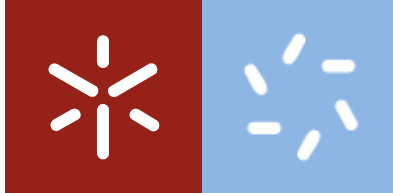


University of Minho
School of Sciences

Diogo Barros Gonçalves

Machine Learning in Analytical Chemistry: applying innovative data analysis methods using chromatographic techniques



University of Minho

School of Sciences

Diogo Barros Gonçalves

Machine Learning in Analytical Chemistry: applying innovative data analysis methods using chromatographic techniques

Msc. in Chemical Analysis and Characterisation Techniques
Chemical Sciences

Supervisors:

Professor Pier Parpot
Professor Nuno Castro

DECLARATION

Name: Diogo Barros Gonçalves

E-mail: diogobarrosgoncalves@gmail.com

Citizen card: 14812577

Supervisors:

Professor Pier Parpot

Professor Nuno Castro

MSc in Chemical Analysis and Characterisation Techniques

Year of conclusion: 2019

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.



<https://creativecommons.org/licenses/by-nc/4.0/>



University of Minho, ____/ ____/ ____

Agradecimentos

Deixo aqui expresso o meu agradecimento a um alargado leque de pessoas cujo contributo para este trabalho foi, de diversas formas, relevante.

Aos meus orientadores, Pier Parpot e Nuno Castro, pela oportunidade dada de trabalhar nesta interface, pelo apoio durante a realização deste trabalho e por me terem conferido a autonomia para explorar o tema da maneira que me pareceu mais adequada.

Gostaria de agradecer também à Fundação para a Ciência e Tecnologia através do projeto POCI-01-0145-FEDER-029147 - PTDC/FIS-PAR/29147/2017 financiado por: OE/FCT, Lisboa 2020, Compete 2020 POCI, Portugal 2020 FEDER.

Ao pessoal do LIP e do laboratório 56 pela camaradagem, em especial ao Tiago Vale pelo apoio e entusiasmo partilhado no desenvolvimento do trabalho.

Ao Rafael pelo companheirismo ao longo do nosso percurso académico, em especial pela partilha de entusiasmo pelas aplicações e potencial de *machine learning* bem como pelo suporte incondicional. À Luz, por tudo, mas especificamente pelo apoio e paciência no processo de escrita. ^^

Parecia-me algo egoísta não expressar uma palavra de apreço aos colaboradores da NVending que, ao longo do ano, foram garantindo que havia café nas máquinas.

Last but not least, à minha família e a todos os meus amigos. Em especial aos meus irmãos, aos meus pais, e à minha avó.

A todos, o meu obrigado!

Resumo

O constante avanço científico-tecnológico permitiu que, ao longo do último século, as técnicas de análise química extraíssem cada vez mais conhecimento das amostras analisadas. Nos últimos anos, a quantidade de dados que as mais recentes técnicas analíticas produzem possui uma dimensão tão elevada que a sua análise é denominada de análise megavariacional. Recentemente, a aplicação de ferramentas de *machine learning* em análises de dados químicos tem permitido extrair informação relevante das amostras analisadas que até recentemente não era possível.

Com isto em mente, o objetivo deste trabalho consiste em classificar condições de manufatura de placas de circuito impresso tendo por base dados provenientes de análise por cromatografia líquida acoplada a espectrometria de massa com extração sólido-líquido. Desta forma, esta dissertação está dividida em duas partes: a primeira sintetiza o trabalho efetuado para garantir que o método de análise produz dados com qualidade adequada para que na segunda parte esses dados sejam usados para construir modelos preditivos. Paralelamente, foi desenvolvida uma técnica de aumento de dados que, até onde o nosso conhecimento vai, constitui a primeira técnica de aumento de dados desenvolvida para problemas de classificação com dados provenientes de análises cromatográficas.

Os resultados dos melhores modelos mostram precisões superiores a 94% para a previsão de todas as condições de manufatura. Adicionalmente, a técnica de aumento de dados desenvolvida mostra desempenhos superiores comparativamente a outras técnicas de aumento de dados.

Em síntese, os resultados obtidos indicam que, para além de distinguir classes com composições químicas diferentes, é possível adquirir informação sobre quais são os compostos químicos que distinguem as classes em estudo. Esta informação pode vir a ter uma importância significativa em áreas como controlo de qualidade, química alimentar e indústria fito-farmacêutica.

PALAVRAS-CHAVE: controlo de qualidade, cromatografia, *machine learning*, quimiometria.

Abstract

Scientific and technological advances allowed the extraction of a growing quantity of knowledge from the analysed samples by means of analytical techniques. Over the last few years, the dimensionality of data that the most recent analytical techniques produce is so high, that its analysis is now called megavariable analysis. Recently, the usage of machine learning tools in chemical data analysis have allowed the extraction of relevant information from samples at a level which, until then, would just not be possible.

The objective of this work consists in classifying manufacturing conditions of printed circuit boards based on data acquired by SLE-HPLC-ESI-MS. As such, this dissertation is divided in two parts: the first synthesizes the work taken to assure the analytical method produces data with adequate quality in such a way the second part shows the development of predictive model using the previous acquired data. At the same time, a data augmentation technique which, to the best of our knowledge, constitutes the first time a data augmentation technique for classification problems using chromatographic data, has been developed.

Best models' results show precisions above 94% for all manufacturing conditions prediction. Moreover, the developed data augmentation technique reports superior performances when compared to three other data augmentation techniques.

In summary, the results show that, besides distinguishing classes with different chemical compositions, it is possible to obtain information about which are the chemical compounds that differentiate the classes. This information might be of significant importance for areas such as quality control, food chemistry, botany and pharmaceutical industry.

KEYWORDS: chemometrics, chromatography, food chemistry, machine learning, quality control.

Contents

| | |
|---|------|
| Agradecimientos | iii |
| Resumo | iv |
| Abstract | v |
| Contents | vi |
| List of abbreviations, acronyms and symbols | viii |
| Chapter 1 – Introduction | 1 |
| 1.1. Motivation for this Master Thesis | 1 |
| 1.2. Document structure | 2 |
| Chapter 2 – Background | 4 |
| 2.1. From problem definition to data interpretation: stages of the Analytical Chemistry Process | 4 |
| 2.1.1. Problem Definition and Method Selection | 4 |
| 2.1.2. Sampling | 5 |
| 2.1.3. Sample Pre-treatment | 6 |
| 2.1.4. Chemical Analysis | 8 |
| 2.1.5. Data Interpretation | 10 |
| 2.2. Towards the fully exploitation of Chemical Data | 12 |
| 2.2.1. Brief overview on the history of Machine Learning | 13 |
| 2.2.2. Learning paradigms in Machine Learning | 14 |
| 2.2.3. Machine Learning Workflow | 16 |
| 2.2.4. Introduction to Learning Algorithms | 19 |
| 2.3. Bridging Analytical Chemistry and Machine Learning | 27 |
| 2.3.1. Motivation and applications of Machine Learning in Analytical Chemistry | 28 |
| Chapter 3 – Experimental | 30 |
| 3.1. Implemented Methodology | 30 |
| 3.2. Samples | 30 |

| | |
|---|-----------|
| 3.3. Reagents | 32 |
| 3.4. Preparation of solutions..... | 32 |
| 3.5. Sample Preparation | 32 |
| 3.6. Instrumentation..... | 33 |
| 3.7. Software and hardware | 33 |
| 3.8. Data mining methodologies | 34 |
| Chapter 4 – Results and Discussion..... | 35 |
| 4.1. Overview | 35 |
| 4.2. Optimization of the analytical method | 36 |
| 4.2.1. Chemical Analysis – HPLC-MS | 36 |
| 4.3. Machine Learning Model Development | 38 |
| 4.3.1. Exploratory Data Analysis | 39 |
| 4.3.2. Time approach | 43 |
| 4.3.3. Mass approach | 47 |
| 4.4. Structure elucidation from feature importance using ESI-MS/MS..... | 52 |
| 4.5. Data augmentation technique..... | 53 |
| Chapter 5 – Conclusions and future work | 59 |
| Bibliography | 62 |

List of abbreviations, acronyms and symbols

| | |
|---------------|--|
| AAS | Atomic absorption spectroscopy |
| ADASYN | Adaptative synthetic sampling |
| AES | Atomic emission spectroscopy |
| AI | Artificial intelligence |
| Avg | Average |
| CE | Capillary electrophoresis |
| DA | Synthetic data |
| DLS | Dynamic light scattering |
| USD | United States dollar |
| DNN | Deep neural network |
| DoE | Design of experimental |
| DSC | Differential scanning calorimetry |
| DT | Decision tree |
| ESI | Electrospray ionisation |
| FIA | Flow injection analysis |
| FTIR | Fourier-transform infrared spectroscopy |
| GBM | Gradient boosting machines |
| GC | Gas chromatography |
| GOF AI | Good old-fashioned AI |
| GSR | Gunshot residue |
| HPLC | High performance liquid chromatography |
| HSSE | Headspace solid extraction |
| IC | Ion chromatography |
| IPA | Isopropyl alcohol |
| IPC | Association connecting electronic industries |
| LLE | Liquid-liquid extraction |
| LR | Logistic regression |
| Max | Maximum |
| MC | Manufacturing condition |
| Min | Minimum |

| | |
|---------------|---|
| ML | Machine learning |
| MS | Mass spectrometry |
| MVA | Multivariate analysis |
| NASA | National Aeronautics and Space Administration |
| NMR | Nuclear magnetic resonance |
| OD | Original data |
| PC | Principal component |
| PCA | Principal component analysis |
| PCB | Printed circuit board |
| PTFE | Polytetrafluoroethylene |
| QPPR | Quantitative pattern-pattern relationship |
| Q-ToF | Quadrupole time-of-flight |
| QqQ | Triple quadrupole |
| QSAR | Quantitative structure-activity relationship |
| QSRR | Quantitative structure-retention relationship |
| RF | Random forest |
| ROS | Random over-sampling |
| SBSE | Stir bar sorptive extraction |
| SLE | Solid-liquid extraction |
| SMOTE | Synthetic minority over-sampling technique |
| SPE | Solid-phase extraction |
| SPME | Solid-phase microextraction |
| SS | Standard scaler |
| SVM | Support vector machines |
| t-SNE | t-distributed stochastic neighbour embedding |
| TGA | Thermal gravimetric analysis |
| TIC | Total ion current |
| UHPLC | Ultra high performance liquid chromatography |
| UV-Vis | Ultraviolet-visible |
| XGB | Extreme gradient boosting |
| XRS | X-ray spectroscopy |

Chapter 1 – Introduction

1.1. Motivation for this Master Thesis

Analytical chemistry is one of the several chemistry fields with deep implications across almost all branches of science and even more importantly, our society. As such, the working methodology is well-defined where five main steps can be identified as defined by Elving back in 1950¹.

Every work starts with the problem definition and method selection. It is supposed to define the motivation of the work, how it will be tackled and by what means. Typically, a vast search across reliable sources - such as peer-reviewed journals and specialty bibliography - comprises its core.

After a proper definition on how the challenge will be tackled, the next two steps concern the sampling process and sample pre-treatment. They are considered crucial to the analytical process since the results obtained from chemical analyses will be as good as the sampling and sample pre-treatment performance. These steps are often tedious and time-consuming requiring high consistency across all samples and days which makes automation a tempting solution massively applied both by industry and academia.

The next stage is the core of the analytical process in which the data necessary to answer the previously stated question is acquired. With the evolution of analytical techniques and equipment, the labour part traditionally done by analytical chemists is being rapidly and steadily replaced by automation.

The last step of the analytical process - data interpretation - is the key point of the analytical process. It is often performed by a skilled operator where a series of complex approaches are executed to achieve the so wanted answer to the stated question. In contrast with the other stages, data Interpretation is falling behind in terms of automation since it requires an intuitive intelligent approach compared to the *easy* programmable tasks which is somehow related with both sampling/sample preparation and chemical analyses procedures. Recently, machine learning methods have been

successfully applied in several science fields, allowing both automation and - even more importantly - the discovery of underlying patterns in data that could hardly be found by other means.

In this context, the aim of this master's thesis consists in combining machine learning methods with chemical analyses focusing in two points:

- Development of models which are able to classify samples according to different relevant parameters (i.e., manufacturing conditions).
- Exploitation of model decision in order to extract new knowledge from sample nature.

1.2. Document structure

The developed work presented in this dissertation focused in the application of machine learning in analytical chemistry. More precisely, the development of machine learning models to classify samples according to chemical data. The first part of the project consists of generating chemical data by selecting, extracting and analysing samples whilst the second relates to all the data mining work carried on to extract knowledge out of the chemical data.

The project work is presented in five chapters. The first serves as guide for orientation purposes to the whole the document.

The aim of Chapter 2 is to bridge analytical chemistry and machine leaning by approaching both analytical chemists to machine learning as well as machine learning researchers to analytical chemistry. This way, the first part focus on a chronological introduction of the methodology carried on by analytical chemists whereas the second aims to introduce basic notions of machine learning for analytical chemists. Chapter 2 is concluded with a brief overview of recent works on the interface between these two areas with a special focus on chromatography techniques.

Chapter 3 presents the technical descriptions of the developed work, both regarding analytical chemistry and data mining.

Chapter 4 reports the discussion of results obtained during work development while Chapter 5 includes the major conclusions drawn as well as suggestions for future work regarding this interface between machine learning and analytical chemistry.

Chapter 2 – Background

2.1. From problem definition to data interpretation: stages of the Analytical Chemistry Process

Analytical chemistry can be defined as the study of substances in a matter of separation, identification and quantification². It has a crucial role across different areas such as environment³⁻⁷ (e.g. analysis of environmental microplastics), medicine⁸⁻¹⁴ (e.g. clinical diagnosis), agriculture¹⁵⁻²¹ (e.g. pesticide analysis) and even some more broad areas such as biology²²⁻²⁶ (e.g. microbiological analysis of food) and geology²⁷⁻³² (e.g. mineral inorganic content). Even though these application areas developed internal analytical processes aiming to increase their work performance according to domain specificities, a common work pipeline can be identified where five main steps are stated: problem definition and method selection, sampling, sample pre-treatment, chemical analysis and data interpretation¹. Along this section, the first four subsections will be briefly discussed with a special focus on data interpretation.

2.1.1. Problem Definition and Method Selection

Intuitively, every work starts with the definition of what question will be subject of study. The goal of this initial stage is to translate broad, domain-free, general questions into well-defined, specific questions whose answers can be achieved using chemical measurements (i.e. “how can printed circuit boards’ (PCBs) manufacturing conditions be related with its chemical composition?” should be translated to something like “how can PCBs’ chemical composition be analysed?”). What type of sample has to be collected/analysed? What kind of sample preparation has to be performed? Which analytical techniques/setups are most suitable for this end?

After proper definition, the operator is taken into domain-specific questions regarding the method selection. What is the budget and time available to achieve desired results? How will samples be collected? Which sample preparation technique should be applied in order to fulfil the pre-stated needs?

What performance requirements threshold should be guaranteed for the analysis (specificity, selectivity, accuracy, precision, etc.)? How will the acquired data address the ultimate question (keep this one in mind!)?

For this end, the operator usually combines experience with published works in peer-reviewed journals and specific bibliography since all of these questions must be well-stated before diving deep into the lab work.

2.1.2. Sampling

'Sample' can be defined in several different ways according to the work stage the analyst is referring to, which led to define 'sample' according to the stage the analyst is mentioning³³. Due to its ubiquitous mentions across all stages - when considering sampling - a more specific sample definition can be used as "a portion of material selected from a larger quantity of material"^{33,34}.

The objective of sampling consists in obtaining a small, representative and homogenous sample.

A schematic sampling process is depicted in **Figure 1**³⁵.



Fig. 1 – Sampling pipeline represented as a flow chart: from lot to aliquot.

The sampling process starts with getting a bulk sample from a lot. A lot represents the total amount of material you have access to regarding your study object (e.g. several PCBs manufactured under different conditions). A bulk sample is still a large sample that it is taken from a lot (e.g. get an adequate number of PCBs produce under the same conditions). This bulk sample must be representative of the lot, i.e.,

must gather chemical properties which illustrate the typical behaviour observed in the lot. The laboratory sample is obtained after the bulk sample has been properly prepared (e.g. cut PCBs in halves and shuffle samples produced under the same manufacturing conditions). The extension/number of stages the sample is exposed during sampling must be kept minimal in order to minimize the sampling error³⁶.

Then, the aliquot is achieved once a small portion is taken from the bulk sample and ready to be submitted to sample pre-treatment.

2.1.3. Sample Pre-treatment

In case the sample is not in a suitable shape to be analysed directly, a middle step between sampling and chemical analysis has to be performed. This stage can have multiple purposes like clean-up, concentration, interference elimination, speciation or extraction. The extension of this stage is mostly dependent on the sample nature, matrix, concentration level of the chemical compounds which are going to be evaluated during the analysis and the employed analytical technique³⁷. This information can be summarized as in **Table 1** where it shows that the pre-treatments a sample is submitted to can be related to the analyte nature.

Table 1 - Sample pre-treatment according to different analytes¹⁶⁸.

| Analyte | Sample pre-treatment |
|---------------------------------|---|
| Organics | Extraction, concentration, clean-up, derivatization |
| Volatile organics | Transfer to vapor phase, concentration |
| Metals | Extraction, concentration, speciation |
| Metals | Extraction, concentration, speciation |
| Ions | Extraction, concentration, derivatization |
| Amino acids, fats carbohydrates | Extraction, clean-up |
| Microstructures | Etching, reactive ion techniques, etc. |

Since it is not a matter of subject for this dissertation to discuss each of them, a brief overview of sample pre-treatment concerning organic/organic volatiles will be carried on with a special focus on solid-liquid extraction (SLE).

There are four widely used techniques for extraction of organic/organic volatile compounds: solid/liquid-liquid extraction (SLE/LLE), solid-phase extraction (SPE), solid-phase microextraction (SPME) and stir bar sorptive extraction (SBSE). Every extraction technique takes advantage of chemical properties which are used to influence the distribution of the analyte between phases. These properties include hydrophobicity, solubility, vapor pressure, molecular weight and dissociation constants of acids and bases. To understand how an extraction can be optimized one must be aware of the chemical equilibrium which is undergoing in the system as



and equilibrium constant,

$$K_D = \frac{[X]_B}{[X]_A} \quad (\text{eq. 2})$$

where **equation 1** denotes the chemical equilibrium between phase A and phase B at a given temperature and **equation 2** represents the equilibrium constant, K_D , where $[X]$ represents the concentration of X at a given temperature. The extraction conditions must be defined in order to increase the analyte concentration in phase B, i. e., to maximize the equilibrium constant.

In SLE, the objective is to extract as much analyte as possible from phase A (solid) to phase B (liquid) using a limited amount of solvent aiming to obtain an extract as concentrated as possible³⁸. In most cases a single-stage extraction is not enough to fulfil the desired specifications and a multi-stage extraction is required. They differ in the number of times phase A is submitted to fresh solvent (phase B). Different modifications can be applied in order to steer different chemical properties such as solubility and vapor pressure³⁹⁻⁴¹. Although more recent techniques such as SBSE or headspace solid extraction

(HSSE) are preferred, recent developments in SLE techniques show great efficiency improvements and greener alternatives when compared with state-of-the-art techniques⁴².

Apart from the employed methodologies, by the end of the pre-treatment, samples must be in a suitable form to maximize the efficiency of the essential step of the analytical chemistry process: the chemical analysis.

2.1.4. Chemical Analysis

*“Chemical analysis began on the 8th day. Adam, recovering after cooperating with god, in creating Eve, felt first pangs of hunger. He went around and harvested different kinds of colorful berries [eyes as detector] and set down for dinner with Eve. Eve rejected some berries due to foul smell (nose as detector). The bitter tasting ones were rejected next (taste as detector), and the delicious ones were consumed. Thus, first chemical detectors were nose, tongue, and eye; the five senses were used as chemical detectors for a long time.”*⁴³

Albeit not even close to the objective truth science looks after, this excerpt exceptionally captures the inquisitive nature of human beings and how it is used to understand the world. In fact, humanity has been using their five senses as detectors for a long time. However, it was not until Dutch scientists discovered how to attach two lenses in line with one another to improve their visual ability that modern analytical science came to be^{44,45}.

Chemical analysis consists in the determination of the chemical composition of substances. “In other words, it is the art and science of determining what matter is it and how much of it exists”⁴⁶. It can be divided in two branches: qualitative and quantitative chemical analysis. Qualitative chemical analysis studies what matter is it, whilst quantitative chemical analysis is responsible for answering how much of it there is. Additionally, when considering the employed technology two more subdivisions come along:

classical, wet chemical methods and modern, instrumental methods⁴⁷. These differ mainly on the used technology and subsequently the magnitude of results one can accomplish⁴⁸. Most classical analytical methods rely on chemical reactions to obtain results (e.g. acid-base titration) whereas modern analytical methods typically measure a certain physical property of the analyte (e.g. UV–Vis spectrophotometry). Obviously, modern methods yield better results (higher sensitivity, specificity, precision, accuracy, low time, ease-of-use, etc.) but they also carry some drawbacks (high cost, higher uncertainties, *black box syndrome*, etc.) when compared to classical methods^{49,50}.

