UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

CLUSTERING AND VISUALIZATION OF LITHIUM-ION BATTERY DATA FOR SECOND LIFE APPLICATIONS

GABRIEL DOMÍNGUEZ PIERNAS

Thesis supervisor: MARTHA IVÓN CÁRDENAS DOMÍNGUEZ (Department of Computer Science)

Degree: Bachelor Degree in Informatics Engineering (Computing)

Thesis report

Facultat d'Informàtica de Barcelona (FIB)

Universitat Politècnica de Catalunya (UPC) - BarcelonaTech

27/06/2023

Acknowledgments

First of all, I would like to express my most sincere gratitude to the tutor of this work, Martha Ivón Cárdenas Domínguez, for giving me the opportunity to develop this work, helping me, giving me the necessary confidence and guiding me throughout the work. I would also like to thank René Alquezar Mancho for his work in advising me throughout the process of development and analysis throughout the project. Without them and their invaluable support this work would not have been possible.

I would like to make a special mention to my loved ones, firstly my family, my parents and my brother, for providing me with support and trust throughout all these academic years and giving me that boost when I needed it, and secondly to my partner, for being by my side at all times, for believing in me. Without you I would not be who I am.

Resum

L'ús de bateries de segona vida és una de les estratègies més prometedores per a resoldre el problema del reciclatge de piles en el futur. L'objectiu d'aquesta recerca és aportar solucions pràctiques de garbellat, reagrupació i reutilització donant un nou ús a les piles de liti retirades. Donar una segona oportunitat a aquestes bateries retirades no sols té beneficis ecològics i mediambientals, sinó que també és rendible des del punt de vista econòmic.

Per a aquest projecte, s'ha aprofitat el conjunt de dades que va registrar la NASA per a observar el comportament de les bateries elèctriques d'ió-liti al llarg de la seva vida per a realitzar un estudi que permet predir la vida útil d'una bateria. Mitjançant diferents tècniques d'aprenentatge automàtic, principalment SVM, s'ha aconseguit entrenar un model que pugui predir aquesta característica.

D'aquesta manera es pretén aportar en un camp on la sostenibilitat i l'economia circular són cada vegada més importants. En el futur, es podrien implementar estratègies basades en l'ús de bateries de segona vida en diferents sectors, des de l'emmagatzematge d'energia fins a la mobilitat elèctrica. Realitzant aquest projecte es pretén continuar pel camí de la recerca a partir de tècniques de ML i IA per a contribuir a una economia més sostenible i responsable.

Resumen

El uso de baterías de segunda vida es una de las estrategias más prometedoras para resolver el problema del reciclaje de pilas en el futuro. El objetivo de esta investigación es aportar soluciones prácticas de cribado, reagrupación y reutilización dando un nuevo uso a las pilas de litio retiradas. Dar una segunda oportunidad a estas baterías retiradas no sólo tiene beneficios ecológicos y medioambientales, sino que también es rentable desde el punto de vista económico.

Para este proyecto, se ha aprovechado el conjunto de datos que registró la NASA para observar el comportamiento de las baterías eléctricas de ión-litio a lo largo de su vida para realizar un estudio que permite predecir la vida útil de una batería. Mediante diferentes técnicas de aprendizaje automático, principalmente SVM, se ha conseguido entrenar un modelo que pueda predecir esta característica.

De esta manera se pretende aportar en un campo donde la sostenibilidad y la economía circular son cada vez más importantes. En el futuro, se podrían implementar estrategias basadas en el uso de baterías de segunda vida en diferentes sectores, desde el almacenamiento de energía hasta la movilidad eléctrica. Realizando este proyecto se pretende continuar por el camino de la investigación a partir de técnicas de ML e IA para contribuir a una economía más sostenible y responsable.

Abstract

The use of second life batteries is one of the most promising strategies to solve the problem of battery recycling in the future. The objective of this research is to provide practical solutions of screening, regrouping and reuse by giving a new use to retired lithium batteries. Giving a second chance to these retired batteries not only has ecological and environmental benefits, but is also cost-effective from an economic point of view.

