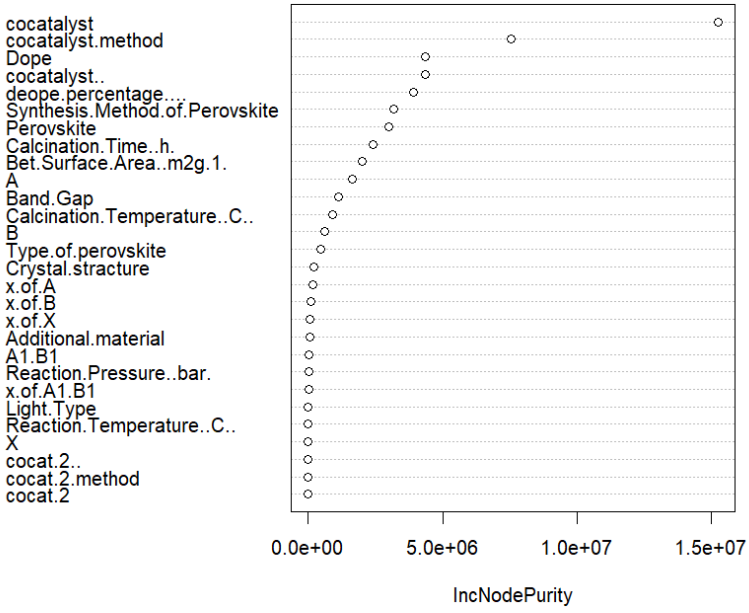

Figure 4.20. Feature importance for liquid phase

4.5.2. Gradient Boosting

The gradient boosting, which is another regression model to predict the yield, was also built for the gas phase and liquid phase dataset separately. For both, 80% of the dataset was train set (27 of 133) while the remaining cases were used for training. In Figure 4.21., the original data versus predicted data for the testing set were plotted for gas phase. Like random forest regression as in Figure 4.15., the data samples are accumulated under 50 μmol.gcat^{-1}.h^{-1}. The R-square and RMSE are 0.65 and 14.75, respectively. The R-square results of random forest and gradient boosting are quite close to each other for the testing set while the gradient boosting result is slightly better.

In Figure 4.22., the experimental versus computed yield plot is presented for the training; again, the samples are accumulated under 200 μmol.gcat^{-1}.h^{-1} like in random forest; however, the effect of the outlier (single) is not as strong as that in the random forest. The R-square and RMSE are 0.95 and 26.5, respectively indicating that the fitness of the model is also higher than random forest model.

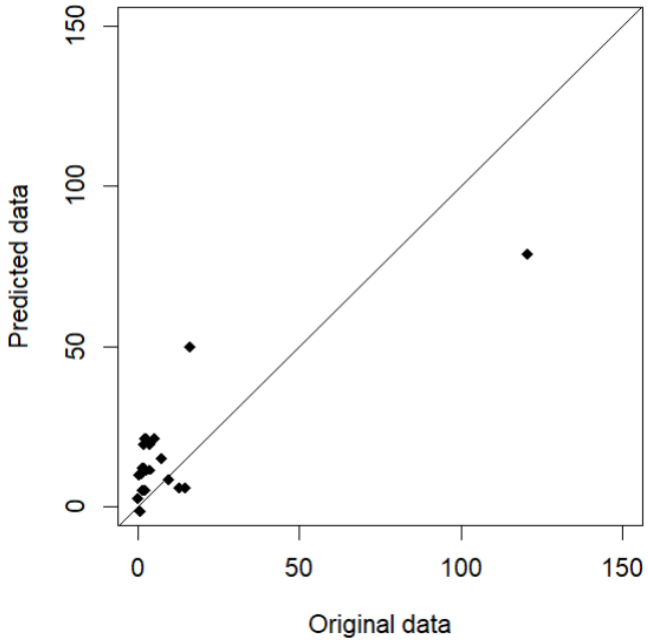

Figure 4.21. The original data versus predicted data at test set for gas phase

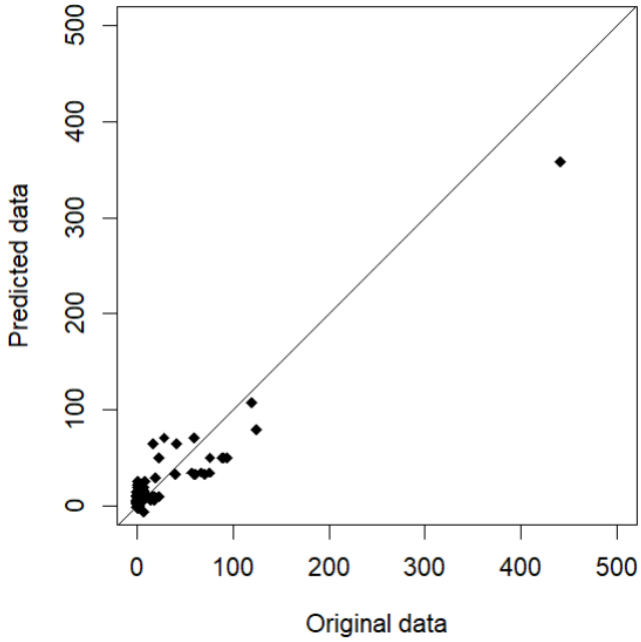

Figure 4.22. The original data versus predicted data at train set for gas phase

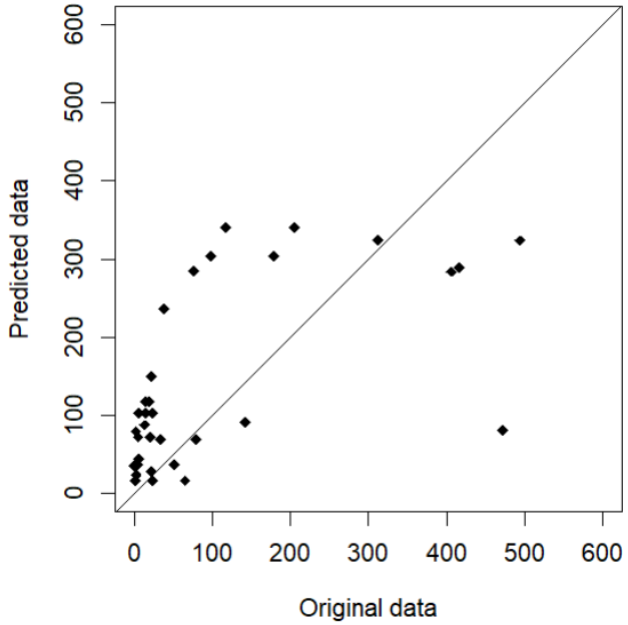

Figure 4.23. The original data versus predicted data at test set for liquid phase

In Figure 4.23., the experimental versus predicted data can be observed for the testing is presented for liquid phase, in which the distribution of the test data is better than the gas phase. The R-square and RMSE are 0.79 and 145.6, respectively. The fitness of the gradient boosting model was also better than random forest. The same plot for training is presented in Figure 4.24, which is quite like that in Figure 4.19 for random forest. However, the learning performance of gradient boosting is much better than random forest, so the outlier sample does not affect learning (R-square and RMSE are 0.94 and 232.7, respectively).