Modern methods comprise a broad range of different methodologies to quantitatively address the pre-stated question. The vast different analytical techniques can be subdivided according to the measured physical property which in turn can be divided in five different families as shown in **Table 2**². It summarizes a wide range of different analytical techniques usually applied in this stage.

Table 2 - Different analytical techniques subdivided by chemistry branches

| Branch | Analytical techniques |
|---------------------------|--|
| Atomic Spectroscopy | AAS, AES, XRS, etc. |
| Molecular Spectroscopy | UV-Vis AS, IR, NMR, MS, etc. |
| Electroanalytical methods | Potentiometry, Coulometry, Voltammetry |
| Separation methods | GC, LC, IC, CE, etc. |
| Miscellaneous methods | TGA, DSC, FIA, DLS, etc. |

These techniques are often combined in order to maximize the chemical information the analyst can get. From all of them, separation methods are one of the most developed branches. Back in 2013, its value market was \$7 billion USD with a prospection for 2018 of \$10 billion USD⁵¹. Within this industry, liquid chromatography (LC) represents the large segment due to its massive use in areas such as pharmaceutical, biotechnology and food chemistry⁵².

Chromatography came a long way since 1903 when a Russian botanist named Mikhail Tsvet discussed his recent research on leaf pigments and a novel way to separate them^{53,54}. Although it was generally well accepted by the community (with even some scientists referring the crucial role Tsvet's research had in the work of Nobel Prize laureates from that epoch⁵⁵) it was not until the Second World War, the Manhattan Project and the urge to find a way to purify rare-earth metals that chromatography gained its momentum, starting with ion chromatography⁵⁶. Gas-liquid chromatography (GC) was developed faster than high pressure liquid chromatography (HPLC) where the first paper by James and Martin on GLC was published in 1952⁵⁷. Fifteen years later the first paper describing an HPLC apparatus was published giving rise to the massive chromatography market we have today⁵⁸. Nowadays, HPLC and Ultra HPLC (UHPLC) coupled with mass spectrometry detectors (MS) have been widely applied in a huge array of works ranging from clinical to beverage industry⁵⁹⁻⁶⁴. The evolution of MS gave rise to a panoply of mass analysers where quadrupole time-of-flight (Q-ToF) and triple quadrupole (QqQ) are currently considered top choices regarding both quantification and structure identification, respectively^{65,66}. The combination of MS with more conventional detectors such as fluorescence or diode array exponentially increased the amount of data the analytical chemist can now acquire regarding its experiments which in turn increased the need to improve his/her arsenal of data interpretation tools in order to tackle these challenges.

2.1.5. Data Interpretation

The last step of the analytical chemistry process consists in understanding what the acquired chemical data allows to conclude about the problem definition. This need led to the employment of statistical tools to study how chemical data relates with the pre-defined question. The first paper describing the use of multivariate regression methods and design of experiment (DoE), in analytical chemistry goes back to 1949 with Mendel⁶⁷. As a consequence of the rise of chemical data's high-

dimensionality associated with areas as spectrophotometric analysis or proteomic, multivariate analysis (MVA) started to be applied in the sixties⁶⁸⁻⁷¹. In its core, MVA consists in simultaneously analysing many variables in order to understand how these variables correlate with each other⁷². **Table 3** shows how data dimensionality relates with the statistical type.

Table 3 - Relationship between dimensionality and statistics type.

| Dimension | Sample set | Statistics |
|------------------|-------------------|-------------------|
| 1-D | Vector | Univariate |
| 2-D | Matrix | Bivariate |
| n-D* | n-D array | Multivariate |

* $n \geq 3$

MVA allowed interesting breakthroughs back then due to its ability to analyse high dimensionality data (a difficult task for humans when $n > 3$) and to allow the analytical chemist to get insights from that.

Although great achievements were performed with MVA's application in data interpretation, its application were mostly based on multivariate regression methods, response surface's and pattern recognition. With the ever-increasing amount of data acquired with novel technology and the rise of artificial intelligence (AI) and machine learning (ML) methods powered by works as Samuel's AI system which was capable to learn how to play checkers, it was a matter of time until ML methods started to be employed in analytical chemistry^{73,74}. Samuel also defined ML as "the field of study that gives computers the ability to learn without being explicitly programmed"⁷³. In fact, an important catalyst in bridging ML and chemistry were NASA's moon missions and their need for organic chemists to develop AI systems for structure elucidation^{70,75}. By the eighties, chemometrics took off as a research field of analytical chemistry with its early applications in high-dimensionality areas such as LC and spectrophotometry⁷⁶⁻⁸¹. "Chemometrics" literally means performing calculations on chemical data⁸². Its wide application not only

allowed a better understanding of high-dimensional chemical data but also generated good practices in the sampling step with the common use of DoE.

In the beginning of this decade, a series of fortunate events led to another revolution concerning data interpretation: the rapid growth of the technology around graphics processing units^{83,84}, powerful cloud-computing systems⁸⁵ and a society where more data is generated in one year than in the entire history of mankind⁸⁶ substantially contributed to major breakthroughs in AI and ML. In turn, the widespread application of these tools led to the development of easy-to-use, open-source software⁸⁷ and a strong community which allowed researchers from fields other than computer science to embrace ML in their works. Analytical chemistry is no exception and the application of these novel tools permitted that the insights hidden in chemical data (often called megavariable data) acquired by several different methods could be strongly scrutinized. In what concerns the analytical chemistry process, this came to be its last big update in a long time.

2.2. Towards the fully exploitation of Chemical Data

After exploring the Analytical Chemistry Process step by step, it became clear how can one benefit from the application of ML tools in the process itself. Along this second section of the chapter, a special focus will be given to the introduction of non-technical, relevant topics of ML for analytical chemists with little to null experience in this area.

The first subsection gives a brief overview on the history of ML from Dartmouth to present days, following the presentation of the most important steps regarding ML workflow, ending with the introduction of the intuition behind ML algorithms which were used in the development of this dissertation.

2.2.1. Brief overview on the history of Machine Learning

The history of ML is deeply connected with AI. One possible definition of intelligence would be “the ability to achieve complex goals”, therefore, AI could be defined as “non-human intelligence”⁸⁸.

It is generally accepted that the term “artificial intelligence” was officially coined in Dartmouth in 1956 by a group of scientists whose question was: “can a machine be capable of thinking?”⁸⁹. To do so, first approaches involved having programmers using their skills to handcraft a series of long and explicit rules. This is known as symbolic AI and it was the dominant paradigm of AI during many years⁹⁰. Of course, symbolic AI (or good old-fashioned AI, GOF AI) proved to be an adequate approach in logical, well-defined problems whose ruling principles are known and, for that reason, *easy* to instruct a machine to do (e.g. having a GOF AI beat the world chess champion⁹¹). However, when considering more complex and intractable problems like image recognition, speech recognition or language translation, GOF AI turned not to be a suitable approach. As a result, a new approach arose in order to surpass these obstacles. Instead of having brilliant programmers instructing a machine, they would give the machine a significant amount of data and its labels (e.g. photos of cats and dogs and its proper labels) and let the machine figure out all of those rules by itself. This approach came to be known as machine learning and has ultimately revolutionize our society.

Nowadays, humanity relies on ML for a panoply of human-level tasks in domains such as communication, healthcare, energy, finance, transportation or manufacturing, to name a few. In fact, ML is so pervasive that we are constantly being exposed to its outputs and (most of the time) not even aware of it. Such controversies gave rise to voices from both sides: some envision a world where AI has a detrimental yet well-oriented role in our society whereas others fear its consequences. A recent example of the latter is whether the ultimate goal of Neuralink of merging humans and AI by developing implantable brain-machine interfaces will be beneficial to humanity and by what means⁹².

Although these questions are, at least, decades away, the need to regulate how these systems will work should be faced in the near future in order to be primed by the time it comes.

2.2.2. Learning paradigms in Machine Learning

Most authors define three paradigms concerning the process of having a machine learning which are schematically depicted in **Figure 2**⁹³.

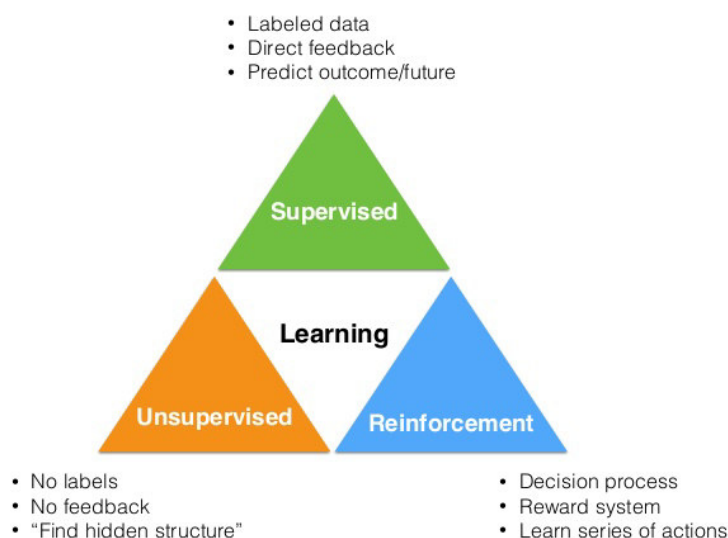


Figure 2 - The three learning paradigms: supervised, unsupervised and reinforcement learning. (adapted from ref 93)

Supervised and unsupervised learning mainly differ in whether labels are given to the algorithm or not. Supervised learning's tasks come down to classification (e.g. predict if a given sample is forgery/contaminated – discrete output) and regression (e.g. predict the concentration of a solution based on an analytical technique signal's response – continuous output)⁹⁴. In unsupervised learning common tasks involve clustering (e.g. group samples according to chemical composition's similarity) or dimensionality reduction (e.g. using less descriptors to explain how data relates)⁹⁴. **Figure 3**⁹⁵ illustrates how learning algorithms (also called learners) from these two paradigms perform.

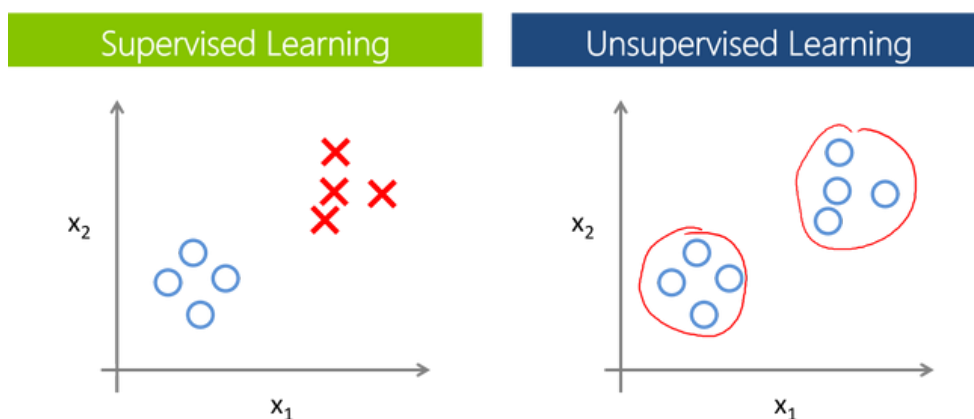


Figure 3 – Main differences between supervised and unsupervised learning. (adapted from ref 95)

While supervised learners output a prediction, unsupervised learners aggregate samples according to a pre-specified metric. In chemometrics, the latter have been used for a long time to find hidden structures related to chemical data⁹⁶. Currently, unsupervised learning is widely used in exploratory data analysis precisely due to its ability to aggregate samples into clusters that are somehow chemically related, which in turn allows the analyst to understand unknown patterns regarding the analysed samples⁹⁷. Other applications consist in data dimensionality reduction which decreases the number of descriptors needed to describe how samples relate with its variables⁹⁸. Synergies between these two paradigms integrate the typical ML workflow in several fields, including analytical chemistry.

Despite these differences, the aforementioned paradigms have a specificity in common: the output they produce is based on their input data, i. e., they both learn from previous knowledge.

This constraint is not applied in reinforcement learning where the agent learns by experience as shown in **Figure 4**.

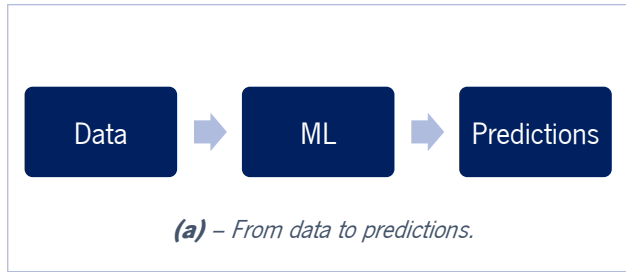


Figure 4 - Reinforcement learning process.

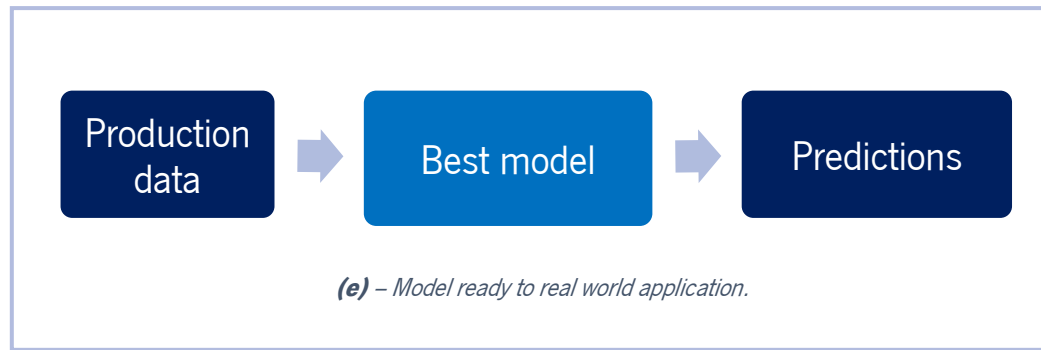
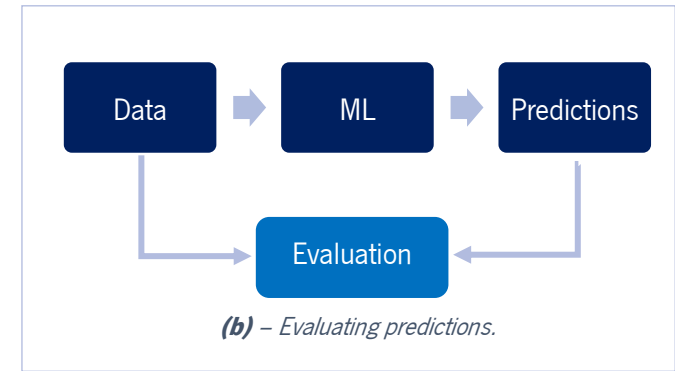
Although reinforcement learning is mostly applied in robotics^{99,100}, text mining^{101,102} and healthcare^{103,104}, several science fields have embraced its application and conducted interesting studies. In organic chemistry, a recent study showed that it is possible to optimize the experimental conditions of chemical reactions by applying this methodology¹⁰⁵.

2.2.3. Machine Learning Workflow

To understand how ML can be a valuable resource in analytical chemistry, it is important to comprehend the basics behind a ML workflow. To do so, this subsection introduces important steps that comprise a supervised learning workflow where **Figure 5** shows a simplistic schematic representation of it. Some steps (e.g. data preparation, data splitting, feature engineering, hyperparameter optimization evaluation, etc.) which would require further explanations were removed for interpretation purposes.



(a) to (b)

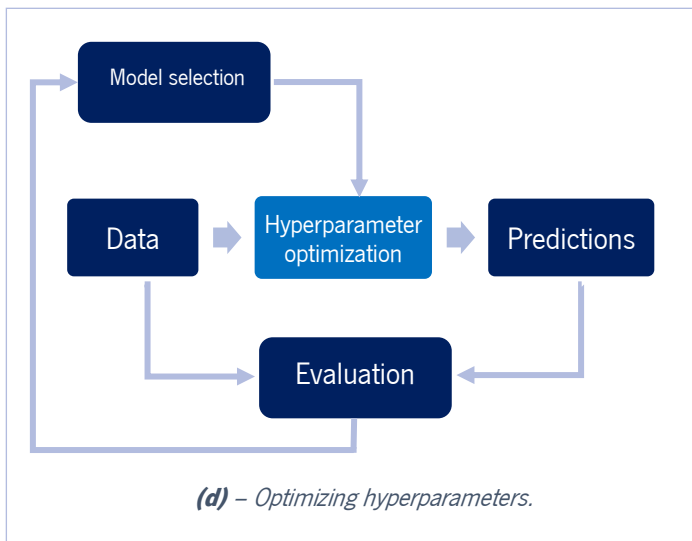


(d) to (e)

ready to deploy!



(b) to (c)



(c) to (d)

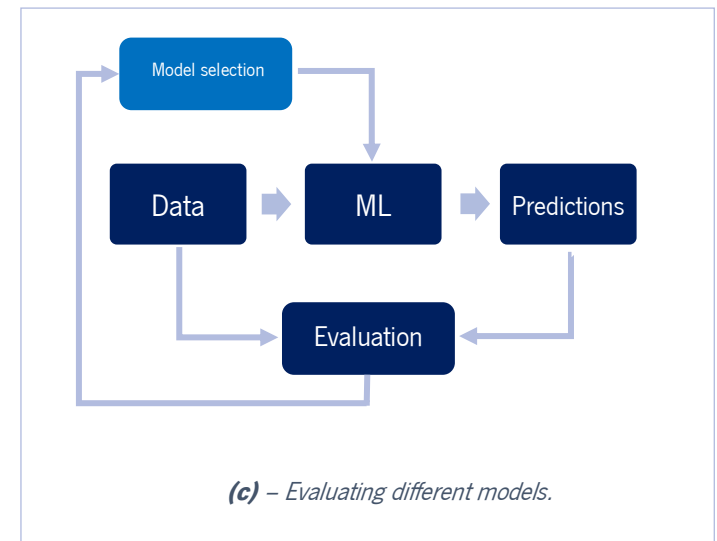


Figure 5 – Schematic representation of the principal steps regarding a supervised learning workflow.

ML model development simplistically consists in feeding the learner an adequate amount of data so it can learn the function which maps how data relates and, ultimately, provide as output an accurate prediction – **5(a)**. The accuracy of the predictions must also be evaluated – **5(b)**. To do so, it is important to use a metric which concisely evaluates the model. Classification tasks typically use confusion matrixes, precision-recall, accuracy and area under the curve metrics whilst root mean square error and root mean absolute error are preferred for regression tasks¹⁰⁶⁻¹⁰⁸. Once a metric is applied model performance can be assessed.

The next step is to apply the same methodology to several algorithms – some of them will be introduced in the next subsection – and measure each performance – **5(c)**. It is important to test different algorithms since distinctive learners will perform better depending on the input data it receives (no free lunch theorem)¹⁰⁹.

Every algorithm has parameters which cannot be learnt by itself and therefore must be defined before training (the process in which the algorithm is learning the mapping function, i. e., where it figures out the rules needed to explain how data is related). These are called hyperparameters and fine tuning them allows the algorithm to more easily capture data patterns, thus, increasing its performance – **5(d)**⁹⁴.

Once the best model (function which best describes how the label is related to its input variables) is obtained and properly evaluated the model is ready to be applied with new data.

Despite the simplistic representation, **Figure 5** allows a quick overview on how most ML models are built.

Another way to increase model performance consists in feeding the learner with more high-quality data. By having more data to train on, the applied learner will be able to capture complex data trends. This can be done by acquiring more data however, sometimes, this is not possible. In these cases, the analyst must resort to synthetic ways to expand the size of his(er) dataset, often called data augmentation techniques. Under an analytical chemist perspective this must be perceived as a means

of increasing the number of events (analysed samples) without further experiments (chemical analysis). Its application in ML research, is responsible for interesting achievements such as top performances in ML competitions. In computer vision (e.g. classifying pictures of dogs and cats) it can be done by applying rotations of the original pictures, changing colours, mirror effects, etc. By doing this, the analyst is feeding the learner with more data so it can find the best model. In chemistry, data augmentation techniques are now applied by adding drifts to the original data^{110,111}. This approach has been successfully applied with NIR and molecular descriptors using public datasets.

2.2.4. Introduction to Learning Algorithms

The last subsection introduced how a supervised ML model can be built. This subsection will introduce the intuition behind the ML algorithms which were used during the elaboration of this dissertation. The intention here is not to present the reader with all the algorithm's mathematical formalism but rather to give an intuition on how they perform the task. If interested in the mathematics behind it, please refer to this book¹¹².