For this project, the data set recorded by NASA to observe the behavior of lithium-ion electric batteries over their lifetime has been used to conduct a study to predict the lifetime of a battery. Using different machine learning techniques, mainly SVM, it has been possible to train a model that can predict this characteristic.

In this way, it aims to contribute to a field where sustainability and circular economy are becoming increasingly important. In the future, strategies based on the use of second life batteries could be implemented in different sectors, from energy storage to electric mobility. This project aims to continue on the path of research based on ML and AI techniques to contribute to a more sustainable and responsible economy.

Contents

List of Figures

List of Tables

Context and scope

1.1 Introduction and contextualization

If the world is to meet its ambitious emissions reduction targets with the urgency that climate change demands, renewable energies and electric transport must be massively adopted [1] [2]. The scarcity of minerals and raw materials has become an increasingly relevant problem in today's society, and one of the most valued resources is lithium [3]. The demand for lithium has increased in recent years due to its use in the production of batteries [4] for electronic devices and electric vehicles, and it represents a key element in energy storage. However, the availability of lithium is limited and the extraction of this resource can have negative impacts on the environment [5]. It is therefore necessary to address this scarcity through recycling and reuse, both of lithium and of all the other minerals and raw materials involved in this process, to ensure a sustainable future.

Figure 1.1: Example image of LiBs [6].

However, retired batteries cannot be used directly for secondary applications for safety reasons due to the inconsistency of both their use and the manufacturing process. Therefore, retired lithium-ion batteries (LiBs) must be sorted and regrouped to improve consistency and ensure their good use and safety in this new life. However, how to obtain these aging characteristics quickly and non-destructively has become a major challenge.
Experimental battery life testing requires significant resources and can take months or years to establish estimates of performance degradation to a given end-of-life (EOL) definition [7]. This is why predicting both the lifetime and condition of batteries from initial characterization data is of great interest to both academia and industry.

Recent studies have shown that ML techniques are able to predict the state of health (SOH) of LiBs from different input attributes such as battery capacity and battery evolution after several charge and discharge cycles [8]. In this project, however, we will focus on identifying, classifying and grouping those batteries focused on giving them a second useful life for another application using different ML techniques.

1.1.1 Context

This project is a final thesis of the Computer Engineering Degree at the Facultat d'Informàtica de Barcelona (FIB) focused on the field of machine learning, a branch of the field of artificial intelligence. This project is a proposal of the project director, Martha Ivón Cárdenas Domínguez, from the Department of Computer Science of the Facultat d'Informàtica de Barcelona (FIB). The objective of the project is to study the classification of LiBs in order to be used for a second life by ML techniques. The study has been carried out from a dataset [9] that has input data of batteries of different types that have been charged and discharged in order to study the evolution of their characteristics due to the deterioration of their use.

1.1.2 Terms and concepts

Artificial Intelligence (AI)

A field of computer science that aims to create intelligent machines that work and learn like humans. AI systems are designed to perform tasks that typically require human intelligence, such as perception, reasoning, learning, and decision-making. AI can be used to automate tasks, solve problems, and make predictions based on data. Examples of AI include natural language processing, computer vision, robotics, and expert systems. The ultimate goal of AI is to create machines that can think and act like humans.

Machine learning (ML)

A branch of artificial intelligence that focuses on the development of algorithms and techniques that allow machines to learn from data and perform tasks without being explicitly programmed to do so. Over time, machine learning systems can improve their performance as they receive more data and learn from it. This enables the automation of processes and decision-making based on patterns and trends in the data.

Li-ion battery (LiB)

A type of rechargeable battery that uses lithium ions as the main charge carrier. Li-ion batteries have a high energy density, meaning they store a lot of energy in a small package, and they have a low self-discharge rate, meaning they can retain their charge for a long time. These characteristics make Li-ion batteries well-suited for portable electronic devices such as smartphones, laptops, and electric vehicles. Li-ion batteries have a high voltage, low memory effect, and are lighter and more compact than other types of rechargeable batteries.