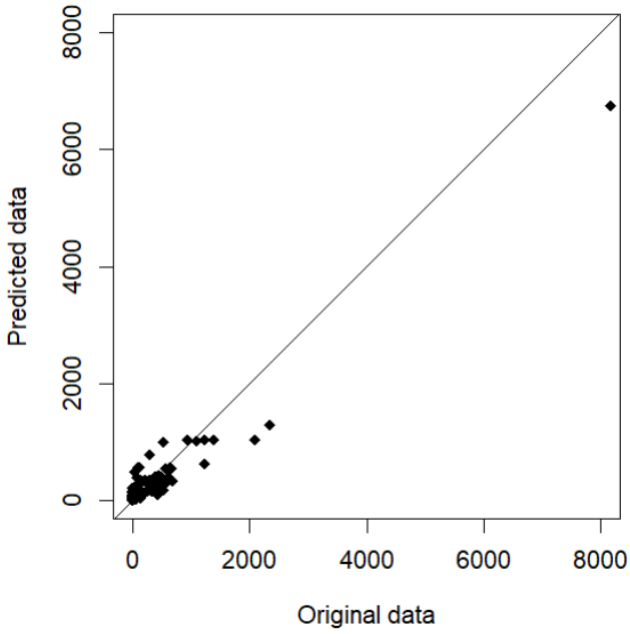

Figure 4.24. The original data versus predicted data at train set for liquid phase

5. CONCLUSION AND RECOMMENDATIONS

5.1. Conclusion

In this study, a database was formed from the published articles about CO_2 photoreduction on perovskites. 309 data points were collected from 61 experimental papers which were searched from Web of Science. Due to the reaction phase, ML analysis was performed for gas and liquid phase separately as the number of descriptors are not the same (29 and 30 descriptors were selected for liquid and gas phase datasets respectively). First, the pre-processing was performed to prepare the data for ML analysis. Both datasets were analyzed using simple descriptive statistics to understand the relations of the inputs and the output variables. Linear regression was used for bandgap prediction to calculate the bandgap for the data points to calculate the missing bandgap values from the available data. The total yield of the CO_2 reduction reaction, however, was predicted by the random forest, and gradient boosting algorithms in R language. For deducing the rules for high total yield, decision tree was used in R. The association between the features and the output was also examined by association rule mining in R.

In linear regression analysis, 17 missing band gap values were predicted with R-square of 0.75. Predicted values were placed in the dataset and the other analysis was done with this form of the dataset. In decision tree analysis, the accuracy for test set was obtained as 0.76 and 0.84 for gas and liquid phase databases, respectively. Feature importance analysis indicated that the perovskite synthesis method for gas phase and type of perovskite for liquid phase were the most important features for the DT model. In random forest, RMSE for testing were found as 24.5 and 221.0 and the most important features were found to be band gap and co-catalysts for the gas and liquid phase, respectively. With gradient boosting, RMSE for testing were 14.75 and 145.6 for gas and liquid phase, respectively.

The performance of the gradient boosting algorithm, *XGBoost,* is better than random forest. DT deduced the rules for liquid phase better than gas phase. Besides, *XGBoost* showed better performance for liquid phase on the other hand RF showed better performance for gas phase.

5.2. Recommendations

Finally, with the results of this work and the experience of the study, the recommendations for the future work can be listed as follows,

- More data can be collected to increase the size of the database. Since the model will be trained with the larger dataset, the accuracy will increase for test set.
- Other machine learning methods can be applied to this dataset to extract more information.
- The features of the dataset can be improved by adding more information such as elemental properties.

REFERENCES

Bazdaric, K., D. Sverko, I. Salaric, A. Martinović, and M. Lucijanic, 2021, "The ABC of Linear Regression Analysis: What Every Author and Editor Should Know", *European Science Editing*, Vol. 47.

Bentéjac, C., A. Csörgő, and G. Martínez-Muñoz, 2021, "A Comparative Analysis of Gradient Boosting Algorithms", *Artificial Intelligence Review*, Vol. 54, No. 3, pp. 1937–1967.

Bi, Y., M. F. Ehsan, Y. Huang, J. Jin, and T. He, 2015, "Synthesis of Cr-doped SrTiO3 Photocatalyst and Its Application in Visible-Light-Driven Transformation of CO2 into CH4", *Journal of CO2 Utilization*, Vol. 12, pp. 43–48.

Bisong, E., 2019, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Apress, California.

Bonaccorso, G., 2017, *Machine Learning Algorithms*, Packt Publishing, Burmingham.

Can, E., A. Jalal, I. G. Zirhlioglu, A. Uzun, and R. Yildirim, 2021, "Predicting Water Solubility in Ionic Liquids Using Machine Learning Towards Design of Hydro-Philic/Phobic Ionic Liquids", *Journal of Molecular Liquids*, Vol. 332, pp. 115848.

Can, E. and R. Yildirim, 2019, "Data Mining in Photocatalytic Water Splitting Over Perovskites Literature for Higher Hydrogen Production", *Applied Catalysis B: Environmental*, Vol. 242, pp. 267–283.

Cao, S., B. Shen, T. Tong, J. Fu, and J. Yu, 2018, "2D/2D Heterojunction of Ultrathin MXene/Bi2WO6 Nanosheets for Improved Photocatalytic CO2 Reduction", *Advanced Functional Materials*, Vol. 28, No. 21, pp. 1800136.

Caruntu, D., T. Rostamzadeh, T. Costanzo, S. Salemizadeh Parizi, & G. Caruntu, 2015, "Solvothermal Synthesis and Controlled Self-Assembly of Monodisperse Titanium-Based Perovskite Colloidal Nanocrystals", *Nanoscale*, Vol. 7, No. 30, pp. 12955–12969.