To properly introduce ML algorithms, it is important to acknowledge what it is first. In practice, an algorithm is a step-by-step way to solve a problem. As previously stated, ML algorithms are usually called learners. In contrast with common algorithms where a list of rules to follow is instructed to a machine, the conception of learners allow them to infer those rules by analysing a considerable amount of data. The plethora of academic and corporate research gave rise to a large variety of learners as stated in **Figure 6**¹¹³.



Figure 6 - Machine learning algorithms representation. (adapted from ref 113)

This large representation intends to group learners according to operational similarity. Different families will typically perform better with distinctive datasets hence the popular no free lunch theorem applied in this domain. From this large number of learners, six of them will be covered here, particularly: principal component analysis, logistic regression, decision tree, random forest, gradient boosting machine and support vector machines.

Principal Component Analysis

Principal component analysis (PCA) is an unsupervised learner whose goals involve reducing data's dimensionality and cluster samples according to its similarity. Widely considered the building block of chemometrics, PCA is commonly found in most chemical data analysis mainly due to its ease of interpretability which can be stated in **Figure 7**^{114,115}.

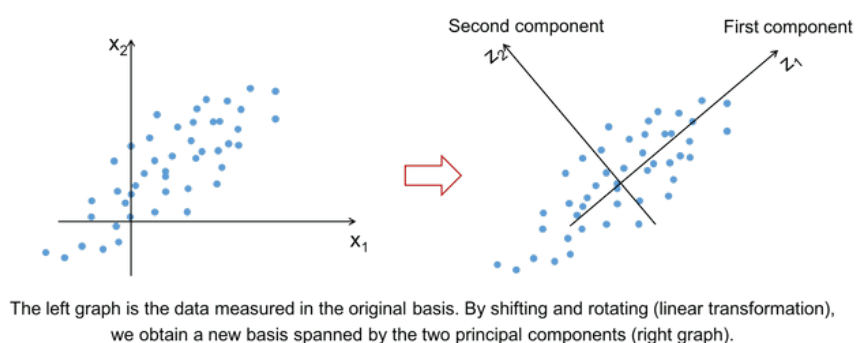


Figure 7 - PCA dataset rotation. (adapted from ref 115)

This methodology consists in the application of algebraic operations which enable dataset's rotation in such a way that the rotated features are statistically uncorrelated. Another definition says "PCA simplifies the complexity in high-dimensionality data while retaining trends and patterns"¹¹⁶.

Applications in analytical chemistry usually tend to plot the first and second/third component in order to explore how samples and features correlate with one another¹¹⁷. PCA's main limitation concerns the fact that the applied rotations are linear transformations of the original data. When looking for more complex data patterns, different algorithms capable of non-linear projections should be taken into account (e.g. t-SNE)¹¹⁸.

Logistic Regression

The first supervised learner to be introduced is logistic regression (LR). Its conception goes back to the first half of the eighteenth century¹¹⁹. For anyone who understands linear regression, LR is just an

upgrade of it. Multiple linear regression can be expressed as in **equation 3** where y denotes the dependent variable (e.g. concentration of Na^+ in a solution), β are the coefficients and X are the dependent variables (e.g. intensity of a measured signal).

$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n \quad (\text{eq. 3})$$

Linear regression enabled analytical chemists to calculate different properties for a long time. However, linear regression has two main limitations: it assumes the relationship between y and X is linear and it outputs a continuous value from $-\infty$ to $+\infty$. This second limitation becomes particularly important if instead of y being a continuous variable (like in the aforementioned problem), y is a categorical value (e.g. given a series of relevant parameters, will a reaction occur or not). This is the kind of problem where LR becomes a valuable tool. As can be seen in **Figure 8**¹²⁰ and explained by **equation 4**, LR has the advantage of outputting a value between two pre-stated values.

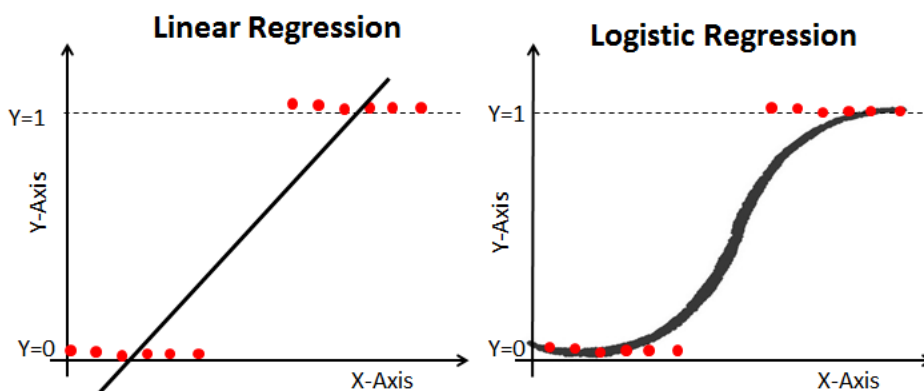


Figure 8 - Logistic regression's advantage over linear regression. (adapted from ref 120)

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}} \quad (\text{eq. 4})$$

This small improvement allowed its wide use in biological sciences back in the early twentieth century and its later application in social sciences¹¹⁹. Despite its simplicity, a recent review showed that the application of more complex learners over LR had no performance benefit concerning clinical prediction models¹²¹. Nevertheless, nature provides endless situations where the problem demands more flexible learners capable of better generalization. The next four learners will address this.

Decision Tree

A decision tree (DT) is no more than a disjunction of conjunctions. In fact, humanity applies DTs in a panoply of different problems. In order to classify rocks, high school students are given a series of rules they have to follow to reach an accurate classification. The intention is to, with each rule the student follows, increase the subset purity, i. e., decrease the number of rock possibilities', until he reaches a subset where only one rock class can fulfil all those specifications – **Figure 9**¹²².

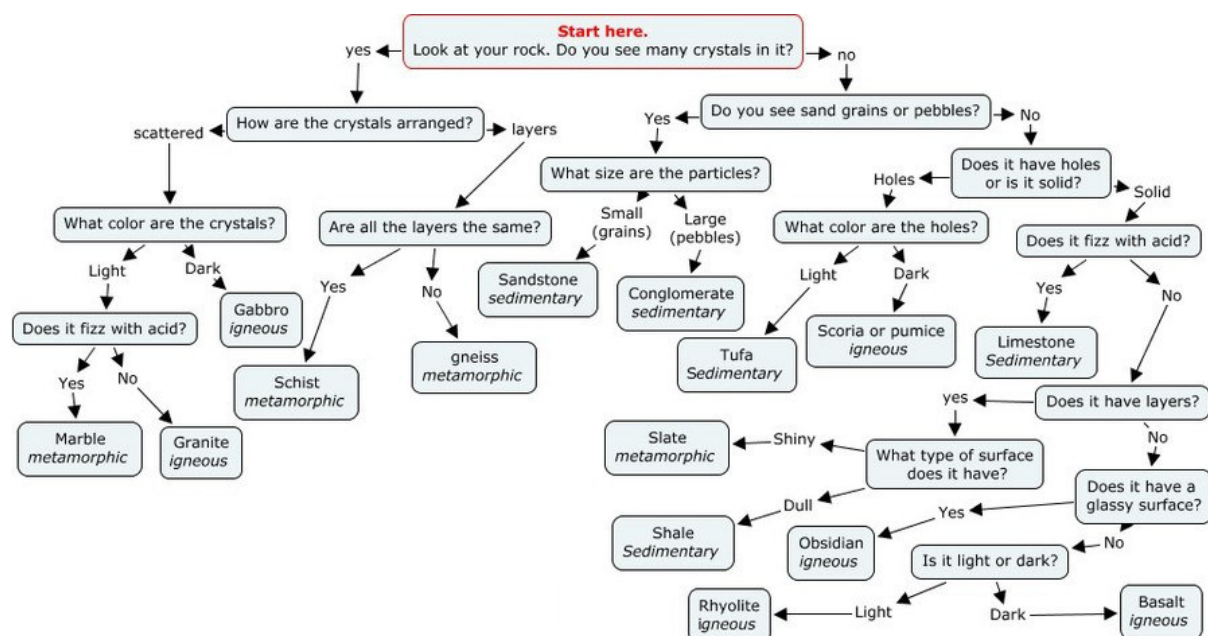


Figure 9 – Schematic representation of a decision tree criteria rules for rock classification task. (adapted from ref 122)

To do so, the student follows a set of rules concerning important features regarding the rock. Some features include its color, particle size, visible crystals, reactivity with certain substances, smell and even its taste. Different features will have distinct importance concerning rock classification. What high school teachers do, is to use their knowledge about geology to write that set of rules. In practice, what a DT learner does is to figure out all of those rules according to the input data it is given.

This same methodology is extensively applied in different areas including analytical chemistry. DT's applications not only involve classification tasks like predicting different types of wines¹²³ but also regression problems like predicting the relationship between structure-activity (QSAR) for a compound¹²⁴. One of the main advantages behind DT comes from the fact that the set of rules it infers enables the analyst to acquire more knowledge about the sample nature. Albeit it looks a normal requirement, state-of-the-art algorithms like deep neural networks (DNN) don't allow such easy intuitions hence DTs widespread use in more simplistic problems.

Random Forest and Gradient Boosting Machines

One of DT's main limitations is its ability to generalize extremely well on the training data. It typically happens when the depth of the tree and/or the number of applied splits is too large which leads to one of the trickiest obstacles in ML called overfitting. Overfitting happens when the learner exceptionally captures data trends during its training – **Figure 10**¹²⁵.

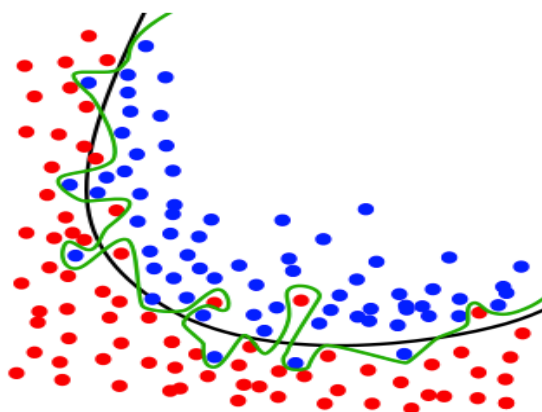


Figure 10 - Overfitting in binary classification. (adapted from ref 125)

While the black line sets an adequate decision boundary, the green line shows what happens when a model over-generalizes during its training. In real world applications, the green line model will tend to perform worse than the black line one. In a certain way, the model is suffering from “hallucinations” since it is capturing trends that don’t really exist which can be attributed to a panoply of sources of error such as noise, mislabelling, detector’s malfunctioning, among others.

In 2001, Breiman¹²⁶ introduced the concept of random forest (RF) which, as the name implies, consists of a large number of DTs that operate as an ensemble – **Figure 11**¹²⁷.

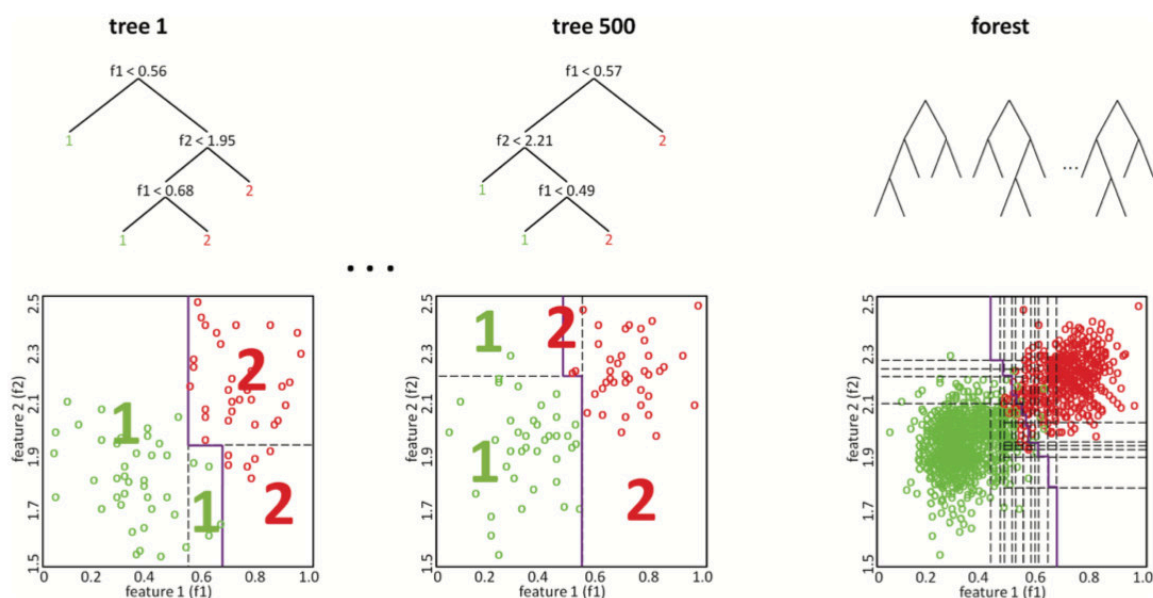


Figure 11 - Random forests are composed by individual decision trees that act together as an ensemble. (adapted from ref 127)

RFs tend to reduce model overfitting. By using a large number of uncorrelated trees operating as a committee, RFs are capable of outperforming any of the individual guesses from the committee. The learner is instructed to apply cut sections on data in order to create distinctive sectors in a hyperdimensional space. RFs are used in analytical chemistry both in classification¹²⁸⁻¹³⁰ and regression¹²⁸ problems.

Gradient boosting machines (GBM) introduces the concept of boosting, but in its core they have similarities with RFs. In fact, they also act as an ensemble. The term *boosting* is related to its major advantage over RFs due to the fact that each tree (classifier) is trained on the last tree's errors – **Figure 12**¹³¹. Datapoints which were mislabelled by the prior classifier are attributed a higher weight in the next classifier's training so the model will be more penalized when mislabelling these instances. This process is done iteratively which will eventually lead to a final model being trained in each classifier's error. GBMs have gained a lot of attention in the last few years, being responsible for a large number of winnings in Kaggle competitions¹³². In analytical chemistry, GBMs are mostly applied in classification problems regarding high-dimension data^{133–135}.

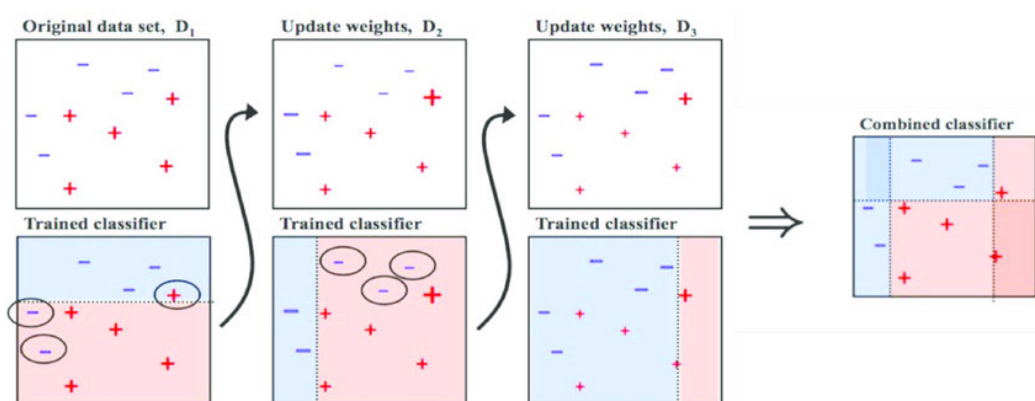


Figure 12 - Training of a gradient boosting machine. (adapted from ref 131)

Support Vector Machines

Support vector machine (SVM) is a supervised learner proposed by Vapnik in the nineties¹³⁶. Although it can be applied in both regression and classification problems it is mainly applied in the latter. In classification tasks, the intuition behind it consist in defining a decision boundary that maximizes the distance between different classes – **Figure 13**.

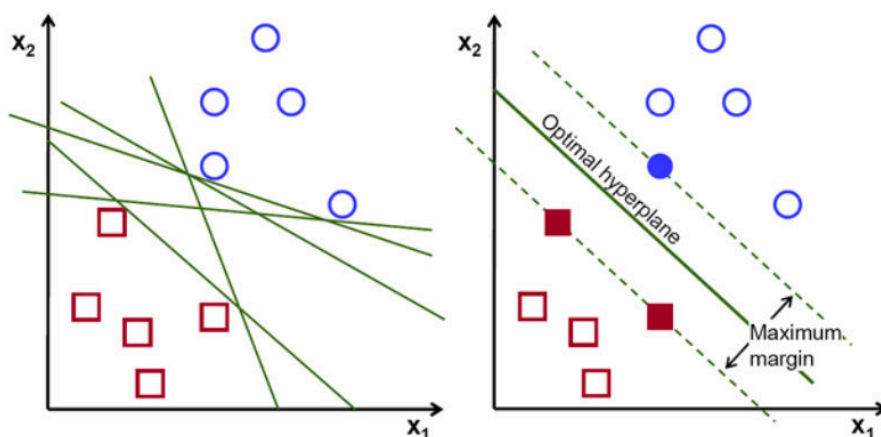


Figure 13 - Intuition behind SVM hyperplane definition. (adapted from ref 136)

Since an infinite number of hyperplanes can be defined (**Figure 13**, left) SVMs take advantage of the closest datapoints from different classes to define the optimal decision boundary by using them to define support vectors to set the optimal hyperplane (**Figure 13**, right), i.e., the hyperplane which maximizes the distance between closest datapoints from different classes. When data is linearly separated without any mislabelled datapoints it consists in a hard margin SVM, the opposite constitutes a soft margin SVM¹³⁷.

SVM became popular in the nineties due to the cheap computation cost and the good performance it implied being broadly applied in topics as QSAR and drug design. Despite the fact that they are now outperformed by more flexible models, they are still applied in analytical chemistry mainly in classification tasks^{138,139}.

2.3. Bridging Analytical Chemistry and Machine Learning

While the first section of this chapter intended to present each step of the analytical chemistry process with a special focus on data interpretation and ML's role in it, the second explored basic concepts regarding ML aiming to understand what it is and how it works. This third and last one aims to bridge

the first two. Its goal is to summarize how ML can be a valuable tool in analytical chemistry and to cover some recent applications regarding this interface.

2.3.1. Motivation and applications of Machine Learning in Analytical Chemistry

Analytical chemistry came a long way since 1860 when Bunsen and Kirchhoff developed the first flame emissive spectrometer which allowed the discovery of the alkali metals rubidium and cesium¹⁴⁰.

In the first half of the twentieth century occurred a revolution in the analytical instrumentation which widened the answers an analyst could gather from those analysis. With novel, upgraded, sophisticated techniques more and more information could be attained which required more complex ways in order to interpret what that information meant. Nowadays, modern analytical instrumentation generates so much data that it is now called megavariable data. All this chemical data brought up the need to implement new ways of examining it.

Simultaneously, social and technological advancements allowed the ordinary application of ML in several science fields including analytical chemistry. This interface where ML is applied in domains such as particle physics¹⁴¹ and biology¹⁴² had already revealed very useful enabling several breakthroughs. In chemistry, numerous areas have benefited from its application ranging from catalysis^{143,144}, drug discovery^{105,145,146} to material science^{147,148}. Some AI experts even claim chemistry should be the next grand challenge for AI¹⁴⁹. Among other things the authors argue the knowledge acquired in conventional AI studies such as two-player board games and human-mimicking tasks as nature language processing or computer vision, places AI community in a good position to tackle chemical challenges with incredible benefits for humanity. In fact, complex chemical tasks as retrosynthesis are now capable of AI automation with a performance at least as good as a skilled chemist¹⁵⁰. In this study, Segler *et. al*/used Monte Carlo tree search and symbolic AI to propose retrosynthetic routes. By training DNNs on more than 12 million single-step reactions the authors developed an AI system capable of understanding the underlying rules

of retrosynthesis in such a way that in double-blind AB test, chemists considered the AI-generated routes to be equivalent to those reported in literature.

An increasing number of works have been done in the interface between chromatography and ML. Cao *et. al*/proposed an approach called quantitative structure-retention relationship (QSRR) to predict the retention time of a compound given a chromatographic setup¹⁵¹. To attain this, the authors used a dataset of 93 molecules where molecular descriptors were used as features and its respective retention times. In contrast with Segler *et. al*/where DNNs were used, this work relies on RFs to build the predictive model.

Recently, another interesting approach called quantitative pattern-pattern relationship (QPPR) was developed to predict the effect that firing a gun has in the chemical composition of the ammunition constituents¹⁵². In forensic sciences, the association of the gunshot residue (GSR) to the person who took the shot constitutes a challenge for forensic experts. Traditional methodologies involve analysing the ammunition content, fire a gun and then analyse GSR from spent cases in order to establish a relationship between GSR and the original content. With QPPR, authors showed it was possible to relate GSR with the initial content without having to fire a gun using ML models. After testing 14 different learners, top performances were obtained with RFs and SVMs.