Neural Networks (NN)

A neural network is a type of machine learning model inspired by the structure and function of the human brain. Neural networks consist of interconnected processing nodes, called artificial neurons, that are connected to each other by weighted connections. The neurons process and transmit information through these connections, allowing the network to learn from data and make predictions. Neural networks can be used for a wide range of tasks, including image and speech recognition, natural language processing, and predictive modeling. The strength of neural networks lies in their ability to automatically learn complex

patterns and relationships in data, making them a powerful tool for solving complex problems.

State of health (SOH)

The state of health of a LiB refers to the ability of a lithium-ion battery to maintain and provide power efficiently and consistently. This state can be affected by factors such as the number of charge and discharge cycles, temperature, charge and discharge rate, and the age of the battery. The SOH can be assessed through different measurements, such as battery capacity, voltage and internal resistance. These measurements can be used to predict long-term battery performance and detect potential problems before they affect battery performance.

Remaining Useful Life (RUL)

Refers to the estimated amount of time that a battery can continue to operate efficiently and reliably before it reaches the end of its useful life. RUL is an important metric for assessing the health and performance of batteries, as it helps to determine when a battery needs to be replaced or serviced in order to prevent potential failures or safety hazards. Accurately predicting RUL can also help to optimize battery usage and reduce costs by minimizing unnecessary replacements and downtime.

1.1.3 Problem to be solved

The problem to be solved is mainly to help in the detection and grouping of the different LiBs to destine them to a second life. The economic and temporal cost and the lack of means and qualified personnel make this process of treatment of the discarded LiBs a tedious and complicated case. In addition to this great difficulty, it must be added that there can be problems in the evaluation of the capacity and quality of the discarded batteries, as well as, for example, a bad inspection of the battery for its diagnosis can generate the impossibility of being used again, generating material waste that must be treated. This whole process generates many complications for this classification and identification, which can make it difficult to make decisions about its reuse. The problem is established mainly in detecting quickly and easily those discarded batteries suitable to be used for another purpose and thus extend their useful life.

1.1.4 Stakeholders

This study is addressed, firstly, to the scientific community, since the study of prediction and clustering of data inputs using different ML techniques is an emerging topic in continuous exploration and growth. Different companies, projects and researchers will be able to access this research to use the results for different purposes in this field. These discoveries and improvements can have a significant impact on scientific research and practical problem-solving in a wide range of disciplines. Thus, this study is directed especially to the community and researchers in this area of study of LiBs. With this study, it is intended to continue and give rise to new research in this field.

Second, the main beneficiary of this research turns out to be the LiB industry. It is estimated that the LiB market may grow to tens of thousands of MWhs and exceed an annual market value of $30 billion by 2020 [10]. Finding solutions for discarded batteries and reusing them creates a market opportunity that could be highly competitive due to the amount of batteries removed. This would have a positive impact on both the environment and the economy. In addition, end users would be able to obtain functional batteries for different purposes at lower prices and recycling, reuse and resource treatment companies would also benefit. Finally, this would have a great impact for all of us as we would see a reduction in the demand for new batteries, which means less resource extraction, less manufacturing and less environmental pollution.

1.2 Justification

1.2.1 Previous studies

As mentioned above, the study in the LiBs sector is growing in recent years due to the increasing demand and the need to use them. Several studies, such as those we find referenced [10] [11] [12], have been conducted using different ML techniques around this field but most of them, almost all of them, are based on predicting the SOH of LiBs for different purposes using such models. In contrast, this study revolves around the ability of different ML techniques to predict and group according to which batteries are ready to be used in a different functionality. This new approach is novel in the field as the study has not been approached from a point of view similar to this one until now.

1.2.1 Justification

The fact of changing the focus and destination of these discarded batteries is established as a new research variant which can establish a new starting point for this sector. As a result of this research, new research can be carried out, or the current research can be expanded with new resources and new approaches. This opens the possibility of a new branch of research where the short and long term projection is very interesting for all parties. As we have justified above, this type of research can be beneficial for multiple sectors and companies and can establish a very competitive market.