Chen, A., X. Zhang, L. Chen, S. Yao, and Z. Zhou, 2020, "A Machine Learning Model on Simple Features for CO2 Reduction Electrocatalysts", *Journal of Physical Chemistry C*, Vol. 124, No. 41, pp. 22471–22478.

Chen, T. and C. Guestrin, 2016, "XGBoost: A Scalable Tree Boosting System", *paper presented at the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Chen, W., Y. Wang, and W. Shangguan, 2019, "Metal (Oxide) Modified (M= Pd, Ag, Au and Cu) H2SrTa2O7 for Photocatalytic CO2 Reduction with H2O: The Effect of Cocatalysts on Promoting Activity Toward CO and H2 Evolution", *International Journal of Hydrogen Energy*, Vol. 44, No. 8, pp. 4123–4132.

Chen, X., J. Wang, C. Huang, S. Zhang, H. Zhang, Z. Li, and Z. Zou, 2015, "Barium Zirconate: A New Photocatalyst for Converting CO2 Into Hydrocarbons Under UV Irradiation", *Catalysis Science & Technology*, Vol. 5, No. 3, pp. 1758–1763.

Chicco, D., M. J. Warrens, and G. Jurman, 2021, "The Coefficient of Determination R-Squared Is More Informative Than SMAPE, MAE, MAPE, MSE And RMSE In Regression Analysis Evaluation", *PeerJ Computer Science*, Vol. 7, pp. 1–24.

Collado, L., M. Gomez-Mendoza, M. García-Tecedor, F. E. Oropeza, A. Reynal, J. R. Durrant, D. P. Serrano, and V. A. de la Peña O'Shea, 2023, "Towards the Improvement of Methane Production in CO2 Photoreduction Using Bi2WO6/TiO2 Heterostructures", *Applied Catalysis B: Environmental*, Vol. 324, pp. 122206.

Dai, W., H. Xu, J. Yu, X. Hu, X. Luo, X. Tu, and L. Yang, 2015, "Photocatalytic Reduction of CO2 into Methanol and Ethanol Over Conducting Polymers Modified Bi2WO6 Microspheres Under Visible Light", *Applied Surface Science*, Vol. 356, pp. 173–180.

Dai, W., J. Yu, Y. Deng, X. Hu, T. Wang, and X. Luo, 2017, "Facile Synthesis of MoS2/Bi2WO6 Nanocomposites for Enhanced CO2 Photoreduction Activity Under Visible Light Irradiation", *Applied Surface Science*, Vol. 403, pp. 230–239.

Dasireddy, V. D. B. C. and B. Likozar, 2022, "Photocatalytic CO2 Reduction to Methanol Over Bismuth Promoted BaTiO3 Perovskite Nanoparticle Catalysts", *Renewable Energy*, Vol. 195, pp. 885–895.

Domor Mienye, I., Y. Sun, and Z. Wang, 2019, "Prediction Performance of Improved Decision Tree-Based Algorithms: A Review", *Procedia Manufacturing*, Vol. 35, pp. 698–703.

Fresno, F., S. Galdón, M. Barawi, E. Alfonso-González, C. Escudero, V. Pérez-Dieste, C. Huck-Iriart, and V. A. de la Peña O'Shea, 2021, "Selectivity in UV Photocatalytic CO2 Conversion Over Bare and Silver-Decorated Niobium-Tantalum Perovskites", *Catalysis Today*, Vol. 361, pp. 85–93.

Fresno, F., P. Jana, P. Reñones, J. M. Coronado, D. P. Serrano, and V. A. De La Peña O'Shea, 2017, "CO2 Reduction Over NaNbO3 and NaTaO3 Perovskite Photocatalysts", *Photochemical & Photobiological Sciences*, Vol. 16, No. 1, pp. 17–23.

He, J., X. Wang, S. Lan, H. Tao, X. Luo, Y. Zhou, and M. Zhu, 2022, "Breaking the Intrinsic Activity Barriers of Perovskite Oxides Photocatalysts for Catalytic CO2 Reduction via Piezoelectric Polarization", *Applied Catalysis B: Environmental*, Vol. 317, pp. 121747.

Huang, C., Z. Li, and Z. Zou, 2016, "A Perspective on Perovskite Oxide Semiconductor Catalysts for Gas Phase Photoreduction of Carbon Dioxide", *Mrs Communications*, Vol. 6, No. 3, pp. 216–225.

Huang, J., Y.-F. Li, and M. Xie, 2015, “An Empirical Analysis of Data Preprocessing for Machine Learning-Based Software Cost Estimation”, *Information and Software Technology*, Vol. 67, pp. 108-127.

Humayun, M., Y. Qu, F. Raziq, R. Yan, Z. Li, X. Zhang, and L. Jing, 2016, “Exceptional Visible-Light Activities of TiO2-Coupled N-Doped Porous Perovskite LaFeO3 for 2,4-Dichlorophenol Decomposition and CO2 Conversion”, *Environmental Science and Technology*, Vol. 50, No. 24, pp. 13600–13610.

Humayun, M., L. Xu, L. Zhou, Z. Zheng, Q. Fu, and W. Luo, 2018, “Exceptional Co-Catalyst Free Photocatalytic Activities of B and Fe Co-Doped SrTiO3 for CO2 Conversion and H2 Evolution”, *Nano Research*, Vol. 11, No. 12, pp. 6391–6404.

Hwang, S. Y., H. J. Jang, Y. J. Kim, J. Y. Maeng, C. K. Rhee, and Y. Sohn, 2023, “Eu(III)–BaTiO3 Nanoparticles and BaTiO3/TiO2/Ti Sheets; Photocatalytic and Electrocatalytic CO2 Reduction”, *Materials Science in Semiconductor Processing*, Vol. 153.

Iizuka, K., T. Wato, Y. Miseki, K. Saito, and A. Kudo, 2011, “Photocatalytic Reduction of Carbon Dioxide Over Ag Cocatalyst-Loaded ALa4Ti4O15 (A = Ca, Sr, and Ba) Using Water as A Reducing Reagent”, *Journal of the American Chemical Society*, Vol. 133, No. 51, pp. 20863–20868.

Im, Y., S. M. Park, and M. Kang, 2017, “Effect of Ca/Ti Ratio on the Core–Shell Structured CaTiO3@basalt Fiber for Effective Photoreduction of Carbon Dioxide”, *Bulletin of the Korean Chemical Society*, Vol. 38, No. 3, pp. 397–400.