Considering quality control, ML has been successfully applied in egg authenticity¹⁵³, adulteration of vegetable oils^{117,154} and citrus fruits' quality¹⁵⁵.

The ever-increasing number of works in this interface strongly indicates that, in the near future, having a basic understanding on how ML can be applied in analytical chemistry will be a valuable skill every analytical chemist should have¹⁵⁶.

Chapter 3 – Experimental

3.1. Implemented Methodology

In this chapter the employed methodologies as well as materials and software used in the development of this dissertation will be described.

The purpose of this project was to study how chemical data acquired from HPLC-MS can be used to attain useful insights regarding sample nature by applying ML tools. More specifically, the goal was to develop ML models able to classify PCBs according to four manufacturing conditions (MCs) by analysing the end product using HPLC-MS. Since ML models need significant amounts of data to train on, a novel data augmentation technique was developed alongside. The used analytical method was performed according to IPC-TM-650 2.3.27.1¹⁵⁷ whose objective is to analyse the chemical composition of PCB's surface.

3.2. Samples

The selected samples for this study consisted in 180 PCBs manufactured under four distinctive conditions (A, B, C and D). There are 18 different combinations of PCB's MCs – **Figure 15**. For each of those 18 different MCs there are 10 replicate samples produced under those same conditions with the exception of A1B3C1D1 and A1B3C1D2 which have 15 and five replicate samples, respectively. These 18 different groups are named according to the MCs that were used in their production (e.g.: a class can be represented as A2B1C2D2. This means conditions A2, B1, C2 and D2 were used during PCB production). For each MC there are two different possibilities except for condition B which has three possibilities (A1/A2, B1/B2/B3, C1/C2, D1/D2).

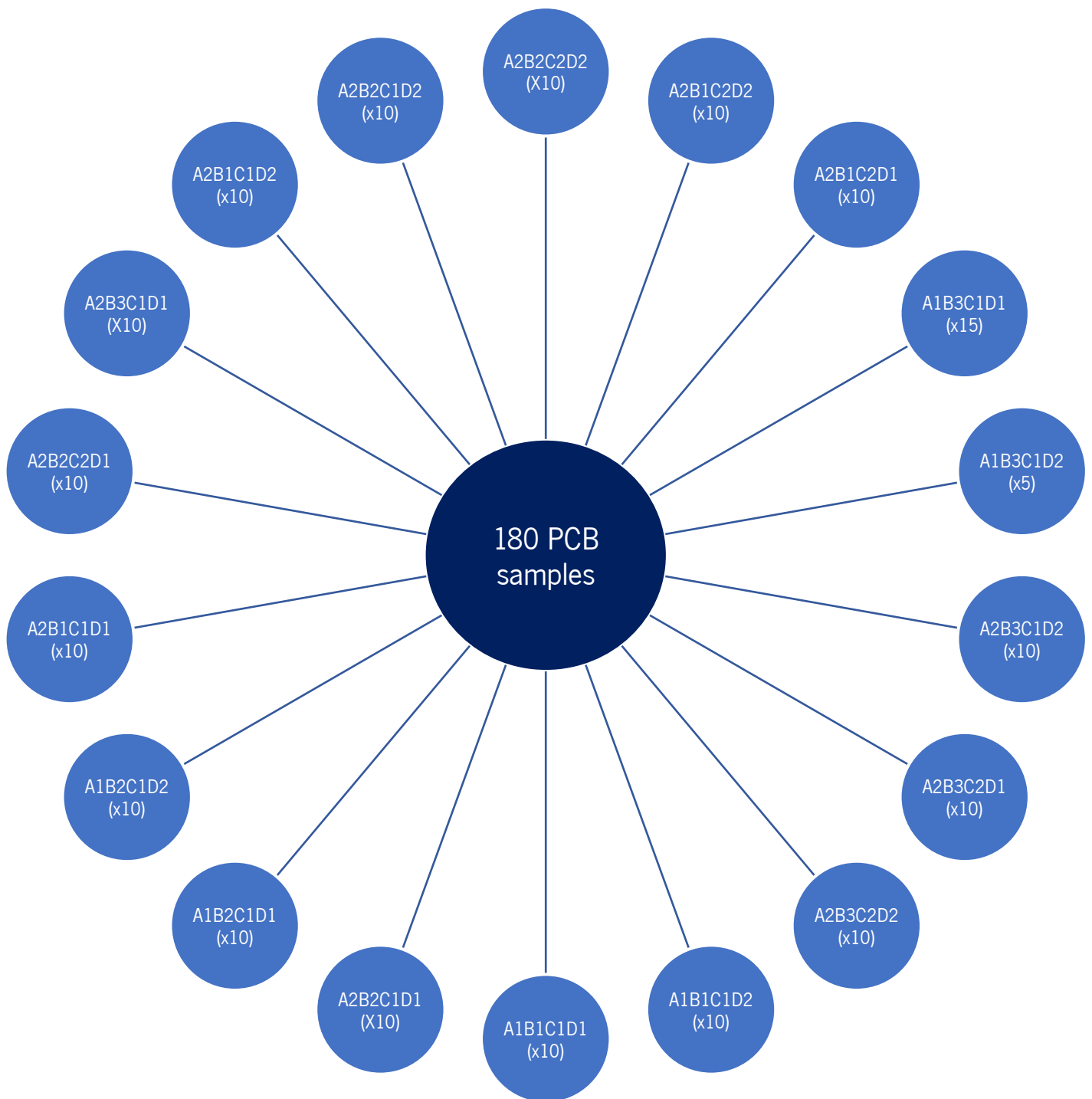


Figure 15 - Diagram of the different manufacturing conditions used in PCB production as well as the manufacturing conditions employed in the 18 different combinations.

3.3. Reagents

Acetonitrile was obtained from Fisher (Loughborough, UK); isopropyl alcohol (IPA) was purchased from Honeywell (Seelze, Germany). Both solvents were HPLC grade. Ultrapure water (18 M Ω .cm) was prepared using a Milli-Q Gradient A10 (Darmstad, Germany). Glacial acetic acid was purchased from Panreac (Barcelona, Spain).

3.4. Preparation of solutions

Extraction solution for SLE (IPA/water, 75:25 (v/v)) was prepared according to IPC-TM-650 in 1 L batches and kept in PTFE gallons in the dark at -4°C. HPLC solvents were prepared individually by adding 500 mL of acetonitrile/water, sonicated in ultrasound bath for 30 min and added 0.1% (v/v) glacial acetic acid in the water shot.

3.5. Sample Preparation

SLE was performed according to IPC-TM-650 2.3.27.1 where PCBs samples were, additionally, cut in halves using a steel blade cutting machine, placed inside a KAPPAK SEALPAK #503 (VWR, USA) extraction bag, added 60 mL of the IPA/water extraction solution, heat-sealed the bag and placed inside a water bath at 75°C for 60 min. After, the extraction bag was allowed to cool down to room temperature before opened, its extracted solution was transferred to 10 mL glass vials after filtered with a 0.2 mm PTFE filter. Extracted solutions were kept in the dark at -4°C prior to analysis.

3.6. Instrumentation

Chromatographic separation was performed on an Kinetex RP-C18 (100x4.6mm, 2.6 μ m) analytical column (Phenomenex, Torrance, CA). An Edwards E2M30 pump (Edwards, West Sussex, UK) was used for gradient elution at a constant flow of 0.3 mL/min.

HPLC solvents were: A (water, 0.1% acetic acid) and B (acetonitrile). The mobile phase was programmed as follows: original conditions 60% A, linear gradient to 10% A in 20 min, linear gradient to 60% A in 5 min. Re-equilibration time was 5 min.

Mass spectrometric measurements were performed on an LXQ (Finnigan, San Jose, CA) linear ion trap mass spectrometer equipped with an electrospray ionisation source (ESI) working in positive ion mode acquisition in a range from 50 to 1000 Da. The ESI parameters were: capillary temperature 250°C, sheath gas flow 50 arbitrary units (a.u.), auxiliary gas flow 10 a.u., sweep gas flow 10 a.u., source voltage 5 kV, source current 100 μ A, capillary voltage 10 V, tube lens 15 V, sheath gas nitrogen (Praxair, PT), auxiliary gas nitrogen (Praxair, PT).

3.7. Software and hardware

HPLC-MS files (chromatograms, MS spectra) were acquired and manipulated with the built-in software version of the equipment XCalibur Quant (version 2.7). Each analysis file is predefined exported in a RAW extension by the built-in software and converted to csv extension with a multi-group internally developed software. All data manipulation was performed with the following software: python v.3.6.8, imbalanced-learn v.0.5.0, matplotlib v.3.1.0, numpy v.1.16.4, pandas v.0.24.2, scikit-learn v.0.21.2, scipy v.1.3.0, seaborn v.0.8.1, xgboost v.0.90. Hardware specifications include: 2.3 GHz dual core Intel Core i5 CPU and 8 GB 2133 MHz LPDDR3 memory.

3.8. Data mining methodologies

Standard scaling was applied before data splitting. Training and test data were divided in 80/20 with class stratification. PCA, in the context of model development, was applied in preprocessing after scaling. The used features allowed to explain 95% of system's variance which corresponds to 11 and 133 features regarding time and mass approach, respectively. Classifiers' performances are measured by precision calculated according to **equation 5**.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (\text{eq. 5})$$

All fifteen different classifiers were submitted to a 5-fold cross validation intending to evaluate model stability to data splitting. This is done by training five different ML models with different training sets and testing them in also five different test sets in an iterative process.

Chapter 4 – Results and Discussion

4.1. Overview

The role of data interpretation in the analytical chemistry process, currently constitutes one of the greatest challenges an analyst has to face. It involves understanding novel, advanced technology and concepts which until recently weren't associated with analytical chemistry. In the past few years, ML have become such a valuable tool regarding this step that numerous works in this interface are now tackling interesting challenges in analytical chemistry. Hyphenated methods, such as LC-MS, are now capable of generating huge amounts of multi-dimensional data. This megavariate data contains much more information than the traditional data interpretation methods could explain which, as a consequence, gave rise to the application of ML tools in order to surpass this limitation.

This chapter is divided in two main parts. The first one presents results regarding laboratory work needed to guarantee the acquisition of high-quality chromatographic data whilst the second part shows results related to the development of ML models capable of predict MCs based on the previously acquired data.

Since this work supports confidential status, results are presented in a generalist, non-specific style aiming to show how the developed tool can be of interest to the analytical chemistry community rather than the problematic which was studied. Also, it is not the scope of this dissertation to study the chemistry of PCBs. Moreover, the confidential details are not relevant for the presented study and, for that reason, this chapter focus on the developed methodology.

4.2. Optimization of the analytical method

This section introduces the work carried on in order to guarantee high quality chemical data. Since the analytical conditions of the employed method were already fine-tuned for a similar problem¹⁵⁸ the intention with this first section is to guarantee that the analytical method, specifically the analytical conditions regarding separation, produce chromatograms in a suitable shape and quality to further feed ML algorithms in order to build predictive models for MCs.

4.2.1. Chemical Analysis – HPLC-MS

Efforts towards guaranteeing a suitable chemical profile capable of capture sample's chemical nature were made by testing three different analytical conditions as described in **Table 4**. Since no methodologies for analytical method validation aiming to build predictive ML models were found in literature, the idea behind these tests is to ensure that the employed separation conditions allow a suitable peak separation and that no sample carryover occurs.

Table 4 – Analytical conditions tested during chemical analysis verification.

| Condition | Analytical column | Separation conditions | Flux (mL/min) |
|------------------|---------------------------------------|--|----------------------|
| 1 | Kinetex C18 (100x4.6mm, 2.6 μm) | original conditions 60% A, linear gradient to 10% A in 20 min, linear gradient to 60% A in 5 min. Re-equilibration time was 5 min | 0.3 |
| 2 | Kinetex C18 (100x4.6mm, 2.6 μm) | original conditions 20% A, linear gradient to 10% A in 20 min, linear gradient to 20% A in 5 min. Re-equilibration time was 5 min | 0.3 |
| 3 | Luna C18 (100x2mm, 3 μm) | original conditions 60% A, linear gradient to 10% A in 20 min, linear gradient to 60% A in 5 min. Re-equilibration time was 5 min | 0.25 |

In order to qualitatively evaluate the produced chromatograms a notation of peaks correspondent to ions at m/z 280 and m/z 375 were kept in each chromatogram (**Figure 16**) with the aim of assessing the degree of peak separation that is achieved with each condition presented in **Table 4**. Condition 1 (black, top) represents the analytical condition which is the base of the analytical method whilst conditions 2 (red, middle) and 3 (green, bottom) were presented for comparison purposes.

Figure 16 shows the resulting chromatograms of the three analytical conditions tested. Condition 3 indicates the employed separation conditions have a higher elution strength which results in a myriad of compounds being eluted in the beginning of the analysis thus decreasing peak separation and, subsequently, the quality of the chemical data. This information is also supported either by the large peak intensity observed in this condition which might happen as a result of having a large number of compounds being eluted at the same time as well as by the relative position of ions m/z 280 and m/z 375.

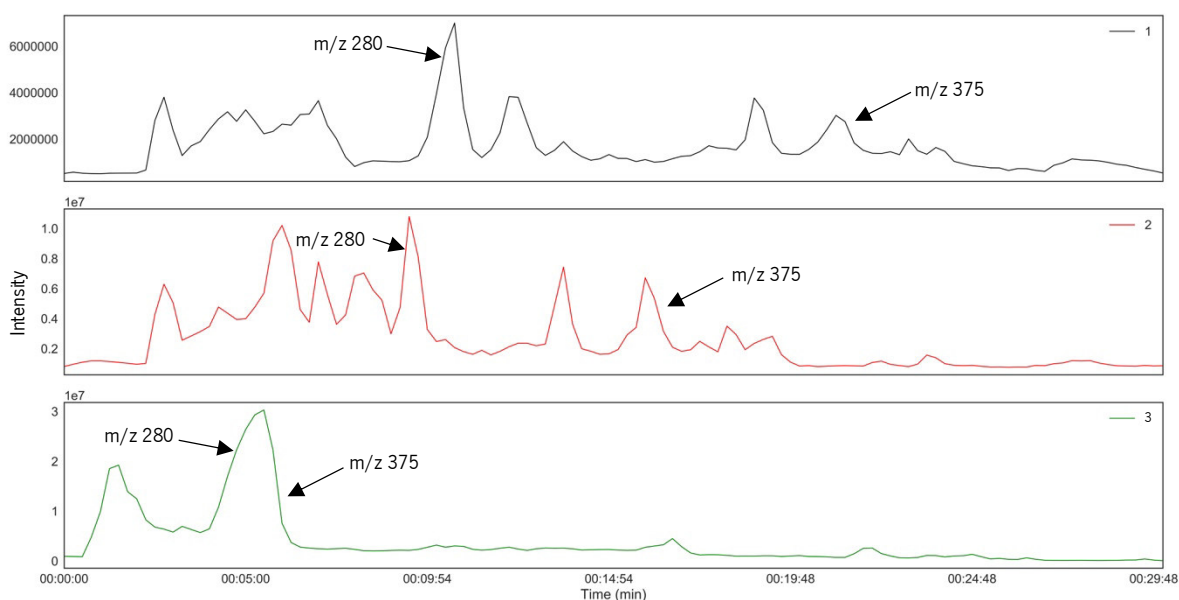


Figure 16 – Chromatograms of the tested analytical conditions. Ions at m/z 280 and m/z 375 are denoted as references for peak separation evaluation.

Condition 2 denotes a condensed chemical profile and a larger elution strength than the one required to produce an adequate peak separation. This information is also supported by the position of ion m/z 375. Although condition 2 arguably denotes the best peak separation regarding the beginning of the analysis, peak separation at mid-time analysis is slightly worst when compared with condition 1. Furthermore, due to the nature of the analytical setup, species eluted at the beginning of the analysis (solvent and unretained species) are more likely to be less important to the matter of classification when compared with mid-time analysis.

Thus, these results show the analytical conditions employed produce chromatograms with a suitable shape and quality for the desired end.

4.3. Machine Learning Model Development

In this section will be presented the obtained results regarding ML model development. The first subsection explores the obtained chemical data and introduces how the problem will be addressed by, regarding two different approaches. Second and third subsections focus on results related to each approach.

4.3.1. Exploratory Data Analysis

Each chromatographic analysis consists in a series of MS scans (ca. 12k) which can be viewed as a 30-min chromatogram. **Figure 17(a)** shows the total ion current (TIC) chromatogram of each MS spectra and **Figure 17(b)** depicts a MS spectrum related to the highest intensity peak at 18.93 min.

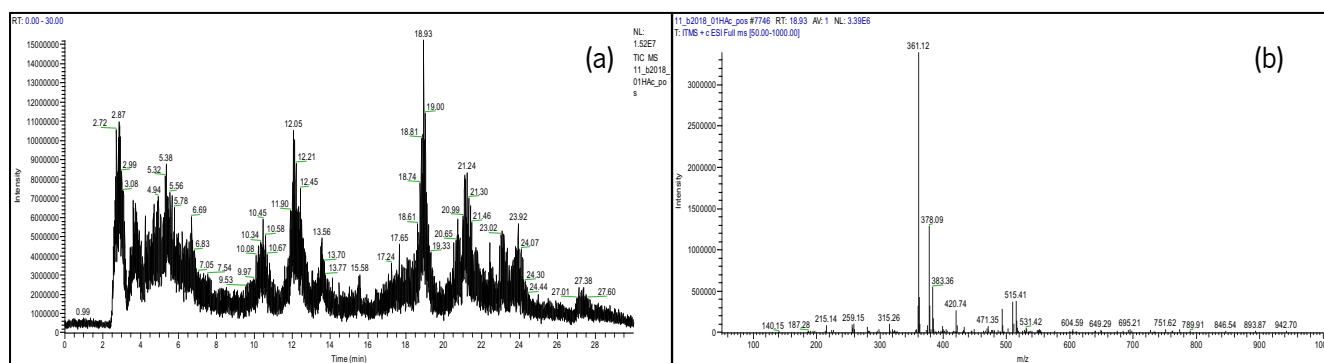


Figure 17 – TIC (a) chromatogram and (b) MS spectrum for the peak at 18.93 min, acquired with the built-in XCalibur software.

When this information is resumed in a tabular format, the same can be viewed as in **Figure 18**. Each column represents the measured intensity of ions ranging from 50 to 1000 Da and each row consists in a time series of the acquired MS spectrum.

| | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | ... | 996 | 997 | 998 | 999 | 1000 |
|---|-----|-----|-----|-----|-----|-----|-----|---------|--------|---------|-----|------------|------------|------------|------------|------------|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 977.69 | 0.00 | 0.00 | ... | 850934.67 | 853711.76 | 1031355.59 | 1309377.85 | 4309489.02 |
| 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 875.72 | 0.00 | 0.00 | ... | 1535313.92 | 1359096.09 | 1331324.44 | 1689264.54 | 1274837.79 |
| 2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.00 | ... | 1406876.91 | 911272.98 | 1224186.87 | 1101283.72 | 1200918.07 |
| 3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.00 | ... | 1037164.06 | 840181.50 | 966045.22 | 1680721.00 | 4514608.39 |
| 4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.00 | ... | 1639359.60 | 1314921.80 | 1393548.49 | 1697286.80 | 2214047.41 |
| 5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 110.71 | 482.54 | 0.00 | ... | 981941.87 | 913529.44 | 1076796.63 | 1601918.74 | 3445054.83 |
| 6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 1912.06 | ... | 1179591.10 | 1206936.95 | 1328766.36 | 1375715.29 | 3904237.78 |
| 7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.00 | ... | 1313127.33 | 1208279.00 | 1120490.58 | 1067881.16 | 1958962.92 |
| 8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.00 | ... | 1450521.64 | 1784865.61 | 1504245.52 | 1706966.98 | 1921701.21 |
| 9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2979.57 | 0.00 | 0.00 | ... | 2238485.25 | 2438093.62 | 2313123.21 | 2132014.58 | 2042966.14 |

Figure 18 - HPLC-MS data from the analysis of one PCB sample resumed in a tabular format.

Features include the ion intensity from 50 to 1000 Da in a time series where each row consists in a scan.

To develop the ML model two different approaches regarding the used features were taken. One consisted in using the sum of TIC intensities – time approach. The second consisted in view each sample as the sum of ion intensities – mass approach. The objective with creating models using these two approaches is to compare both and to evaluate which one allows better ML model performances. The following subsections will present the results regarding both approaches.