The need to treat these discarded LiBs will remain as long as they continue to be produced, but the treatment during their useful life will depend on the results of research such as this. In addition, and from an environmental point of view, extending the life of all these batteries will absorb the market of those services or products that require LiBs without such restrictive conditions as electric cars, thus generating a decrease in the production and exploitation of natural resources.

1.3 Scope

1.3.1 Objectives and sub-objectives

This project consists of two generic objectives that divide the project into two stages.
A first stage of initiation in the world of machine learning and its applications in the field of LiBs and contextualization in the context and problems, and a second block of experimentation and deep research in this field.

The first objective is to investigate, learn and become familiar with the context of the treatment of LiBs, their use, recycling and reuse. Also, to document through previous studies in order to understand the current starting point in relation to the research carried out and the results obtained. In this way it is possible to have a broad vision of the problem to be faced and to be able to understand the best way to approach it.

The second objective is to investigate in depth the classification and clustering techniques, visualize the plots obtained and determine the possibility, based on the results obtained, of classifying LiBs discarded for second life through ML techniques.

Within the first objective, the following sub-objectives can be included in order to facilitate the achievement of this objective:

- Understand the characteristics of the context, social and economic situation.
- Understand the difficulties of the treatment of discarded LiBs and the need for an alternative solution for their treatment.
- Understand and be able to analyze the different results obtained from previous studies.
- Understand the characteristics of the LiBs of the dataset and to be able to analyze and detect the most relevant variables for the practical treatment.

Within the second objective, the following sub-objectives can be included in order to facilitate the achievement of this objective:

- Understand how ML works and apply it within the project.
- Experiment with different ML models to be able to evaluate the different results.
- Be able to analyze the results obtained and identify the problems of those that do not fit correctly.
- Experiment with different parameterization in those models where the data does not fit correctly or anomalies have been detected.
- Be able to visualize the results graphically and understand the nature of the results.

1.3.2 Requirements

Since this is a research project, there are some requirements that must be met in order for the research to be useful and profitable for any interested person. The results must have the following characteristics:

- Reproducibility: All experiments must be reproducible and obtain similar results. To ensure accurate comparison of performance in different scenarios, we must ensure that the experiments have been performed under the same conditions.
- Understandable: One of the objectives of the project is to improve the understanding of the subject and to be understood by as many people as possible, for this reason the results should be clear and concise.
- Realistic: Although this is an experiment with specific data from specific LiBs, it must be taken into account that there are multiple types of LiBs and there are certain limitations for each one of them.

1.3.3 Obstacles and risks

Like in all scientific research, obstacles may appear during the development of the thesis.
First, the lack of experience and understanding in the field of ML may cause difficulties and limitations, especially at the beginning of the study. Moreover, this area of research is relatively new, so there aren't many studies with the focus given to this research, the information available is not as extensive and deep as in other disciplines, as there is still a lot of research to be done, especially in the LiBs area. However, this can also be taken as an opportunity to innovate and contribute as much as possible to this branch of ML and this sector.

Secondly, the dataset with all the data entries that we have selected to be used for the investigation of this project may have limitations or may not be optimal for the study that we are going to deploy. Even if other studies have been carried out in the same branch of the LiBs with this same dataset, and it is the most completed one available, when carrying out the study with a different approach we may encounter obstacles or drawbacks when modeling the data. The adaptation of this data may not be suitable for use with the different ML techniques. This fact can be considered as another possible problem, the limitation of any ML technique when being used with our data. Some of the algorithms used may have some limitations in relation to the accuracy or adaptability and treatment with the data.

Another possible obstacle to take into account is the computational cost of working with very large datasets. The dataset we are working with has numerous record entries, and the computational cost of using such a dataset with some elaborate ML technique can be a drawback. These issues result in a decrease in efficiency and speed and, in the worst case, an abortion of the process due to the impossibility of performing it with the computational power available for the volume of input data. With this in mind, the data set may have to be processed in subsets to overcome the computational complexity.

Finally, there is one last threat inherent in any final project of a degree, which is the deadline. This project, not being an independent study with a group of researchers who can dedicate full time, but a final project that must meet predetermined deadlines, may limit the content and depth of the project. Therefore, in case of unforeseen events, it will be necessary to readjust the task plans in order to stick to the schedule and avoid unfortunate rushes.