Jain, A., S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, 2013, “Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation”, *APL Materials*, Vol. 1, No. 1, pp. 11002.

Jeyalakshmi, V., R. Mahalakshmy, K. R. Krishnamurthy, and B. Viswanathan, 2016, "Photocatalytic Reduction of Carbon Dioxide in Alkaline Medium on La Modified Sodium Tantalate with Different Co-Catalysts Under UV–Visible Radiation", *Catalysis Today*, Vol. 266, pp. 160–167.

Jeyalakshmi, V., R. Mahalakshmy, K. R. Krishnamurthy, and B. Viswanathan, 2018, "Strontium Titanates with Perovskite Structure as Photo Catalysts for Reduction of CO2 by Water: Influence of Co-Doping with N, S & Fe", *Catalysis Today*, Vol. 300, pp. 152–159.

Jeyalakshmi, V., R. Mahalakshmy, K. Ramesh, P. V. C. Rao, N. V. Choudary, G. Sri Ganesh, K. Thirunavukkarasu, K. R. Krishnamurthy, and B. Viswanathan, 2014, "Visible Light Driven Reduction of Carbon Dioxide with Water on Modified Sr3Ti2O7 Catalysts", *RSC Advances*, Vol. 5, No. 8, pp. 5958–5966.

Jia, P., Y. Li, Z. Zheng, Y. Wang, and T. Liu, 2022, "Ferroelectric Polarization Promotes the Excellent CO2 Photoreduction Performance of Bi4Ti3O12 Synthesized by Molten Salt Method", *Journal of Alloys and Compounds*, Vol. 920.

Jiang, S., K. Zhao, M. Al-Mamun, Y. L. Zhong, P. Liu, H. Yin, L. Jiang, S. Lowe, J. Qi, R. Yu, D. Wang, and H. Zhao, 2019, "Design of Three-Dimensional Hierarchical TiO2/SrTiO3 Heterostructures Towards Selective CO2 Photoreduction", *Inorganic Chemistry Frontiers*, Vol. 6, No. 7, pp. 1667–1674.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, 2017, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 3146–3154.

Khan, H., H. Charles, and C. S. Lee, 2022, "Synergistic Effect Stemming from Vertically Anchored Seamless 2D MoSe2 Nanosheets on 1D NiTiO3 Nanofibers Toward CO2 Photoreduction", *Journal of CO2 Utilization*, Vol. 61, pp. 102058.

Khan, H., S. Kang, H. Charles, and C. S. Lee, 2022, "Epitaxial Growth of Flower-Like MoS2 on One-Dimensional Nickel Titanate Nanofibers: A "Sweet Spot" for Efficient Photoreduction of Carbon Dioxide", *Frontiers in Chemistry*, Vol. 10, pp. 837915.

Kong, X. Y., W. Q. Lee, A. R. Mohamed, and S. P. Chai, 2019 "Effective Steering of Charge Flow Through Synergistic Inducing Oxygen Vacancy Defects and P-N Heterojunctions in 2D/2D Surface-Engineered Bi2WO6/BiOI Cascade: Towards Superior Photocatalytic CO2 Reduction Activity", *Chemical Engineering Journal*, Vol. 372, pp. 1183–1193.

Kong, X. Y., W. L. Tan, B. J. Ng, S. P. Chai, and A. R. Mohamed, 2017, "Harnessing Vis–NIR Broad Spectrum for Photocatalytic CO2 Reduction over Carbon Quantum Dots-Decorated Ultrathin Bi2WO6 Nanosheets", *Nano Research*, Vol. 10, No. 5, pp. 1720–1731.

Kumar, A., A. Kumar, and V. Krishnan, 2020, "Perovskite Oxide Based Materials for Energy and Environment-Oriented Photocatalysis", *ACS Catalysis*, Vol. 10, No. 17, pp. 10253–10315.

Kumar, A., G. Sharma, M. Naushad, T. Ahamad, R. C. Veses, and F. J. Stadler, 2019, "Highly Visible Active Ag2CrO4/Ag/BiFeO3@RGO Nano-Junction for Photoreduction of CO2 and Photocatalytic Removal of Ciprofloxacin and Bromate Ions: The Triggering Effect of Ag and RGO", *Chemical Engineering Journal*, Vol. 370, pp. 148–165.

Kwak, B. S., J. Y. Do, N. K. Park, and M. Kang, 2017, "Surface Modification of Layered Perovskite Sr2TiO4 for Improved CO2 Photoreduction with H2O to CH4", *Scientific Reports*, Vol. 7, No. 1, pp. 1–15.

Kwak, B. S. and M. Kang, 2015, "Photocatalytic Reduction of CO2 with H2O Using Perovskite CaxTiyO3", *Applied Surface Science*, Vol. 337, pp. 138–144.

Kwak, B. S. and M. Kang, 2017, "Evaluation of Photoreduction Performance of CO2 to CH4 with H2O over Alkaline-Earth-Metal-Based Perovskite Nanoparticles", *Journal of Nanoscience and Nanotechnology*, Vol. 17, No. 10, pp. 7351–7357.

Lee, C. H., J. Sim, and D. H. Lim, 2019, "A Conducting Polymer Coated Perovskite Supported on Glass Fiber Substrate for Gas-Phase CO2 Conversion to Methane", *Energy Procedia*, Vol. 158, pp. 534–540.

L'Heureux, A., K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, 2017, "Machine Learning with Big Data: Challenges and Approaches", *IEEE Access*, Vol. 5, pp. 7776–7797.

Li, D., S. Ouyang, H. Xu, D. Lu, M. Zhao, X. Zhang, and J. Ye, 2016, "Synergistic Effect of Au and Rh on SrTiO3 in Significantly Promoting Visible-Light-Driven Syngas Production from CO2 and H2O", *Chemical Communications*, Vol. 52, No. 35, pp. 5989–5992.

Li, M., P. Li, K. Chang, T. Wang, L. Liu, Q. Kang, S. Ouyang, and J. Ye, 2015, "Highly Efficient and Stable Photocatalytic Reduction of CO2 to CH4 over Ru Loaded NaTaO3", *Chemical Communications*, Vol. 51, No. 36, pp. 7645–7648.