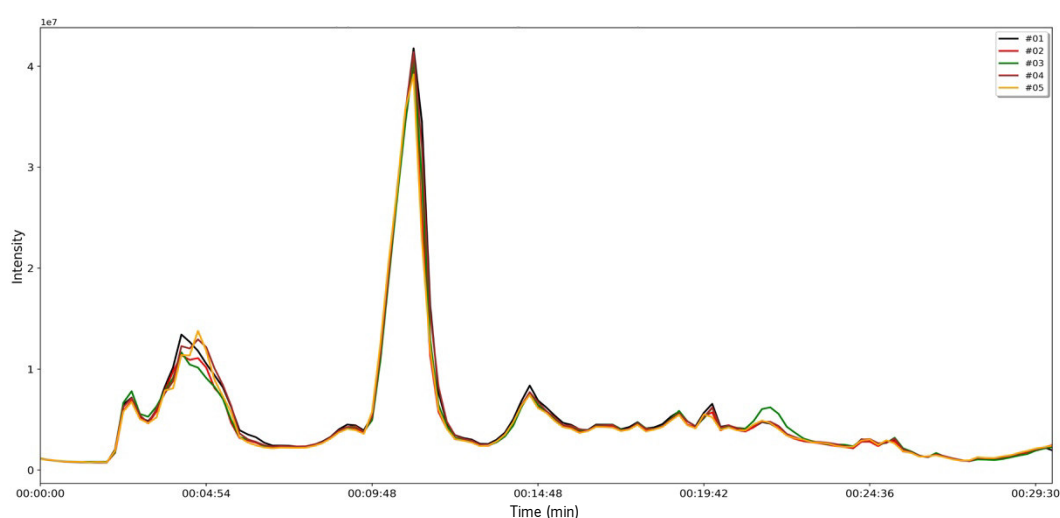


Figure 19 – Chromatograms of five replicate sample of A2B3C1D2 regarding time approach.

Considering time approach, **Figure 19** shows chromatograms of five independent replicate samples, where a similar chemical profile can be observed across all samples. Slight variations in peak shape and area are expected as described in literature¹⁵⁹. These can be related with some HPLC components (detector, column, autosampler) but the main reasons are usually pressure and autosampler variations. Furthermore, for this approach the number of scans was reduced 100 times (from 12k to ca. 120). This way, noise can be reduced to improve model performance while still keeping the chemical information.

Figure 20 depicts the 20 highest intensity ions regarding three independent replicate samples. At this point, it is crucially to understand that peak intensity variations in mass approach are expected as a result of the ESI-MS detection setup that allows a nominal mass precision which in turn enables the

possibility of having different isomers being considered the same compound. Nevertheless, the same hierarchical ion intensity relationship among samples can be stated.

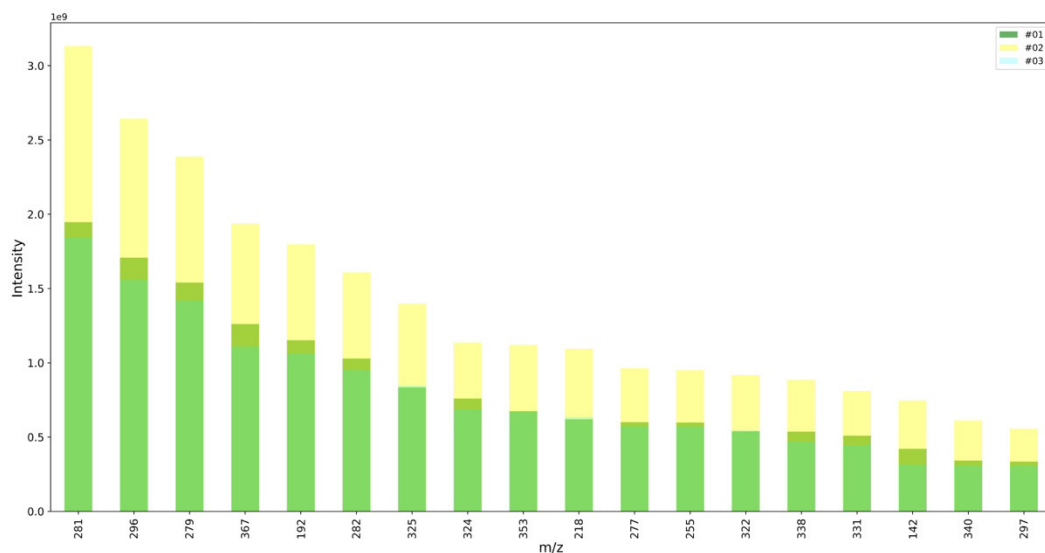


Figure 20 - Ion distribution of the highest intensity ions for 3 replicate samples (mass approach).

Finding and exploring trends within samples is an important step to define which features are more promising to use in ML model development. **Figures 19** and **20** suggest TIC (time approach) and ion intensity (mass approach) as suitable approaches to be used as features to feed ML models with.

Figure 21 shows a portion of the used datasets for both approaches.

| 00:01:42 | 00:02:00 | 00:02:12 | ... | 00:29:00 | 00:29:18 | 00:29:30 | 00:29:48 | 00:30:00 | sample |
|--------------|--------------|--------------|-----|--------------|--------------|--------------|--------------|------------|----------|
| 7.991387e+05 | 7.902258e+05 | 1.552634e+06 | ... | 1.700168e+06 | 1.935095e+06 | 2.040455e+06 | 2.058520e+06 | 2370082.59 | A2B2C2D2 |
| 7.678751e+05 | 7.433682e+05 | 1.600130e+06 | ... | 1.762859e+06 | 1.920391e+06 | 2.077753e+06 | 2.106003e+06 | 1839966.93 | A2B1C2D2 |
| 6.548013e+05 | 6.540540e+05 | 1.699338e+06 | ... | 1.208580e+06 | 1.408682e+06 | 1.562559e+06 | 1.557216e+06 | 1615971.67 | A2B1C2D1 |
| 1.078850e+06 | 1.058425e+06 | 2.150201e+06 | ... | 2.253817e+06 | 2.617136e+06 | 2.923317e+06 | 2.965594e+06 | 3262562.48 | A1B3C1D1 |
| 7.132024e+05 | 7.195547e+05 | 2.171150e+06 | ... | 1.701592e+06 | 1.854413e+06 | 2.062076e+06 | 2.185079e+06 | 2316800.79 | A1B3C1D2 |

| 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | ... | 996 | 997 | 998 | 999 | 1000 | sample |
|-----|-----|-----|-----|-----|-----|--------|--------|-----|---------|-----|------------|------------|------------|------------|------------|----------|
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.0 | 0.00 | ... | 1366629.52 | 1191830.87 | 1208733.56 | 1291423.64 | 3872068.43 | A2B2C2D2 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.0 | 1213.18 | ... | 1440790.05 | 1143541.51 | 963677.52 | 1470470.18 | 3020333.31 | A2B1C2D2 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 795.72 | 0.00 | 0.0 | 0.00 | ... | 1212624.83 | 1342357.12 | 1176706.50 | 1202995.57 | 1275746.38 | A2B1C2D1 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 0.00 | 0.0 | 0.00 | ... | 1917269.69 | 1904425.15 | 1916397.26 | 1941449.31 | 1663126.33 | A1B3C1D1 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.00 | 174.33 | 0.0 | 115.52 | ... | 2011548.56 | 1505143.77 | 1504687.26 | 1515248.48 | 1131924.22 | A1B3C1D2 |

Figure 21 - Portions of the original datasets used in model development. Time approach (up), mass approach (down).

Time approach dataset (**Figure 21, up**) contains 120 features (0 to 30 min) whereas mass approach dataset (**Figure 21, down**) contains 950 features (50 to 1000 Da). Both datasets have 168 rows which corresponds to the number of samples that were successfully analysed. 12 samples were not analysed due to contaminations related with the extraction process.

For each MC, a series of models were developed aiming to evaluate them at the prediction task. Models differ in the combination of data preprocessing and learner used, where for each data preprocessing technique all of the learners described in **Table 5** were applied, independently.

Table 5 – Data preprocessing and learners employed in the models.

| Data preprocessing | Learners |
|------------------------------|---------------------------|
| None | Logistic regression |
| Standard scaler (SS) | Decision tree |
| Principal component analysis | Random forest |
| | Extreme gradient boosting |
| | Support vector machines |

These combinations result in fifteen different ML models. The choice for these learners was done in order to cover tree-based and support vectors modelling techniques.

4.3.2. Time approach

This subsection describes the obtained model performances for all MCs regarding time approach measured by the averaged precision (n = 5). **Table 6** describes the obtained model performances results on test data. The attained performances across all classifiers indicate data is linearly separated. SVM weak performance's without data pre-processing are expected since SVM are highly sensitive to data scaling⁹⁴.

Table 6 - Results of model performances on test data regarding time approach, measured by the averaged precision (n = 5).

| Manufacturing Condition | Preprocessing | LR | DT | RF | XGB | SVM |
|-------------------------|------------------|-----------|-----------|-----------|-----------|-----------|
| A | None | 89 (± 9) | 92 (± 7) | 83 (± 8) | 86 (± 3) | 67 (± 16) |
| A | SS | 89 (± 10) | 92 (± 7) | 89 (± 10) | 86 (± 3) | 86 (± 3) |
| A | PCA (.95) | 80 (± 11) | 70 (± 18) | 79 (± 10) | 77 (± 15) | 68 (± 15) |
| B | None | 97(± 4) | 93 (± 7) | 93 (± 7) | 91 (± 6) | 29 (± 6) |
| B | SS | 96 (± 4) | 93 (± 7) | 96 (± 4) | 91 (± 6) | 91 (± 6) |
| B | PCA (.95) | 93 (± 8) | 77 (± 15) | 80 (± 14) | 85 (± 8) | 74 (± 14) |
| C | None | 100 (± 0) | 95 (± 4) | 99 (± 3) | 95 (± 4) | 66 (± 8) |
| C | SS | 99 (± 3) | 95 (± 4) | 99 (± 3) | 95 (± 4) | 95 (± 4) |
| C | PCA (.95) | 99 (± 3) | 93 (± 5) | 99 (± 3) | 91 (± 9) | 91 (± 6) |
| D | None | 97 (± 3) | 94 (± 3) | 96 (± 4) | 96 (± 4) | 47 (± 8) |
| D | SS | 94 (± 10) | 94 (± 3) | 94 (± 3) | 96 (± 4) | 96 (± 3) |
| D | PCA (.95) | 94 (± 3) | 84 (± 8) | 91 (± 9) | 89 (± 9) | 91 (± 6) |

The same information can be graphically depicted as a clustermap grouped according to the Euclidean distance among results – **Figure 22**. SVM results obtained without preprocessing were removed for interpretation purposes. **Figure 22(a)** shows model average performance and **Figure**

22(b) the correspondent standard deviation of each model. Results indicate that an overall good linear separation is achieved with excellent performances predicting MCs C and D. Classifiers' performances tend to be slightly worst on MCs A and B which suggest these conditions don't have a predominant impact on samples' chemical composition.

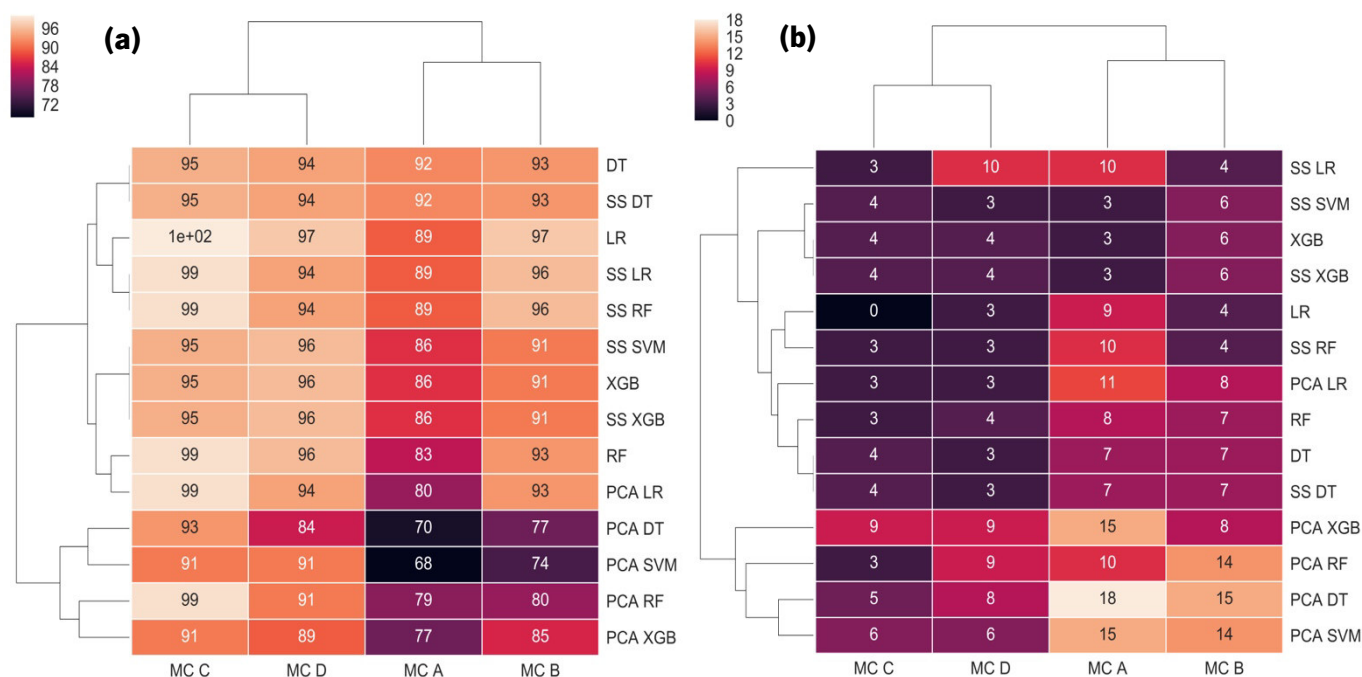


Figure 22 - Clustermap with (a) model average performances' and (b) standard deviation regarding time approach measured by the averaged precision (n = 5).

Regarding data preprocessing, best results are achieved with standard scaling or with no preprocessing at all. PCA tends to decrease classifiers performances' contrary with other applications^{160,161}. This may be attributed to the fact that PCA does not consider the target variable but rather independent variables' variance. Thus, by using less descriptors the model is not able to capture all data trends which also suggests the model is using a large number of features in its decision process. This seems particularly important regarding MC A where PCA has the most negative impact on classifiers' performances. Data suggests the best classifiers for this problem to be DT and LR both with and without standard scaling as these algorithms are not sensitive to scaling¹²⁶.

ML models seem to be stable to data splitting (**Figure 22(b)**) where MCs C and D show best stability. MCs A and B results show less stability to data splitting. Considering model's architecture, better results can be achieved with tree-based models as XGB, DTs or RFs with and without data scaling. PCA also tends to decrease classifiers' stability to data splitting especially in MCs A and B prediction which is in agreement with classifiers' performances when PCA is applied as preprocessing technique.

Score plots of the first two principal components (PCs) (**Figure 23**) allow to reproduce some of the intuition behind classifiers' performances where the first two PCs explain together 62.9% of system's variance. **Figures 23(c)** and **23(d)** graphically show why classifiers performance are usually better at predicting MCs C and D compared to MCs A and B. These results strongly indicate that the employed analytical conditions should be fine-tuned according to the MC that wants to be predicted.

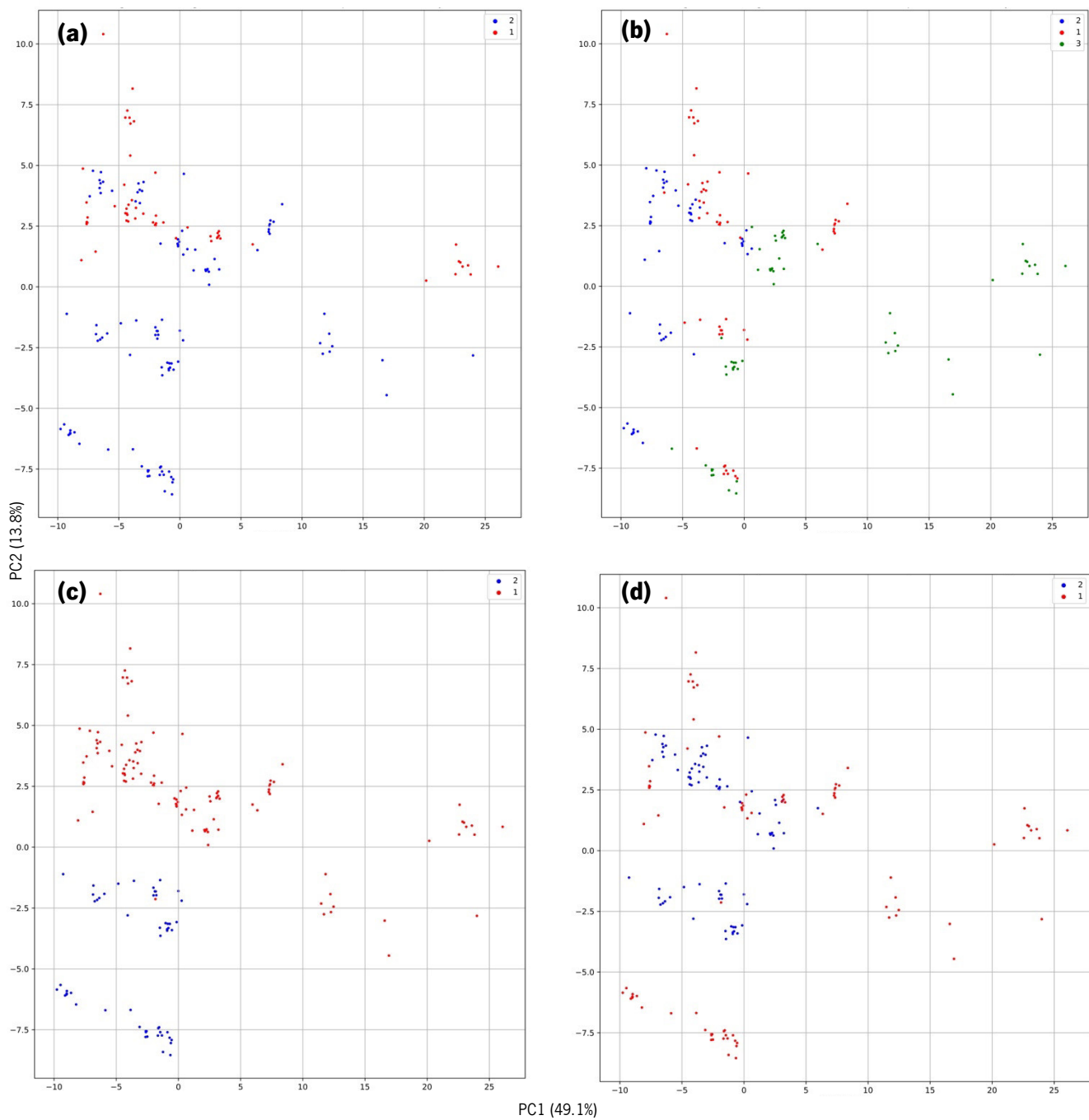


Figure 23 - Score plot of the first two principal components. (a) MC A, (b) MC B, (c), MC C, (d) MC D.

* Loadings were not included for interpretation purposes since all classifiers used a large number of features which would result in a complicated visualization. Selecting a small portion of them would distort the plot resulting in misleading conclusions.

4.3.3. Mass approach

Here are described the obtained ML model performances for all MCs regarding mass approach measured by the averaged precision ($n = 5$). **Table 7** shows model performance of the different classifiers on all MCs. Data shows excellent linear class separation across all MCs for several models and high stability to model splitting.

Table 7 – Results of model performances on test data regarding mass approach, measured by the averaged precision ($n = 5$).

| Manufacturing Condition | Preprocessing | LR | DT | RF | XGB | SVM |
|-------------------------|------------------|-----------------|----------------|-----------------|-----------------|-----------------|
| A | None | 98 (± 3) | 98 (± 3) | 99 (± 2) | 99 (± 2) | 65 (± 10) |
| A | SS | 99 (± 2) | 98 (± 3) | 99 (± 2) | 99 (± 2) | 99 (± 2) |
| A | PCA (.95) | 100 (± 0) | 94 (± 3) | 98 (± 2) | 95 (± 4) | 94 (± 4) |
| B | None | 100 (± 0) | 98 (± 3) | 95 (± 4) | 98 (± 3) | 28 (± 5) |
| B | SS | 100 (± 0) | 98 (± 3) | 100 (± 0) | 98 (± 3) | 98 (± 3) |
| B | PCA (.95) | 100 (± 0) | 94 (± 4) | 98 (± 3) | 100 (± 0) | 88 (± 4) |
| C | None | 100 (± 0) | 99 (± 3) | 100 (± 0) | 100 (± 0) | 65 (± 8) |
| C | SS | 100 (± 0) | 99 (± 3) | 100 (± 0) | 100 (± 0) | 100 (± 0) |
| C | PCA (.95) | 100 (± 0) | 95 (± 5) | 99 (± 2) | 96 (± 5) | 95 (± 5) |
| D | None | 97 (± 3) | 90 (± 5) | 94 (± 5) | 92 (± 4) | 44 (± 9) |
| D | SS | 98 (± 5) | 90 (± 5) | 98 (± 5) | 92 (± 4) | 92 (± 4) |
| D | PCA (.95) | 98 (± 5) | 96 (± 4) | 94 (± 3) | 98 (± 2) | 89 (± 7) |

Its graphical visualization (**Figure 24**) shows a perfect class separation regarding MCs C and B for several models. Models tend to perform worst predicting MC D which may also suggest that this condition doesn't have a predominant impact on samples chemical composition. Preprocessing doesn't seem to have a large impact on model performances since different combinations showed perfect separations in MCs B and C. Applying PCA as a preprocessing technique to feed SVM decreases model performance compared with the other combinations. Despite this, applying PCA and LR shows to be the best combination for MCs A, B and C.