1.4 Methodology and rigor

1.4.1 Methodology

The development of this work is linked to the research of classification techniques using ML, therefore, it is very important to find a flexible methodology that allows to implement functionalities, test them, and determine whether they should be discarded, terminated, or on the contrary, open a new avenue of research.

On the other hand, taking into account the study framework where the research has to be carried out in a relatively short period of time, around 4-5 months, it has been determined to carry out an agile work methodology where the steps to be carried out are subdivided into sprints of around 2 weeks (time between meetings with the tutor). In each sprint, the aim is to start from the previous sprint, taking into account the different results obtained in order to continue progressing along the best possible path, but focusing on the main and core trajectory of the work.

In each new meeting, the objectives established for the sprint completed are analysed and the results obtained are observed. Depending on the success of both variables, a new assessment is made for the next sprint with a view to the next meeting.

1.4.2 Validation

As it is a research project, the validation of the results is done jointly with the tutor. Meetings are held every 2 weeks to see the changes implemented and how they have worked. An initial validation and assessment of the data is carried out with the results obtained by means of parameterisation values, accuracy of the result, percentage of correctness of the classification, etc.

On the other hand, given that these parameters may sometimes not be significant, due to cases of overfitting, incorrect treatment of the variables generating an overly optimistic classification, etc., a second manual validation is carried out to assess the nature of the result obtained. In this second assessment, confirmation of the results is awaited from the tutor.

Secondly, the programming environment of the code used is in Google Colaboratory, an online development environment where multiple users can access the same environment simultaneously and observe the results of the different classification runs. In this way, a history of the programmed code and the possibility of retrieving previous versions is also maintained. In addition, all the tests performed are fully recoverable and reproducible, thus providing a greater solidity to the results obtained regardless of their nature.

Temporal planning

2.1 Description of the tasks

This project has an approximate duration of 585 hours, spread over 167 days starting from 9 January 2023, date until 25 June 2023. The exact date for the oral defense has not yet been defined, so the deadline is the earliest possible date. It's expected to work approximately 3.5 hours every day. Accordingly, the number of hours devoted to each task in order to ensure the successful completion of the thesis.

2.1.1 Task definition

Throughout this section we will discuss the tasks that will be carried out during the project. In addition, each task is described taking into account the different dependencies with other tasks and the duration of each task. The tasks to be performed have been grouped into modules for a better classification. There are time dependencies between them, for example, in order to start the practical application it is essential to have established a theoretical and contextual basis of the problem to be tackled and so on.

Project planning

Good planning is essential to have an order and a global vision of the different sections of the work. For this reason, the following is a definition of the project management tasks, documentation and work to be carried out are defined below.

- **Contextualization and scope of the project.** Definition of the project, objectives, stakeholders and scope of the project in the context of your study. Approach to the current state of the area and context and how it will be developed.
- **Time planning.** Organization of the structure of the thesis by describing the phases, resources and requirements for its completion.
- **Economic management and sustainability.** Analysis of the environmental and sustainable context of the thesis as well as its economic dimension and cost.
- **Meetings.** Scheduling of regular meetings (every two weeks) with the tutor or those responsible for the work. The aim of the meetings is to guarantee the constant and correct evolution of the study of the project, meeting the deadlines and assessing the results.
- **Integration in the final document.** Grouping and organization of all the documents and tasks carried out throughout the work.

Research

Since the main purpose of the thesis is to investigate LiBs that can be used for second life, knowledge and contextualisation in the framework of this type of batteries is essential. This is why it is necessary to carry out a thorough research process to understand the problem, how it is addressed, where our solution is framed and how to implement it. It is also essential to search for and analyze studies similar to ours that have obtained promising results. Mainly the following tasks will be carried out:

- Analysis of the LiBs, context, attributes and functioning.
- Search for similar information and studies
- Study of possible solutions and their implementation

Practical implementation

For a correct process of experimentation and analysis of results, it is essential to previously program all the architectures involved in the research. Although there are numerous Python libraries that make the work easier, it is advisable to program the scripts according to the specific needs. We divide the practical implementation into the following tasks.