Li, M., L. Zhang, X. Fan, Y. Zhou, M. Wu, and J. Shi, 2015, "Highly Selective CO2 Photoreduction to CO over g-C3N4/Bi2WO6 Composites Under Visible Light", *Journal of Materials Chemistry A*, Vol. 3, No. 9, pp. 5189–5196.

Li, P., S. Ouyang, G. Xi, T. Kako, and J. Ye, 2012, "The Effects of Crystal Structure and Electronic Structure on Photocatalytic H2 Evolution and CO2 Reduction over Two Phases of Perovskite-Structured NaNbO3", *Journal of Physical Chemistry C*, Vol. 116, No. 14, pp. 7621–7628.

Li, P., S. Ouyang, Y. Zhang, T. Kako, and J. Ye, 2012, "Surface-Coordination-Induced Selective Synthesis of Cubic and Orthorhombic NaNbO3 and Their Photocatalytic Properties", *Journal of Materials Chemistry A*, Vol. 1, No. 4, pp. 1185–1191.

Li, P., H. Xu, L. Liu, T. Kako, N. Umezawa, H. Abe, and J. Ye, 2014, "Constructing Cubic–Orthorhombic Surface-Phase Junctions of NaNbO3 Towards Significant Enhancement of CO2 Photoreduction", *Journal of Materials Chemistry A*, Vol. 2, No. 16, pp. 5606–5609.

Li, X., H. Pan, W. Li, and Z. Zhuang, 2012, "Photocatalytic Reduction of CO2 to Methane over HNb3O8 Nanobelts", *Applied Catalysis A: General*, Vol. 413–414, pp. 103–108.

Liang, L., F. Lei, S. Gao, Y. Sun, X. Jiao, J. Wu, S. Qamar, and Y. Xie, 2015, "Single Unit Cell Bismuth Tungstate Layers Realizing Robust Solar CO2 Reduction to Methanol", *Angewandte Chemie International Edition*, Vol. 54, No. 47, pp. 13971–13974.

Luo, C., J. Zhao, Y. Li, W. Zhao, Y. Zeng, and C. Wang, 2018, "Photocatalytic CO2 Reduction over SrTiO3: Correlation Between Surface Structure and Activity", *Applied Surface Science*, Vol. 447, pp. 627–635.

Mahesh, B., 2018, "Machine Learning Algorithms-A Review", *International Journal of Science and Research*, Vol. 9, No. 1, pp. 381–386.

Mantovani, R., T. Horváth, R. Cerri, S. B. Junior, J. Vanschoren, and A. Carvalho, 2018, "An Empirical Study on Hyperparameter Tuning of Decision Trees", arXiv:1812.02207 [cs.LG].

Mateo, D., J. Albero, and H. García, 2019, "Titanium-Perovskite-Supported RuO2 Nanoparticles for Photocatalytic CO2 Methanation", *Joule*, Vol. 3, No. 8, pp. 1949–1962.

Mora-Hernandez, J. M., A. M. Huerta-Flores, and L. M. Torres-Martínez, 2018, "Photoelectrocatalytic Characterization of Carbon-Doped NaTaO3 Applied in the Photoreduction of CO2 Towards the Formaldehyde Production", *Journal of CO2 Utilization*, Vol. 27, pp. 179–187.

Morais, E., K. Stanley, K. R. Thampi, and J. A. Sullivan, 2021, "Scope for Spherical Bi2WO6 Quazi-Perovskites in the Artificial Photosynthesis Reaction—The Effects of Surface Modification with Amine Groups", *Catalysis Letters*, Vol. 151, No. 1, pp. 293–305.

Murcia-López, S., V. Vaiano, M. C. Hidalgo, J. A. Navío, and D. Sannino, 2015, "Photocatalytic Reduction of CO2 over Platinised Bi2WO6-Based Materials", *Photochemical & Photobiological Sciences*, Vol. 14, No. 4, pp. 678–685.

Nakanishi, H., K. Iizuka, T. Takayama, A. Iwase, and A. Kudo, 2016, "Highly Active NaTaO3-Based Photocatalysts for CO2 Reduction to Form CO Using Water as the Electron Donor", *Chemistry Sustainability Energy Materials*, Vol. 10, No. 1, pp. 112–118.

Odabaşı Özer, Ç. and R. Yıldırım, 2019, "Performance Analysis of Perovskite Solar Cells in 2013–2018 Using Machine-Learning Tools", *Nano Energy*, Vol. 56, pp. 770–791.

Oku, T., 2020, "Crystal Structures of Perovskite Halide Compounds Used for Solar Cells", *Reviews on Advanced Materials Science*, Vol. 59, No. 1, pp. 264–305.

Oral, B., E. Can, and R. Yildirim, 2022, "Analysis of Photoelectrochemical Water Splitting Using Machine Learning", *International Journal of Hydrogen Energy*, Vol. *47*, No. 45, pp. 19633–19654.

Parida, K. M., K. H. Reddy, S. Martha, D. P. Das, and N. Biswal, 2010, "Fabrication of Nanocrystalline LaFeO3: An Efficient Sol–Gel Auto-Combustion Assisted Visible Light Responsive Photocatalyst for Water Decomposition", *International Journal of Hydrogen Energy*, Vol. 35, No. 22, pp. 12161–12168.

Paulista, L. O., J. Albero, R. J. E. Martins, R. A. R. Boaventura, V. J. P. Vilar, T. F. C. V. Silva, and H. García, 2021, "Turning Carbon Dioxide and Ethane into Ethanol by Solar-Driven Heterogeneous Photocatalysis over RuO2-And NiO-Co-Doped SrTiO3", *Catalysts*, Vol. 11, No. 4, pp. 461.

Phokha, S., S. Pinitsoontorn, S. Maensiri, and S. Rujirawat, 2014, "Structure, Optical and Magnetic Properties of LaFeO3 Nanoparticles Prepared by Polymerized Complex Method", *Journal of Sol-Gel Science and Technology*, Vol. *71*, No. 2, pp. 333–341.

Pino-Mejías, R., A. P. Erez-Fargallo, C. Rubio-Bellido, and J. A. Pulido-Arcas, 2016, "Comparison of Linear Regression and Artificial Neural Networks Models to Predict Heating and Cooling Energy Demand, Energy Consumption and CO2 Emissions", *Energy*, Vol. 118, pp. 24-36.