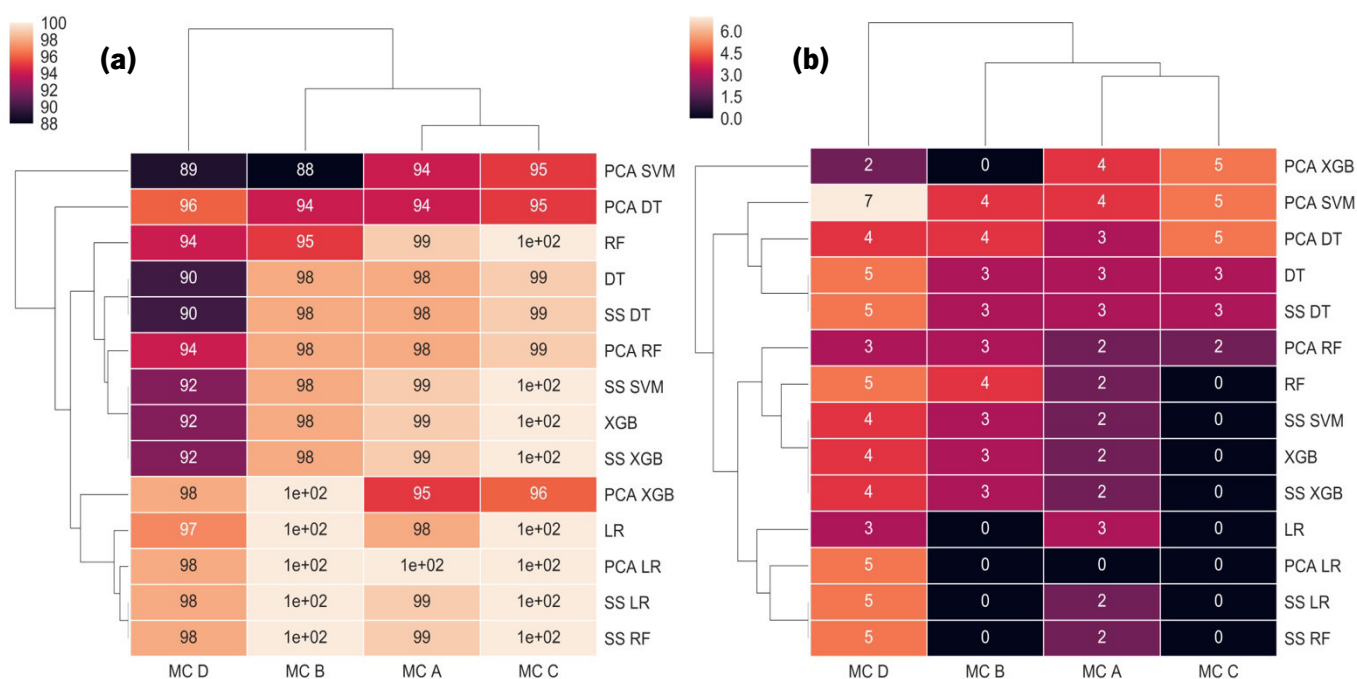


Figure 24 - Clustermap with (a) model's average performances and (b) standard deviation regarding mass approach, measured by the averaged precision ($n = 5$).

ML models show high stability to data splitting across all MCs with excellent results regarding MC C. Results show model stability is worst when predicting MC D. Applying PCA in preprocessing seems to decrease model stability especially for XGB, SVMs and DTs.

Score plots of the first two PCs (**Figure 25**) reproduce the intuition behind model working. The first two PCs explain 72% of system's variance. **Figure 25(c)** illustrates perfect class separation regarding MC C which suggests this condition has relevant differences regarding the chemical composition of both classes.

Mass approach results are overall better than time approach. The main reason for this may come from the fact that time approach contains all information present in mass approach in a processed way, thus leading to a decrease in model performance. Mass approach uses ion intensity resulting in 950 features while time approach only has 120 features which ends up with having less statistics to work on.

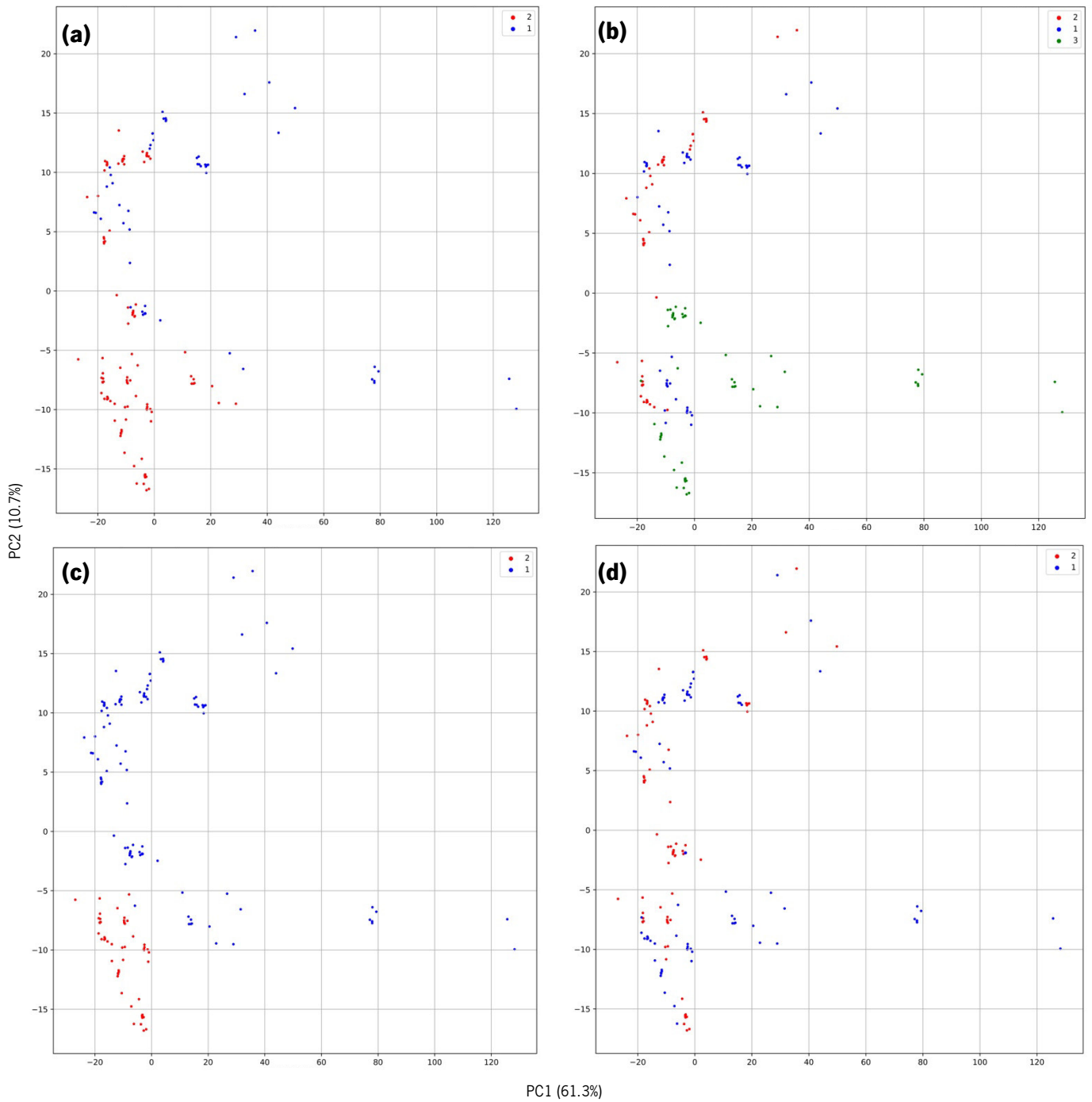


Figure 25 - Score plot of the first two principal components. (a) MC A, (b) MC B, (c), MC C, (d) MC D.

* Loadings were not included for interpretation purposes since all classifiers used a large number of features which would result in a complicated visualization. Selecting a small portion of them would distort the plot resulting in misleading conclusions.

Another interesting property regarding mass approach is the importance of the used features (in this case, ion intensities from 50 to 1000 Da) can be assessed with some models. Considering tree-based models, this is done by iteratively measure how much the precision changes when a feature is excluded from model training. **Figure 26** shows feature importance related to models that combine SS and RF for each MC concerning the 20 ions with largest importance. As previously suggested, data corroborates the fact that models are using a large number of features in MC prediction with the exception of **Figure 26(c)** where less features are used and, therefore, a large importance to the used features is given. This might be representative of the complexity of the chemical composition for different classes of the same MC. It seems the more similar chemical compositions are (e.g. A1 vs A2 samples), the large the number of features the model uses to classify. Data also suggests that heavier ions have a detrimental impact at predicting MCs A and B compared to MCs C and D. This might suggest the possibility of dimer formation with the employed conditions in the MS interface as pointed out by other authors¹⁶².

As previously mentioned, one limitation of this approach and specifically of assessing feature importance relates with the employed analytical setup which has a nominal precision and may in turn lead to misclassify isomers as the same compound. Nevertheless, with a high-resolution MS setup this limitation might be tackled enabling the analytical chemist to know the main differences regarding the chemical composition of the analysed samples.

Identifying which ions the model uses to classify samples rather than just predicting the response using unknown rules decided by a learner might be of interesting for different applications.

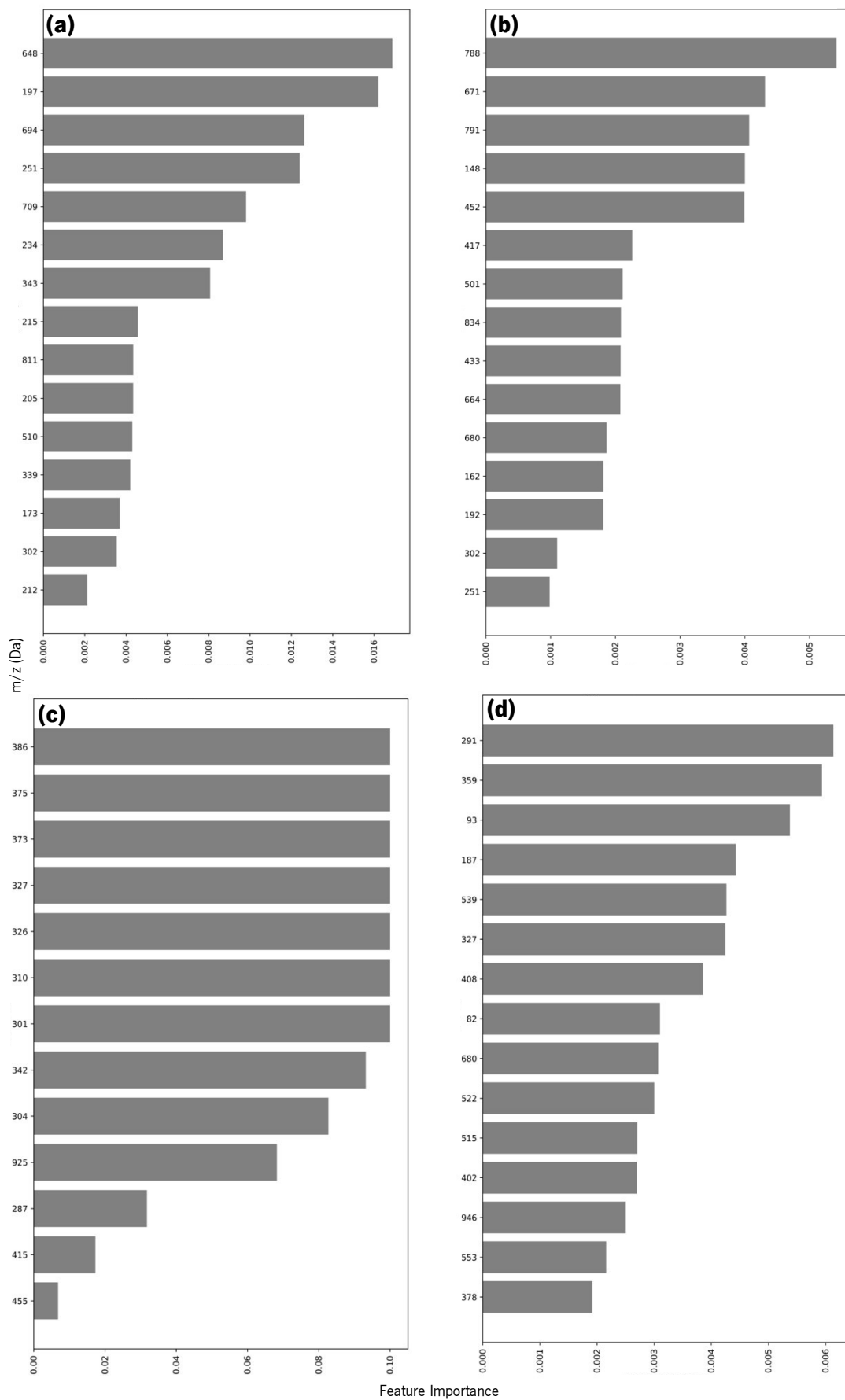
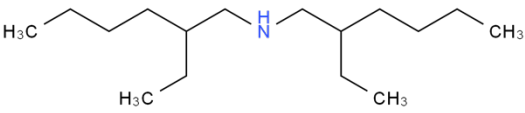
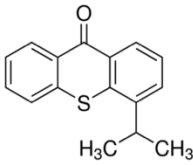
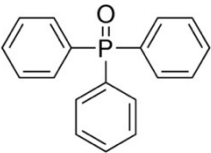
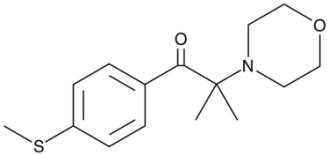


Figure 26 - Feature importance of models with SS and RF for all MCs regarding mass approach.

4.4. Structure elucidation from feature importance using ESI-MS/MS

Following the feature importance attributed by models composed by SS and RF regarding mass approach, ESI-MS/MS of several ions was performed where some structure suggestions were successfully found. **Table 8** summarizes the four chemical structure suggestions that we were able to accomplish with a certain degree of confidence measured by HighRes metric provided by the mzcloud software^{163,164}. This metric measures the relationship between two spectra by means of a match factor which is determined from the m/z value and the abundance correlation coefficients where matches higher than 50 should be understood as probable structures.

Table 7 – Results of model performances on test data regarding mass approach.

| m/z (Da) | Structure suggestion [HighRes metric] | Name |
|-------------|---|---|
| 242 |  [70.4] | Bis-(2-ethylhexyl)-amine |
| 255 |  [87.8] | Isopropyl-9H-thioxanthene-9-one |
| 279 |  [81.4] | Triphenylphosphine oxide |
| 280 |  [97.3] | 2-methyl-4-(methylthio)-2-morpholinopropiophenone |

From the found structures, most of them show aromatic structures with C=O/ P=O bonds, with the exception of the m/z 242 which consists in a di-substituted n-alkyl amine. Ions with m/z 280 and m/z 279 were related to specific parts of the PCB's manufacturing process, namely, solder paste reflux and UV curing¹⁶³⁻¹⁶⁵.

4.5. Data augmentation technique

A data augmentation technique was developed alongside aiming to generate more samples for all classes in order to increase model performance. This way, the first experiment involved picking different samples after performing SLE, mix same class samples and then analyse the new synthetic samples. **Figure 27** shows three chromatograms where – after performing SLE – two samples (black and red line, both A2 samples) are mixed in a proportion 1:1 (v/v) in order to generate a third synthetic sample (blue, synthetic A2 sample) which in theory would have the same properties as its “parents”.

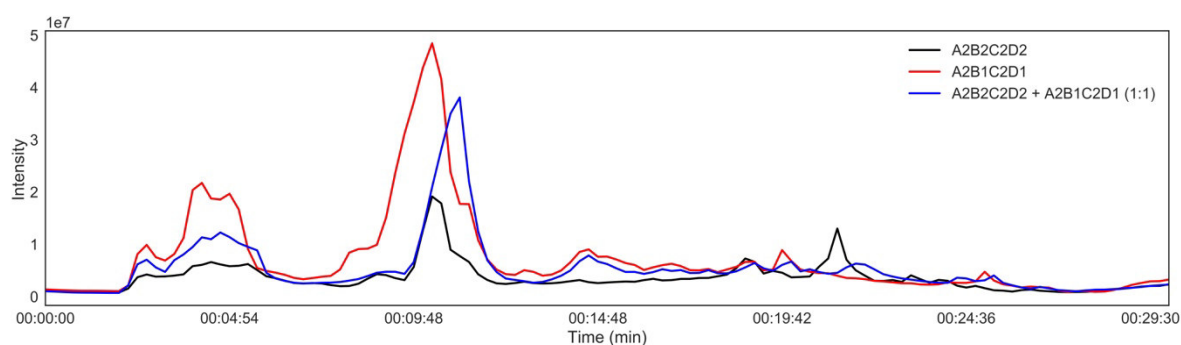


Figure 27 - First experiment in data augmentation technique development based on mixing same class samples and analyse the synthetic sample.

Figure 27 shows the generated synthetic sample whose chemical profile is different from the rest of the original samples. Applying this methodology would in turn decrease model performance by introducing false insights in the data which drove to abort this experiment. The next intuitive step would be to mix same replicate samples in order to generate a new synthetic replicate sample.

This led to a new, innovative trial based on the chromatographic notion that the same sample analysed multiple times generates chemical profiles with slight differences in each analysis as a result of the analytical setup, as previously described¹⁵⁹. In this work, the same notion can be applied to all independent replicates of the same sample. **Figure 28** schematically depicts the intuition behind this data augmentation technique regarding time approach. Results related to data augmentation technique are all presented for time approach since model performances for mass approach already allowed perfect separation for some models.

For each of the 120 features, a minimum, an average and a maximum value can be recorded based on chromatograms from the 10 independent replicates. The developed data augmentation technique takes advantage of these three values to iteratively generate new datapoints within these boundaries according to a triangular distribution (**Figure 28(b)**).

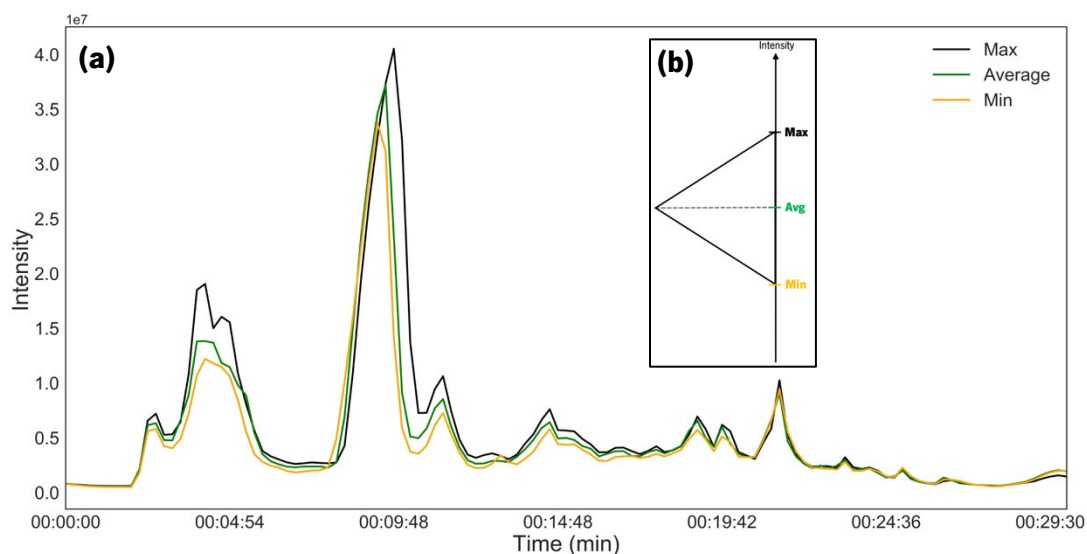


Figure 28 - Second experiment in data augmentation technique development based on generating new datapoints within the observed intensity boundaries for same class samples.

*Chromatograms should be interpreted as a schematic representation of the maximum, average and minimum value for all replicate samples.

This way, new synthetic chromatograms can be generated based on the interval between the minimum and maximum value. Triangular distributions are typically used when the real sample

distribution is not acknowledged but the minimum and maximum values as well as the most likely value are known.