Praveen Kumar, D., A. Putta Rangappa, K. H. Do, Y. Hong, M. Gopannagari, K. Arun Joshi Reddy, P. Bhavani, D. Amaranatha Reddy, and T. Kyu Kim, 2022, "Noble Metal Free Few-Layered Perovskite-Based Ba2NbFeO6 Nanostructures on Exfoliated g-C3N4 Layers as Highly Efficient Catalysts for Enhanced Solar Fuel Production", *Applied Surface Science*, Vol. 572, pp. 151406.

Priyam, A., R. Gupta, A. Rathee, and S. Srivastava, 2013, "Comparative Analysis of Decision Tree Classification Algorithms", *International Journal of Current Engineering and Technology*, Vol. 3, No. 2.

Probst, P., M. N. Wright, and A. L. Boulesteix, 2019, "Hyperparameters and Tuning Strategies for Random Forest", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 9, No. 3.

Prokhorenkova, L., G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, 2017, "CatBoost: Unbiased Boosting with Categorical Features", arXiv:1706.09516 [cs.LG].

Qiu, J., Q. Wu, G. Ding, Y. Xu, and S. Feng, 2016, "A Survey of Machine Learning for Big Data Processing", *EURASIP Journal on Advances in Signal Processing*, Vol. 67.

Ran, J., M. Jaroniec, and S. Z. Qiao, 2018, "Cocatalysts in Semiconductor-based Photocatalytic CO2 Reduction: Achievements, Challenges, and Opportunities", *Advanced Materials*, Vol. 30, No. 7, pp. 1704649.

Reshmi Varma, P. C., 2018, *Perovskite Photovoltaics*, Elsevier, Kerala.

Rokach, L. & O. Maimon, 2005, *Data Mining and Knowledge Discovery Handbook*, Springer, Boston.

Rossum, G., 1995, *Python Reference Manual*, Centrum voor Wiskunde en Informatica, Amsterdam.

Saadetnejad, D., B., Oral, E., Can, and R. Yıldırım, 2022, "Machine Learning Analysis of Gas Phase Photocatalytic CO2 Reduction for Hydrogen Production", *International Journal of Hydrogen Energy*, Vol. 47, No. 45, pp. 19655–19668.

Schneider, A., G. Hommel, and M. Blettner, 2010, "Linear Regression Analysis", *Deutsches Ärzteblatt International*, Vol. 107, No. 44, pp. 776–782.

Schonlau, M., and R. Y. Zou, 2020, "The Random Forest Algorithm for Statistical Learning", *The Stata Journal*, Vol. 20, No. 1, pp. 3–29.

Shao, X., X. Yin, and J. Wang, 2018, "Nanoheterostructures of Potassium Tantalate and Nickel Oxide for Photocatalytic Reduction of Carbon Dioxide to Methanol in Isopropanol", *Journal of Colloid and Interface Science*, Vol. 512, pp. 466–473.

Shi, H., G. Chen, C. Zhang, and Z. Zou, 2014, "Polymeric g-C3N4 Coupled with NaNbO3 Nanowires Toward Enhanced Photocatalytic Reduction of CO2 into Renewable Fuel", *ACS Catalysis*, Vol. 4, No. 10, pp. 3637–3643.

Shi, H., C. Zhang, C. Zhou, and G. Chen, 2015, "Conversion of CO2 into Renewable Fuel over Pt–g-C3N4/KNbO3 Composite Photocatalyst", *RSC Advances*, Vol. 5, No. 113, pp. 93615–93622.

Sun, Z., H. Wang, Z. Wu, and L. Wang, 2018, "g-C3N4 Based Composite Photocatalysts for Photocatalytic CO2 Reduction", *Catalysis Today*, Vol. 300, pp. 160–172.

Teramura, K., S. Okuoka, H. Tsuneoka, T. Shishido, and T. Tanaka, 2010, "Photocatalytic Reduction of CO2 Using H2 as Reductant over ATaO3 Photocatalysts (A = Li, Na, K)", *Applied Catalysis B: Environmental*, Vol. 96, No. 3–4, pp. 565–568.

Tu, W., Y. Zhou, and Z. Zou, 2014, "Photocatalytic Conversion of CO2 into Renewable Hydrocarbon Fuels: State-of-the-Art Accomplishment, Challenges, and Prospects", *Advanced Materials*, Vol. 26, No. 27, pp. 4607–4626.

Vu, N. N., C. C. Nguyen, S. Kaliaguine, and T. O. Do, 2017, "Reduced Cu/Pt–HCa2Ta3O10 Perovskite Nanosheets for Sunlight-Driven Conversion of CO2 into Valuable Fuels", *Advanced Sustainable Systems*, Vol. 1, No. 9.

Wang, J., C. Huang, X. Chen, H. Zhang, Z. Li, & Z. Zou, 2015, "Photocatalytic CO2 Reduction of BaCeO3 with 4f Configuration Electrons", *Applied Surface Science*, Vol. 358, pp. 463–467.

Wang, L., J. Zhao, H. Liu, and J. Huang, 2018, "Design, Modification and Application Of Semiconductor Photocatalysts", *Journal of the Taiwan Institute of Chemical Engineers*, Vol. 93, pp. 590–602.

Wang, S., K. Teramura, T. Hisatomi, K. Domen, H. Asakura, S. Hosokawa, and T. Tanaka, 2020, "Effective Driving of Ag-Loaded and Al-Doped SrTiO3 under Irradiation at $\lambda >$ 300 nm for the Photocatalytic Conversion of CO2 by H2O", *ACS Applied Energy Materials*, Vol. 3, No. 2, pp. 1468–1475.

Wang, Y., M. Liu, W. Chen, L. Mao, and W. Shangguan, 2019, "Ag Loaded on Layered Perovskite H2SrTa2O7 To Enhance the Selectivity of Photocatalytic CO2 Reduction with H2O", *Journal of Alloys and Compounds*, Vol. 786, pp. 149–154.

Wang, Z., K. Teramura, S. Hosokawa, & T. Tanaka, 2015, "Photocatalytic Conversion of CO2 in Water over Ag-Modified La2Ti2O7", *Applied Catalysis B: Environmental*, Vol. 163, pp. 241–247.

Wu, J., Y. Huang, W. Ye, and Y. Li, 2017, "CO2 Reduction: From the Electrochemical to Photochemical Approach", *Advanced Science*, Vol. 4, No. 11.

Wu, X., C. Wang, Y. Wei, J. Xiong, Y. Zhao, Z. Zhao, J. Liu, and J. Li, 2019, "Multifunctional Photocatalysts of Pt-Decorated 3DOM Perovskite-Type SrTiO3 with Enhanced CO2 Adsorption and Photoelectron Enrichment for Selective CO2 Reduction with H2O to CH4", *Journal of Catalysis*, Vol. 377, pp. 309–321.