By applying this data augmentation technique, the original dataset can be synthetically oversampled to contain more samples. The visual result is depicted in **Figure 29** where from the four A2B1C2D1 chromatograms shown, two are real and two are synthetic.

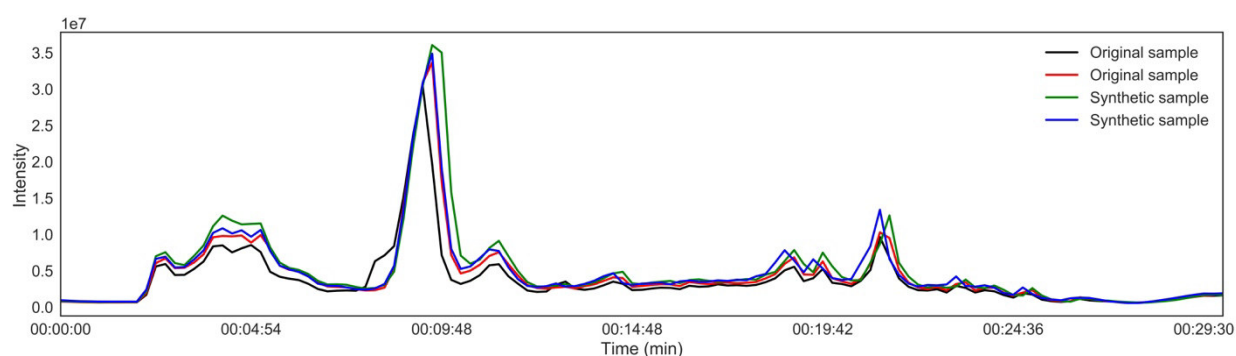


Figure 29 – Chromatograms of real and synthetic samples belonging to A2B1C2D1 samples.

Chemical profiles reveal homogeneous across all samples thus indicating the data augmentation technique can be applied and model performance results compared with previous results and other oversampling techniques.

Figure 30 shows the score plot of the first two PCs and compares the original dataset (**Figure 30(a)**) with the synthetic dataset generated with the developed oversampling technique (**Figure 30(b)**). The augmented dataset contains 10 times more samples than the original one. With the application of the data augmentation technique 18 different clusters are now well-defined, corresponding to all different ABCD MCs combinations as described in Chapter 3.

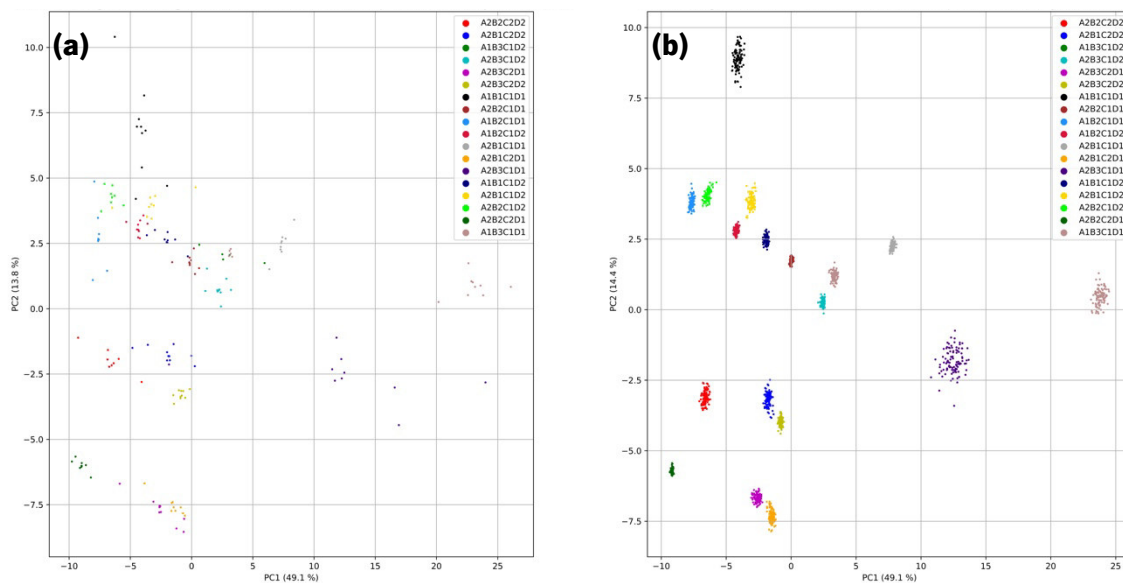


Figure 30 - Comparison of the effect of data augmentation. Score plots of the first two PCs. (a) original dataset, (b) synthetic dataset (x10 more instances).

The impact of the data augmentation technique in the overall model performance was also assessed. **Figure 31** shows the comparison of model performances with original data (OD) and with synthetic data (DA). During model development the original dataset was divided in 2 parts. One was used for oversampling and to subsequently train models on and the other was used as validation set. Results show that the application of the data augmentation technique overall increases model performance across all MCs. Regarding model stability to data splitting, models which were trained with synthetic data show overall higher stability when compared with models trained on the original data.

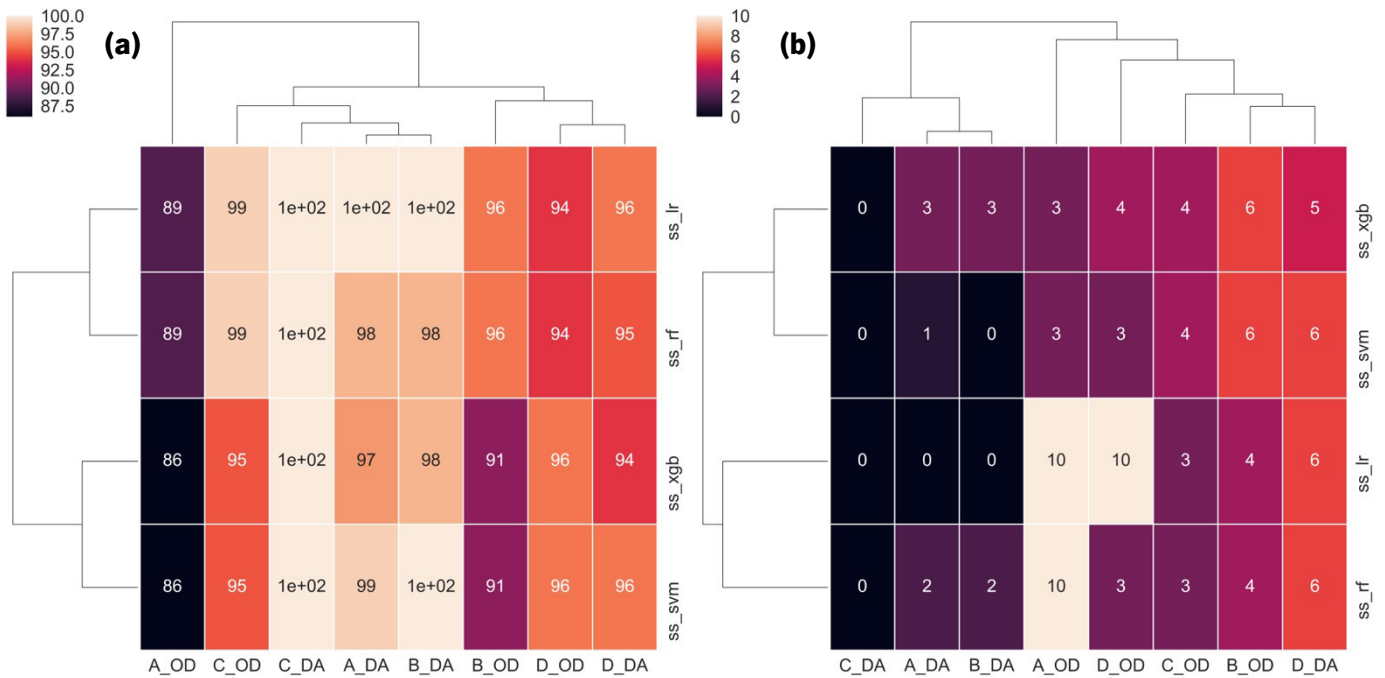


Figure 31 - Comparison of model performance with original data and synthetic data regarding time approach. **(a)** shows model performance measured by the averaged precision ($n = 5$) and **(b)** shows its correspondent standard deviation.

On the other hand, it is important to compare the developed technique with other over-sampling techniques. Random over-sampling (ROS), synthetic minority over-sampling technique¹⁶⁶ (SMOTE) and adaptative synthetic sampling¹⁶⁷ (ADASYN) were applied to the original dataset. The idea behind these techniques is different from the developed one. Whilst our data augmentation technique over-samples all classes, these techniques aim to over-sample minority classes, so all classes have the same number of instances. This way, the presented results for ROS, SMOTE and ADASYN show a balanced class dataset, whilst results concerning our over-sampling technique show an imbalanced dataset with 10 times more instances per class. **Figure 32** shows the comparison of four different oversampling techniques regarding four different models for all MCs. Results show an overall better performance of the developed technique compared with the others. Results also show that the application of ROS, ADASYN and SMOTE in MC A tend to worsen model performance. Nevertheless, perfect class separations are achieved at predicting MC C using other techniques with models whose learners are LR or SVM.

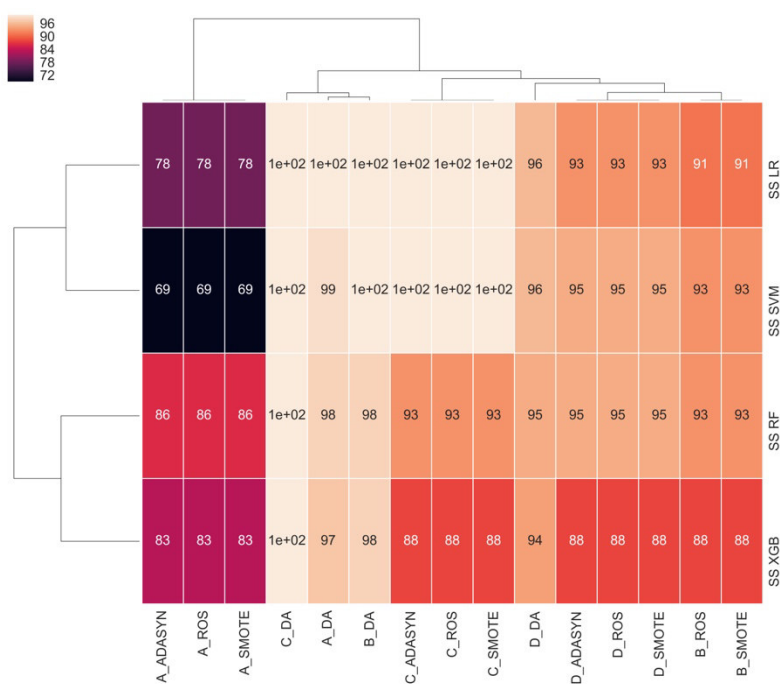


Figure 32 – Comparison of the effect of different over-sampling techniques on model performance's results regarding time approach, measured by the averaged precision (n = 5).

To the best of our knowledge it is the first time an over-sampling technique was developed for chemical classification problems using chromatographic data. This may explain the observed improvement in model performances when compared with state-of-the-art techniques such as SMOTE and ADASYN, concerning this problematic.

Chapter 5 – Conclusions and future work

The objective of this Master thesis is to use chemical data acquired with chromatographic techniques to extract knowledge out of samples by building ML models to classify a relevant parameter (in this case, a MC). The project was developed in two complementary parts. The first one consisted in generating high-quality chemical data through HPLC-MS, while the second part focused on using the acquired chromatographic data to build ML models capable of predicting the employed MCs as well as extracting knowledge of the developed models.

Regarding the first part, the obtained results of the second part confirmed that the employed analytical conditions were adequate for the purposed end. This is supported either by score plots of the first two PCs as well as by model performance. These results also show that the analytical conditions should be fine-tuned according to the MC the model is predicting when top performances are required, considering that model performance is dependent on the employed analytical conditions (chemical fingerprint).

An important note relates to the fact that despite the ever-increasing number of works in this interface, currently, there aren't any methodologies on how to validate an analytical method with focus on ML model development as far as our knowledge could go. Given the rise of works in this interface and the potential it comprises, further work should be done in order to create methodologies for method validation.

The second part of the developed work compared two complementary featurization approaches. Both allowed up to perfect separations (model precision of 100%) for some MCs. Time approach results were overall slightly worse compared to mass approach.

Assessing mass approach's feature importance came to be one of the main interesting topics covered in this study. Further work should be done in understanding the relationship between feature importance and the impact a feature (ion intensity) has in differentiating the chemical composition of the

studied samples. Ultimately, this relationship might enable an analyst to understand what are the exact chemical compounds that make the studied samples different in terms of chemical composition. On the other hand, mass approach application is conditioned by the analytical equipment available. As mentioned, mass approach requires high-resolution MS in order to avoid false insights regarding isomers. In cases where high-resolution MS is not available, time approach might be the best option to explore when developing predictive models, in order to prevent isomers with different retention times to be acknowledged as the same ion.

Both approaches showed excellent model performances whereas mass approach revealed itself as the overall best approach considering both the sole purpose of MC prediction as well as sample's chemical nature elucidation. A third, more complete approach that could, eventually, cover both retention time as well as mass spectra should be target of further work in order to tackle limitations regarding the developed approaches.

Main limitations of this work relate with the fact that all samples must be analysed with the same analytical method and with the time-consuming lab work required to generate chemical data. In this work, ca. 10 replicate samples of 18 different classes were extracted and analysed which might represent an obstacle to the application of this methodology both by industry and academia due to this time-consuming step. However, the proposed data augmentation technique might enable surpassing this limitation since less samples are required to generate enough data to develop predictive models.

Regarding the developed data augmentation technique, results suggest that it might become a detrimental part of related works in this interface as it allows the use of less samples, which drastically diminishes the time required to build datasets. Nonetheless, further work should be done in order to answer open questions such as the minimum number of samples required to be representative of the whole MC class while maintaining accurate boundaries for oversampling. Equally important, a threshold for the largest difference between the maximum and minimum value should also be studied. Additionally,

further work should be done in the application of the developed over-sampling technique with other analytical techniques (e.g. NIR, GC, NMR, etc.) in order to verify if the same behaviour exists with other types of chemical data.

In conclusion, the developed work showed that new insights regarding sample nature can be acquired by applying ML in analytical chemistry. Areas in which gathering knowledge regarding sample nature is of significant importance will find value in this work. The developed work might have a crucial role in routine lab works in areas such as quality control regarding sample contamination, sample forgery, among others. Furthermore, the large-scale application of this methodology could enable interesting discoveries regarding what are the main differences in the chemical composition of different yet related samples (e.g.: natural plants, beverage industry, biotechnology, pharmaceutical, etc.).

Bibliography

1. Elving, P. J. The Analytical Process in Chemistry. *Anal. Chem.* **22**, 962–965 (1950).
2. Skoog, D. A., West, D. M., Holler, F. J. & Crouch, S. R. *Fundamentals of Analytical Chemistry*. (Mary Finch, 2014).
3. Potter, D. W. & Pawliszyn, J. Rapid Determination of Polyaromatic Hydrocarbons and Polychlorinated Biphenyls in Water Using Solid-Phase Microextraction and GC/MS. *Environ. Sci. Technol.* **28**, 298–305 (1994).
4. Samuelsson, L. M. *et al.* Using NMR metabolomics to identify responses of an environmental estrogen in blood plasma of fish. *Aquat. Toxicol.* **78**, 341–349 (2006).
5. Chitescu, C. L. *et al.* High sensitive multiresidue analysis of pharmaceuticals and antifungals in surface water using U-HPLC-Q-Exactive Orbitrap HRMS. Application to the Danube river basin on the Romanian territory. *Sci. Total Environ.* **532**, 501–511 (2015).
6. Käppler, A. *et al.* Analysis of environmental microplastics by vibrational microspectroscopy: FTIR, Raman or both? *Anal. Bioanal. Chem.* **408**, 8377–8391 (2016).
7. Ma, J. *et al.* Metal organic frameworks (MOFs) for magnetic solid-phase extraction of pyrazole/pyrrole pesticides in environmental water samples followed by HPLC-DAD determination. *Talanta* **161**, 686–692 (2016).
8. Mathew, T. *et al.* Technologies for Clinical Diagnosis Using Expired Human Breath Analysis. *Diagnostics* **5**, 27–60 (2015).
9. Nakhleh, M. K. *et al.* Diagnosis and Classification of 17 Diseases from 1404 Subjects via Pattern Analysis of Exhaled Molecules. *ACS Nano* **11**, 112–125 (2017).
10. Zijdenbos, A. P. *et al.* Automatic ‘pipeline’ analysis of 3-D MRI data for clinical trials: Application to multiple sclerosis. *IEEE Trans. Med. Imaging* **21**, 1280–1291 (2002).
11. Baumann, J. M. *et al.* Nucleation and aggregation of calcium oxalate in whole urine; spectrophotometric sedimentation analysis: A new approach to study the aggregation of calcium oxalate dihydrate. *Urol. Res.* **28**, 147–154 (2000).
12. Jin, D. *et al.* Quantitative determination of uric acid using CdTe nanoparticles as fluorescence probes. *Biosens. Bioelectron.* **77**, 359–365 (2016).
13. Elliott, S. P. *et al.* The first reported fatality associated with the synthetic opioid 3,4-dichloro-N-[2-(dimethylamino)cyclohexyl]-N-methylbenzamide (U-47700) and implications for forensic analysis. *Drug Testing and Analysis* **8**, 875–879 (2016).
14. Lyan, B. *et al.* Simple method for clinical determination of 13 carotenoids in human plasma using an isocratic high-performance liquid chromatographic method. *J. Chromatogr. B Biomed. Sci. Appl.* **751**, 297–303 (2001).
15. Albero, B. *et al.* Application of matrix solid-phase dispersion followed by GC–MS/MS to the analysis of emerging contaminants in vegetables. *Food Chem.* **217**, 660–667 (2017).
16. Rabelo, S. N. *et al.* FTIR Analysis for Quantification of Fatty Acid Methyl Esters in Biodiesel Produced by Microwave-Assisted Transesterification. *Int. J. Environ. Sci. Dev.* **6**, 964–969 (2015).
17. Masi, E. *et al.* PTR-TOF-MS and HPLC analysis in the characterization of saffron (*Crocus sativus* L.) from Italy and Iran. *Food Chem.* **192**, 75–81 (2016).

18. Azeena, S. *et al.* Antibacterial activity of agricultural waste derived wollastonite doped with copper for bone tissue engineering. *Mater. Sci. Eng. C* **71**, 1156–1165 (2017).
19. Lazzari, E. *et al.* Classification of biomass through their pyrolytic bio-oil composition using FTIR and PCA analysis. *Ind. Crops Prod.* **111**, 856–864 (2018).
20. Guo, L. *et al.* Simultaneous determination of five synthetic antioxidants in edible vegetable oil by GC-MS. *Anal. Bioanal. Chem.* **386**, 1881–1887 (2006).
21. Lehotay, S. J. *et al.* Comparison of QuEChERS sample preparation methods for the analysis of pesticide residues in fruits and vegetables. *J. Chromatogr. A* **1217**, 2548–2560 (2010).
22. Taurino, A. M. *et al.* Analysis of dry salami by means of an electronic nose and correlation with microbiological methods. in *Sensors and Actuators, B: Chemical* **95**, 123–131 (Elsevier, 2003).
23. Jaffrès, E. *et al.* Sensory characteristics of spoilage and volatile compounds associated with bacteria isolated from cooked and peeled tropical shrimps using SPME-GC-MS analysis. *Int. J. Food Microbiol.* **147**, 195–202 (2011).
24. Lianou, A. *et al.* Rapid assessment of the microbiological quality of pasteurized vanilla cream by means of Fourier transform infrared spectroscopy in tandem with support vector machine analysis. *Food Anal. Methods* **11**, 840–847 (2018).
25. Rocha, J. E. *et al.* HPLC-DAD analysis and antifungal effect of *Hyptis martiusii* Benth (Lamiaceae) against *Candida* strains. *Asian Pac. J. Trop. Biomed.* **9**, 123 (2019).
26. Fuller, M. E. *et al.* Evaluation of ATR-FTIR for analysis of bacterial cellulose impurities. *J. Microbiol. Methods* **144**, 145–151 (2018).
27. Li, X. *et al.* Sequential extraction of soils for multielement analysis by ICP-AES. *Chem. Geol.* **124**, 109–123 (1995).
28. Knowles, E. *et al.* A comparative analysis of potential biosignatures in basalt glass by FIB-TEM. *Chem. Geol.* **330–331**, 165–175 (2012).
29. Smith, C. *et al.* Scanning electron microscope (SEM) microtextural analysis as a paleoclimate tool for fluvial deposits: A modern test. *GSA Bull.* **130**, 1256–1272 (2018).
30. Anderson, C. J. *et al.* Natural solid-state ion conduction induces metal isotope fractionation. *Geol. Soc. Am. / Geol.* **47**, (2019).
31. Fetter, N., Blichert-Toft, J., Télouk, P. & Albarède, F. Extraction of Pb and Zn from crude oil for high-precision isotopic analysis by MC-ICP-MS. *Chem. Geol.* **511**, 112–122 (2019).
32. Nardella, F. *et al.* Chemical investigations of bitumen from Neolithic archaeological excavations in Italy by GC/MS combined with principal component analysis. *Anal. Methods* **11**, 1449–1459 (2019).
33. Horwitz, W. *Nomenclature for sampling in analytical chemistry (Recommendations 1990)*. *Pure and Applied Chemistry* **62**, (1990).
34. Guilbart, G. G. & Hjelm, M. *Nomenclature for automated and mechanised analyses*. *Pure and Applied Chemistry* **61**, (1989).
35. EPA. *Guidance on choosing a sampling design for Environmental Data Collection*. (2002).
36. Ramsey, M. H. & Ellison, S. L. R. *Measurement uncertainty arising from sampling: A guide to methods and approaches*. (2007).
37. Zhang, J. & Zhang, C. Sampling and sampling strategies for environmental analysis. *Int. J. Environ. Anal. Chem.* **92**, 466–478 (2012).