Xin, L., J. Wen, J. Low, Y. Fang, and J. Yu, 2014, "Design and Fabrication of Semiconductor Photocatalyst for Photocatalytic Reduction of CO2 to Solar Fuel", *Science China Materials*, Vol. 57, pp. 70-100.

Xu, G., X. Huang, V. Krstic, S. Chen, X. Yang, C. Chao, G. Shen, and G. Han, 2014, "Hydrothermal synthesis of single-crystalline tetragonal perovskite PbTiO3 nanosheets with dominant (001) or (111) facets", *CrystEngComm*, Vol. 16, No. 21, pp. 4373–4376.

Yang, G., J. Xiong, M. Lu, W. Wang, W. Li, Z. Wen, S. Li, W. Li, R. Chen, and G. Cheng, 2022, "Co-Embedding Oxygen Vacancy and Copper Particles into Titanium-Based Oxides (TiO2, BaTiO3, and SrTiO3) Nanoassembly for Enhanced CO2 Photoreduction Through Surface/Interface Synergy", *Journal of Colloid and Interface Science*, Vol. 624, pp. 348–361.

Yin, S., Y. Zhu, Z. Ren, C. Chao, X. Li, X. Wei, G. Shen, Y. Han, and G. Han, 2014, "Facile Synthesis of PbTiO3 Truncated Octahedra via Solid-State Reaction and Their Application in Low-Temperature CO Oxidation by Loading Pt Nanoparticles", *Journal of Materials Chemistry A*, Vol. 2, No. 24, pp. 9035–9039.

Yoshida, H., L. Zhang, M. Sato, T. Morikawa, T. Kajino, T. Sekito, S. Matsumoto, and H. Hirata, 2015, "Calcium Titanate Photocatalyst Prepared by A Flux Method for Reduction of Carbon Dioxide with Water", *Catalysis Today*, Vol. 251, No. 132–139.

Zangeneh, N. P., S. Sharifnia, and E. Karamian, 2020, "Modification of Photocatalytic Property of BaTiO3 Perovskite Structure by Fe2O3 Nanoparticles for CO2 Reduction

in Gas Phase", *Environmental Science and Pollution Research*, Vol. 27, No. 6, pp. 5912–5921.

Zeng, S., P. Kar, U. K. Thakur, and K. Shankar, 2018, "A Review on Photocatalytic CO2 Reduction Using Perovskite Oxide Nanomaterials", *Nanotechnology*, Vol. 29.

Zhang, L., Y. Yang, Z. Zhou, J. Li, G. Chen, L. Zhou, Y. Qiu, and Y. Sun, 2023, "Redispersion of Exsolved Cu Nanoparticles on LaFeO3 Photocatalyst for Tunable Photocatalytic CO2 Reduction", *Chemical Engineering Journal*, Vol. 452, pp. 139273.

Zhao, Q. and S. S. Bhowmick, 2003, *Association Rule Mining: A Survey*, Technical Report, Nanyang Technical University.

Zhou, H., J. Guo, P. Li, T. Fan, D. Zhang, and J. Ye, 2013, "Leaf-Architectured 3D Hierarchical Artificial Photosynthetic System of Perovskite Titanates Towards CO2 Photoreduction Into Hydrocarbon Fuels", *Scientific Reports*, Vol. 3, No. 1, pp. 1–9.

Zhou, H., P. Li, J. Guo, R. Yan, T. Fan, D. Zhang, and J. Ye, 2014, "Artificial Photosynthesis on Tree Trunk Derived Alkaline Tantalates with Hierarchical Anatomy: Towards CO2 Photo-Fixation into CO and CH4", *Nanoscale*, Vol. 7, No. 1, pp. 113–120.

Zhou, L., S. Pan, J. Wang, and A. V. Vasilakos, 2017, "Machine Learning on Big Data: Opportunities and Challenges", *Neurocomputing*, Vol. 237, pp. 350–361.

APPENDIX A: THE INPUT AND OUTPUT GROUPS FOR ARM

Table A.1. The range of the groups of ARM for liquid and gas phase

Feature	Class	Liquid	Gas
Total yield	C	(-0.1,9.85]	(-0.1,1.02]
	B	(9.85,142]	(1.02,6.2]
	A	(142,8184]	(6.2,655]
x of A	1	(-0.1,1]	(0.749,1]
	2	(1,4]	(1,2]
x of A1/B1	0	(-0.1,0]	(-0.1,0]
	1	(0,4]	(0,2]
x of B	0	(-0.1,1]	(0.497,1]
	1	(1,4]	(1,3]
x of X	0	(-0.1,3]	(2.99,3]
	1	(3,15]	(3,10]
Dope percentage	0	(-0.1,0]	(-0.1,0]
	1	(0,16]	(0,0.5]
Cocatalyst percentage	0	(-0.1,0.1]	(-0.1,0]
	1	(0.1,5]	(0,2.05]
2nd cocat percentage	0	0	(-0.1,0]
	1	-	(0,0.971]
Bet surface area	1	(0.422,1.67]	(-0.1,3]
	2	(1.67,2.04]	(3,4.04]
	3	(2.04,3.7]	(4.04,6.49]
	4	(3.7,12]	(6.49,17.8]
	5	(12,20]	(17.8,24]
	6	(20,25]	(24,26.4]
	7	(25,359]	(26.4,30]
	8	-	(20,39.8]

	9	-	(39.8,72.6]
	10	-	(72.6,207]
Calcination time	1	(0.977,1]	1
	2	(1,2]	2
	3	(2,4]	3
	4	(4,10]	4
	5	(10,24]	5
	6	-	6
	7	-	8
	8	-	10
	9	-	20
	10	-	24
Calcination temp	1	(23.8,310]	(119,150]
	2	(310,600]	(150,350]
	3	(600,675]	(350,400]
	4	(675,900]	(400,500]
	5	(900,1000]	(500,550]
	6	(1000,1100]	(550,600]
	7	(1100,1200]	(600,700]
	8	-	(700,880]
	9	-	(880,1100]
Reaction temp	0	(24.99,25]	(3.8,25]
	1	(25,35]	(25,200]
Reaction pressure	0	(0.136,1]	(0.999,1]
	1	(1,2]	(1,2]
Band gap	1	(1.74,2.34]	(2.06,2.26]
	2	(2.34,2.9]	(2.26,2.7]
	3	(2.9,3.2]	(2.7,2.8]
	4	(3.2,3.45]	(2.8,2.9]
	5	(3.45,3.6]	(2.9,3.2]
	6	(3.6,3.9]	(3.2,3.3]