38. Berk, Z. *Food Process Engineering and Technology*. (Elsevier, 2018). doi:10.1016/C2016-0-03186-8
39. Romdhane, M. & Gourdon, C. Investigation in solid–liquid extraction: influence of ultrasound. *Chem. Eng. J.* **87**, 11–19 (2002).
40. Bucić-Kojić, A. *et al.* Study of solid–liquid extraction kinetics of total polyphenols from grape seeds. *J. Food Eng.* **81**, 236–242 (2007).
41. Fincan, M. *et al.* Pulsed electric field treatment for solid–liquid extraction of red beetroot pigment. *J. Food Eng.* **64**, 381–388 (2004).
42. Naviglio, D. *et al.* Rapid Solid-Liquid Dynamic Extraction (RSLDE): A Powerful and Greener Alternative to the Latest Solid-Liquid Extraction Techniques. *Foods* **245**, 1–22 (2019).
43. Gupta, R. C. *Veterinary toxicology: basic and clinical principles*. (Academic Press, 2012).
44. Hajdu, S. I. The first use of the microscope in medicine. *Ann. Clin. Lab. Sci.* **32**, 309–10 (2002).
45. Lawson, A. E. What Does Galileo’s Discovery of Jupiter’s Moons Tell Us About the Process of Scientific Discovery? *Sci. Educ.* **11**, 1–24 (2002).
46. ACS. American Chemical Society Web page. (2019). Available at: <https://www.acs.org/content/acs/en/careers/college-to-career/areas-of-chemistry/analytical-chemistry.html>. (Accessed: 3rd August 2019)
47. Skoog, D. A., West, D. M., Holler, F. J. & Crouch, S. R. *Fundamentals of Analytical Chemistry*. (Brooks/Cole, Cengage Learning, 2014).
48. Doménech-Carbó, M. T. & Osete-Cortina, L. Another beauty of analytical chemistry: chemical analysis of inorganic pigments of art and archaeological objects. *ChemTexts* **2**, 14 (2016).
49. Wilson, H. N. *An Approach to Chemical Analysis: Its Development and Practice*. (Elsevier, 1966).
50. Harvey, D. Analytical Chemistry 2.0—an open-access digital textbook. *Anal. Bioanal. Chem.* **399**, 149–152 (2011).
51. *Chromatography Instruments Market by Type (Systems, Detectors), Consumables (Columns) and Accessories (Auto-sampler Accessories), Applications (Life science Research, Environmental Testing, Food & Beverage Testing), and Region - Global Forecast to 2022*. (2018).
52. ChromatographyToday. How Big is the Chromatography Industry? *Chromatography Today 2* (2015). (Accessed: 3rd August 2019)
53. Tsevt, M. On a new category of adsorption phenomena and their application to biochemical analysis. in *Biological Section of the Warsaw Society of Natural Sciences* (1903).
54. Berezkin, V. G. Biography of Mikhail Semenovitch Tswett and translation of Tswett’s preliminary communication on a new category of adsorption phenomena. *Chem. Rev.* **89**, 279–285 (1989).
55. Hais, I. M. Tswett’s letters to Claparède. *J. Chromatogr. A* **440**, 509–531 (1988).
56. Tompkins, E. R. *et al.* Ion-Exchange as a Separations Method. I. The Separation of Fission-Produced Radioisotopes, Including Individual Rare Earths, by Complexing Elution from Amberlite Resin. *J. Am. Chem. Soc.* **69**, 2769–2777 (1947).
57. James, A. T. *et al.* Gas-liquid partition chromatography: the separation and micro-estimation of ammonia and the methylamines. *Biochem. J.* **52**, 238–242 (1952).
58. Horvath, C. G. *et al.* Fast Liquid Chromatography: An Investigation of Operating Parameters and the Separation of Nucleotides on Pellicular Ion Exchangers. *Anal. Chem.* **39**, 1422–1428 (1967).
59. Campone, L. *et al.* Rapid and automated on-line solid phase extraction HPLC–MS/MS with peak

- focusing for the determination of ochratoxin A in wine samples. *Food Chem.* **244**, 128–135 (2018).
60. Tohma, H. *et al.* Antioxidant activity and phenolic compounds of ginger (*Zingiber officinale* Rosc.) determined by HPLC-MS/MS. *J. Food Meas. Charact.* **11**, 556–566 (2017).
 61. Kaiser, M. *et al.* An Innovative Approach to the Preparation of Plasma Samples for UHPLC–MS Analysis. *J. Agric. Food Chem.* **67**, 6665–6671 (2019).
 62. Lin, M. *et al.* Mathematical Model-Assisted UHPLC-MS/MS Method for Global Profiling and Quantification of Cholesteryl Esters in Hyperlipidemic Golden Hamsters. *Anal. Chem.* **91**, 4504–4512 (2019).
 63. Malaca, S. *et al.* Dilute and shoot ultra-high performance liquid chromatography tandem mass spectrometry (UHPLC–MS/MS) analysis of psychoactive drugs in oral fluid. *J. Pharm. Biomed. Anal.* **170**, 63–67 (2019).
 64. Scollo, E. *et al.* UHPLC-MS/MS analysis of cocoa bean proteomes from four different genotypes. *Food Chem.* 125244 (2019).
 65. Lucini, L. *et al.* QqQ and Q-TOF liquid chromatography mass spectrometry direct aqueous analysis of herbicides and their metabolites in water. *Int. J. Mass Spectrom.* **392**, 16–22 (2015).
 66. Chen, S. *et al.* Metabolite Profiling of 14 Wuyi Rock Tea Cultivars Using UPLC-QTOF MS and UPLC-QqQ MS Combined with Chemometrics. *Molecules* **23**, 104 (2018).
 67. Mandel, J. Statistical methods in analytical chemistry. *J. Chem. Educ.* **26**, 534 (1949).
 68. Sternberg, J. C. *et al.* Spectrophotometric Analysis of Multicomponent Systems Using Least Squares Method in Matrix Form. Ergosterol Irradiation System. *Anal. Chem.* **32**, 84–90 (1960).
 69. Hansch, C. & Fujita, T. ρ - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **86**, 1616–1626 (1964).
 70. Kowalski, B. R. *et al.* Computerized learning machines applied to chemical problems. Multicategory pattern classification by least squares. *Anal. Chem.* **41**, 695–700 (1969).
 71. Kankare, J. J. Computation of equilibrium constants for multicomponent systems from spectrophotometric data. *Anal. Chem.* **42**, 1322–1326 (1970).
 72. Maeder, M. & Neuhold, Y. M. *Practical Data Analysis in Chemistry.* (2007).
 73. Samuel, A. Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.* **3**, 535–554 (1959).
 74. Samuel, A. L. Some studies in machine learning using the game of checkers. II-Recent progress. *IBM J. Res. Dev.* **11**, 601–617 (1967).
 75. Brereton, R. G. *et al.* Chemometrics in analytical chemistry—part I: history, experimental design and data analysis tools. *Anal. Bioanal. Chem.* **409**, 5891–5899 (2017).
 76. Delaney, M. F. *et al.* Chemometric classification of reversed-phase high-performance liquid chromatography columns. *J. Chromatogr. A* **410**, 31–41 (1987).
 77. Osten, D. W. & Kowalski, B. R. Multivariate curve resolution in liquid chromatography. *Anal. Chem.* **56**, 991–995 (1984).
 78. Scoble, H. A. *et al.* Chemometrics and liquid chromatography in the study of acute lymphocytic leukemia. *Anal. Chim. Acta* **150**, 171–181 (1983).
 79. Mark, H. Chemometrics in near-infrared spectroscopy. *Anal. Chim. Acta* **223**, 75–93 (1989).
 80. Salamin, P. A. *et al.* Identification of chemical substances by their near-infrared spectra. *Chemom.*

- Intell. Lab. Syst.* **3**, 329–333 (1988).
81. Otto, M. & Wegscheider, W. Spectrophotometric multicomponent analysis applied to trace metal determinations. *Anal. Chem.* **57**, 63–69 (1985).
 82. Brown, S. D., Tauler, R. & Walczak, B. *Comprehensive Chemometrics. Chemical and Biochemical Data Analysis.* (2009).
 83. Raina, R. *et al.* Large-scale deep unsupervised learning using graphics processors. in *26th Annual International Conference on Machine Learning* 873–880 (2009).
 84. LeCun, Y. *et al.* Deep learning. *Nature* **521**, 436–444 (2015).
 85. Bahrami, M. & Singhal, M. The Role of Cloud Computing Architecture in Big Data. in 275–295 (Springer, Cham, 2015). doi:10.1007/978-3-319-08254-7_13
 86. Helbing, D. *et al.* Will Democracy Survive Big Data and Artificial Intelligence? in *Towards Digital Enlightenment* 73–98 (2019)
 87. Al., P. *et.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
 88. Legg, S. & Hutter, M. *A Collection of Definitions of Intelligence.* (2007).
 89. Moor, J. The Dartmouth College Artificial Intelligence Conference: The Next Fifty years. *AI Mag.* **27**, 87–89 (2006).
 90. Chollet, F. *Deep Learning with Python.* (Manning Publications Co., 2018).
 91. Campbell, M. *et al.* Deep Blue. *Artif. Intell.* **134**, 57–83 (2002).
 92. Musk, E. *An integrated brain-machine interface platform with thousands of channels.* (2019).
 93. Parikh, D. Learning Paradigms in Machine Learning. (2018). Available at: <https://medium.com/datadriveninvestor/learning-paradigms-in-machine-learning-146ebf8b5943>. (Accessed: 9th August 2019)
 94. Mueller, A. & Guido, S. *Introduction to Machine Learning with Python.* (O'Reilly, 2017).
 95. Zhou, L. Simplify Machine Learning Pipeline Analysis with Object Storage. (2018). Available at: <https://blog.westerndigital.com/machine-learning-pipeline-object-storage/>. (Accessed: 9th August 2019)
 96. Kryger, L. Interpretation of Analytical Chemical Information by Pattern Recognition Methods. **28**, 871–887 (1981).
 97. Wenning, R. J. & Erickson, G. A. Interpretation and analysis of complex environmental data using chemometric methods. *Trends Anal. Chem.* **13**, 446–457 (1994).
 98. Biancolillo, A. & Marini, F. Chemometric Methods for Spectroscopy-Based Pharmaceutical Analysis. *Front. Chem.* **6**, 576 (2018).
 99. Singh, A., Yang, L., Hartikainen, K., Finn, C. & Levine, S. End-to-End Robotic Reinforcement Learning without Reward Engineering. (2019).
 100. Haarnoja, T. *et al.* Learning to Walk via Deep Reinforcement Learning. (2018).
 101. Chen, R. *et al.* Word-level sentiment analysis with reinforcement learning. *IOP Conf. Ser. Mater. Sci. Eng.* **490**, 062063 (2019).
 102. Sharma, D. *et al.* Trend Analysis in Machine Learning Research Using Text Mining. in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* 136–141 (2018)
 103. Wang, L. *et al.* Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation. in *Proceedings of the 24th ACM SIGKDD International Conference*

- on *Knowledge Discovery & Data Mining* 2447–2456 (2018).
104. Gottesman, O. *et al.* Guidelines for reinforcement learning in healthcare. *Nat. Med.* **25**, 16–18 (2019).
 105. Zhou, Z. *et al.* Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **3**, 1337–1344 (2017).
 106. Hossin, M. & Sulaiman. A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process* **5**, (2015).
 107. Gopalakrishna, A. K. *et al.* Relevance as a metric for evaluating machine learning algorithms. in *Machine Learning and Data Mining in Pattern Recognition* 195–208 (2013).
 108. Flach, P. Performance Evaluation in Machine Learning: The Good, The Bad, The Ugly and The Way Forward. in *33rd AAAI Conference on Artificial Intelligence* (2019).
 109. Sewel, M. No Free Lunch Theorems. Available at: <http://www.no-free-lunch.org>. (Accessed: 11th August 2019)
 110. Cortes-Ciriano, I. & Bender, A. Improved Chemical Structure–Activity Modeling Through Data Augmentation. *J. Chem. Inf. Model.* **55**, 2682–2692 (2015).
 111. Bjerrum, E. J. *et al.* Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics. 1–10 (2017).
 112. Deisenroth, M. *et al.* *Mathematics for Machine Learning*. (Cambridge University Press, 2019).
 113. Kaggle team. Kaggle - Getting started. (2019). Available at: <https://www.kaggle.com/getting-started/83518>. (Accessed: 13th August 2019)
 114. Cordella, C. PCA: The Basic Building Block of Chemometrics. in *Analytical Chemistry* **47** (2012).
 115. Kumar, K. *Principal Component Analysis: Most Favourite Tool in Chemometrics*. (2017).
 116. Lever, J. *et al.* Principal Component Analysis. *Nat. Methods* **14**, 641–642 (2017).
 117. Zhang, L. *et al.* Classification and adulteration detection of vegetable oils based on fatty acid profiles. *J. Agric. Food Chem.* **62**, 8745–8751 (2014).
 118. Laurens van der Maaten & Geoffrey E., H. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **164**, 10 (2008).
 119. Cramer, J. S. The Origins of Logistic Regression. in *Logit Models from Economics and Other Fields* (Cambridge University Press, 2002).
 120. Navlani, A. Understanding Logistic Regression in Python. (2018). Available at: <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>. (Accessed: 14th August 2018)
 121. Christodoulou, E. *et al.* A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **110**, 12–22 (2019).
 122. Rocks and Rock identification. Available at: <https://www.vagabondgeology.com/rocks-rock-identification.html>. (Accessed: 14th August 2019)
 123. Buratti, S. *et al.* Characterization and classification of Italian Barbera wines by using an electronic nose and an amperometric electronic tongue. *Anal. Chim. Acta* **525**, 133–139 (2004).
 124. Ashman, W. P. *et al.* Decision tree for chemical detection applications. *Anal. Chem.* **57**, 1951–1955 (1985).
 125. Overfitting in machine learning. Available at: <https://elitedatascience.com/overfitting-in-machine-learning>. (Accessed: 14th August 2019)

126. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
127. Hanselmann, M. *et al.* Toward digital staining using imaging mass spectrometry and random forests. *J. Proteome Res.* **8**, 3558–3567 (2009).
128. Svetnik, V. *et al.* Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003).
129. de Santana, F. B. *et al.* Rapid Discrimination Between Authentic and Adulterated Andiroba Oil Using FTIR-HATR Spectroscopy and Random Forest. *Food Anal. Methods* **11**, 1927–1935 (2018).
130. Deng, X. *et al.* Predictive geographical authentication of green tea with protected designation of origin using a random forest model. *Food Control* **107**, 106807 (2020).
131. Training of a GBM.
132. Tree Boosting With XGBoost – Why Does XGBoost Win “Every” Machine Learning Competition? Available at: <https://medium.com/syncedreview/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition-ca8034c0b283>. (Accessed: 15th August 2019)
133. Svetnik, V. *et al.* Boosting: An ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **45**, 786–799 (2005).
134. Lu, J. *et al.* Estimation of elimination half-lives of organic chemicals in humans using gradient boosting machine. *Biochim. Biophys. Acta - Gen. Subj.* **1860**, 2664–2671 (2016).
135. Zhou, Z. & Zare, R. N. Personal Information from Latent Fingerprints Using Desorption Electrospray Ionization Mass Spectrometry and Machine Learning. *Anal. Chem.* **89**, 1369–1372 (2017).
136. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297 (1995).
137. Yue, S. *et al.* SVM classification: Its contents and challenges. *Appl. Math. J. Chinese Univ.* **18**, 332–342 (2003).
138. Bona, E. *et al.* Support vector machines in tandem with infrared spectroscopy for geographical classification of green arabica coffee. *LWT - Food Sci. Technol.* **76**, 330–336 (2017).
139. Chen, S. *et al.* Rapid determination of soil classes in soil profiles using vis-NIR spectroscopy and multiple objectives mixed support vector classification. *Eur. J. Soil Sci.* **70**, 42–53 (2019).
140. Arikawa, Y. Basic Education in Analytical Chemistry. *Anal. Sci.* **17**, 571–573 (2001).
141. Radovic, A. *et al.* Machine learning at the energy and intensity frontiers of particle physics. *Nature* **560**, 41–48 (2018).
142. Webb, S. Deep learning for biology. *Nature* **554**, (2018).
143. Kitchin, J. R. Machine learning in catalysis. *Nat. Catal.* **1**, 230–232 (2018).
144. Jensen, Z. *et al.* A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Cent. Sci.* (2019).
145. Chen, H. *et al.* The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250 (2018)
146. Yang, X. *et al.* Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* (2019)
147. Elton, D. C. *et al.* Applying machine learning techniques to predict the properties of energetic materials. *Sci. Reports - Nat.* **8**, 1–12 (2018).
148. Butler, K. T. *et al.* Machine learning for molecular and materials science. *Nature* **559**, 547–555

- (2018).
149. Kishimoto, A. *et al.* AI meets chemistry. in *32nd AAAI Conference on Artificial Intelligence 7978–7982* (2018).
 150. Segler, M. H. S. *et al.* Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
 151. Cao, M. *et al.* Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* **11**, 696–706 (2015).
 152. Gallidabino, M. D. *et al.* Quantitative profile-profile relationship (QPPR) modelling: A novel machine learning approach to predict and associate chemical characteristics of unspent ammunition from gunshot residue (GSR). *Analyst* **144**, 1128–1139 (2019).
 153. Barbosa, R. M. *et al.* The Use of Decision Trees and Naive Bayes Algorithms and Trace Element Patterns for Controlling the Authenticity of Free-Range-Pastured Hens' Eggs. *J. Food Sci.* **79**, 1672–C1677 (2014).
 154. Elzey, B. *et al.* Determination of adulterated neem and flaxseed oil compositions by FTIR spectroscopy and multivariate regression analysis. *Food Control* **68**, 303–309 (2016).
 155. Qiu, S. & Wang, J. Application of Sensory Evaluation, HS-SPME GC-MS, E-Nose, and E-Tongue for Quality Detection in Citrus Fruits. *J. Food Sci.* **80**, S2296–S2304 (2015).
 156. Martens, H. Quantitative Big Data: Where chemometrics can contribute. *J. Chemom.* **29**, 563–581 (2015).
 157. Circuits, T. I. for I. and P. E. IPC-TM-650 Test Methods Manual - Rosin Flux Residue - HPLC Method. 3 (1995).
 158. Gonçalves, D. & Parpot, P. *Identification of BMW 35 up display contamination - internal report.* (2018).
 159. Meyer, V. *Practical HPLC.* (Wiley, 2004).
 160. Howley, T., Madden, M. G., O'connell, M. L. & Ryder, A. G. *The Effect of Principal Component Analysis on Machine Learning Accuracy with High Dimensional Spectral Data.*
 161. Granato, D. *et al.* Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends Food Sci. Technol.* **72**, 83–90 (2018).
 162. Medzihradzky, K. F. Noncovalent Dimer formation in liquid chromatography-mass spectrometry analysis. *Anal. Chem.* **86**, 8906–8909 (2014).
 163. CIBA Specialty Chemicals. Photoinitiators for UV Curing. 1–8 (2003).
 164. Leggesse, E. G. *et al.* Theoretical study on photochemistry of Irgacure 907. *J. Photochem. Photobiol. A Chem.* **347**, 78–85 (2017).
 165. Kubo, N. *et al.* Flux for lead-free solder, and lead-free solder paste. 13 (2018).
 166. Chawla, N. *et al.* SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
 167. Haibo, H. *et al.* ADASYN: Adaptive synthetic sampling approach for imbalanced learning. in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* 1322–1328 (2008).
 168. Winefordner, J. D. *Sample Preparation Techniques in Analytical Chemistry.* (2003).