	7	(3.9,4.05]	(3.3,3.44]
	8	(4.05,4.1]	(3.44,3.75]
	9	(4.1,4.8]	(3.75,4.05]
	10	-	(4.05,5.1]
H2O:CO2	1	-	0.08
	2	-	0.1
	3	-	0.138
	4	-	0.25
	5	-	0.33
	6	-	2
	7	-	4
	8	-	9.7
	9	-	13.8

APPENDIX B: ASSOCIATION RULES FOR CLASS C IN ARM

Table B.1. Association rules with RHS is class C (<9.85 μmol.gcat^{-1}.h^{-1}) in liquid phase

LHS	RHS	Support	Confidence	Lift	Count
{Perovskite=BaZrO3}	Class C	0.06	1	2.98	10
{B=Zr}		0.06	1	2.98	10
{Type.of.perovskite=other}		0.1	1	2.98	18
{Synthesis.Method.of.Perovskite= Pechini}		0.1	1	2.98	18
{Perovskite=CaTiO3}		0.08	0.93	2.78	14
{Band.Gap=4}		0.07	0.86	2.56	12
{Bet.Surface.Area..m2g.1.=2}		0.06	0.85	2.52	11
{A=Ca}		0.08	0.83	2.48	15
{Calcination.Temperature..C..=5}		0.1	0.81	2.41	17
{Synthesis.Method.of.Perovskite= flux}		0.07	0.8	2.39	12
{cocatalyst.method=chemical reduction}		0.05	0.75	2.24	9
{Calcination.Temperature..C..=1}		0.07	0.72	2.15	13
{Band.Gap=5}		0.07	0.7	2.1	12
{Type.of.perovskite=titanate}		0.16	0.7	2.06	29
{Perovskite=SrTiO3}		0.07	0.67	1.99	12

Table B.2. Association rules with RHS is class C (<1.02 μmol.gcat^{-1}.h^{-1}) in gas phase

LHS	RHS	Support	Confidence	Lift	Count
{Synthesis.Method.of.Perovskite=Sol-gel}	Class C	0.05	1	2.96	7
{Type.of.perovskite=tantalate}		0.08	0.92	2.71	11
{B=Ta}		0.13	0.71	2.09	17
{Calcination.Time..h.=8}		0.11	0.68	2.02	15
{Bet.Surface.Area..m2g.1.=3}		0.06	0.67	1.97	8
{cocatalyst=Ag}		0.05	0.64	1.88	7
{Calcination.Time..h.=10}		0.05	0.64	1.88	7
{Synthesis.Method.of.Perovskite=solid-state}		0.06	0.62	1.82	8
{Bet.Surface.Area..m2g.1.=1}		0.1	0.56	1.67	13
{Calcination.Temperature..C..=1}		0.05	0.54	1.59	7
{Calcination.Temperature..C..=6}		0.1	0.54	1.59	14
{Band.Gap=8}		0.06	0.53	1.58	8
{cocatalyst.method=photodeposition}		0.12	0.52	1.52	16
{Band.Gap=5}		0.06	0.5	1.48	8
{Bet.Surface.Area..m2g.1.=7}		0.05	0.47	1.38	7

APPENDIX C: ARTICLES INVOLVED IN DATASET

Table C.1. Article involved in datasets

Article Number	Reference
1	(Kwak & Kang, 2015)
2	(H. Zhou et al., 2013)
3	(Kwak & Kang, 2017)
4	(Jiang et al., 2019)
5	(Dasireddy & Likozar, 2022)
6	(X. Wu et al., 2019)
7	(P. Li, Ouyang, Xi, et al., 2012)
8	(Collado et al., 2023)
9	(Hwang et al., 2023)
10	(Vu et al., 2017)
11	(Fresno et al., 2017)
12	(Morais et al., 2021)
13	(Fresno et al., 2021)
14	(Z. Wang et al., 2015)
15	(H. Zhou et al., 2014)
16	(Mateo et al., 2019)
17	(D. Li et al., 2016)
18	(Teramura et al., 2010)
19	(Shi et al., 2015)
20	(Shi et al., 2014)
21	(Y. Wang et al., 2019)
22	(W. Chen et al., 2019)
23	(Kong et al., 2019)
24	(Cao et al., 2018)
25	(Dai et al., 2017)

26	(Kong et al., 2017)
27	(M. Li, Zhang, et al., 2015)
28	(X. Li et al., 2012)
29	(P. Li, Ouyang, Zhang, et al., 2012)
30	(Paulista et al., 2021)
31	(Murcia-López et al., 2015)
32	(P. Li et al., 2014)
33	(Luo et al., 2018)
34	(J. Wang et al., 2015)
35	(Jeyalakshmi et al., 2018)
36	(Jeyalakshmi et al., 2014)
37	(Im et al., 2017)
38	(Lee et al., 2019)
39	(Jia et al., 2022)
40	(Yang et al., 2022)
41	(Shao et al., 2018)
42	(Bi et al., 2015)
43	(Zhang et al., 2023)
44	(Kwak et al., 2017)
45	(Humayun et al., 2016)
46	(Praveen Kumar et al., 2022)
47	(Iizuka et al., 2011)
48	(Jeyalakshmi et al., 2016)
49	(He et al., 2022)
50	(Khan, Kang, et al., 2022)
51	(Khan, Charles, et al., 2022)
52	(Humayun et al., 2018)
53	(S. Wang et al., 2020)
54	(Yoshida et al., 2015)
55	(Mora-Hernandez et al., 2018)
56	(Nakanishi, Iizuka, et al., 2016)
57	(Kumar et al., 2019)

58	(Dai et al., 2015)
59	(Liang et al., 2015)
60	(X. Chen et al., 2015)
61	(M. Li, Li, et al., 2015)

APPENDIX D: COPYRIGHT LICENSES FOR FIGURES

The licenses of all adopted figures are given below. The rest of the figures and tables which are not mentioned are the original.

Number of figures	Licensed content publisher	License Number
Figure 2.1	Elsevier	5620720833319
	Springer Nature	5620730480356
Figure 2.2	Elsevier	5620721253819
Figure 2.3	Springer Nature	5620730108904