- Learn different ML algorithms. To become familiar with the different ML algorithms, especially SVM and Linear Regression, as well as their characteristics and hyperparameters.
- Program the algorithms for RUL prediction.
- Test the correct operation of the algorithms and programmed networks to ensure the correct operation of each one of them.

Experimentation, analysis and conclusion

A coherent and organized process of experimentation, analysis of results and conclusion is very important for the thesis. Therefore, this module has been organized in the following steps.

- Selecting and preparing the dataset. Choose the dataset to be used and preprocess it in order to optimize it as much as possible to avoid later problems.
- Experimentation. Carry out a process of experimentation and adjustment of parameters and hyperparameters until the optimum ones are found for the programmed architectures.
- Conclusion. Analyze the results obtained through experimentation.

Documentation is an implicit task that has to be performed throughout the thesis. It is important to document each task recurrently in order to have a memory to avoid the rush of the last weeks. The last task of the work is the preparation of the **oral defense**, taking into account the different parts of the thesis that need to be explained.

2.1.2 Summary of the tasks

The following table summarizes the tasks explained above, showing the dependencies between them and also including the number of hours dedicated to each one.

Id.	Task	Time (h)	Dependences
T1	**Project planning**	**30**	-
T2	- Contextualization and project scope	10	-
T3	- Temporal planning	5	-
T4	- Economic management and sustainability	5	-
T5	- Integration in final document	10	-
T6	**Meetings**	**10**	-
T7	**Research**	**100**	-
T8	- Analysis of the LiBs, context, attributes and functioning	30	-
T9	- Search for similar information and studies	50	T8
T10	- Study of possible solutions and their implementation	20	T8,T9
T11	**Practical implementation**	**190**	T7
T12	- Learn different ML algorithms	20	-
T13	- Program the algorithms	120	-
T14	- Test the algorithms	50	-
T15	**Experimentation, analysis and conclusion**	**200**	T11
T16	- Selecting and preparing the dataset	20	-
T17	- Experimentation	150	-
T18	- Conclusion	30	-
T19	**Documentation**	**40**	-
T20	**Oral exposition**	**15**	T15

Table 2.1: Summary of the tasks.

2.1.3 Resources

In order to realize the project there are a number of resources that are necessary.

Human resources

Mainly one human resource is required for the realization of the thesis and that is the researcher, who will be fully involved in the work. On the other hand, both the main tutor Martha Ivón Cárdenas Domínguez and the tutor of the subject GEP Paola Lorenza Pinto will tutor the main researcher in the development of the thesis, so they are also fundamental resources.

Material resources

Although this project is a novel approach, it is necessary to build on and/or document previous studies in the same field. For this reason, scientific articles and papers or books have been necessary to help contextualize the problem. In addition, I will need some software and hardware resources to carry out the practical part and the experiments.

- Google Drive: This digital platform will be used because it offers numerous tools, including a text editor to create the memory of the work and generous storage to classify and maintain various files in the cloud, being easy to collaborate and share with other people. The latter facilitates communication and tutoring by the work tutor.
- Atenea: Used to communicate with the professor in charge of GEP and deliver the work.
- Google Colaboratory: This resource will be used to provide an environment in the cloud where scripts can be programmed and executed in the same environment. Further, it is very useful for structuring, programming online and executing them in the same environment. Moreover, it gives you the ability to share and visualize the code with the tutor by allowing the sharing of the notebook and the editing and execution of the same by different people. This means that no extra material resources or computers with high computing power are needed to run the tests, as they are carried out in an online environment.

2.2 Gantt

Below (figure 2.1) I attach the Gantt chart showing the tasks distributed over the semester, taking into account the number of hours for each task (see 2.1.2) and the dependencies described in the previous section. The start date of the project corresponds to the first meeting with the tutor once the work has been agreed and formalized and, given that I will defend the thesis in the first call, the end date will be the last week of June of this year.
The meetings with the tutors are periodically scheduled every two weeks although, if required, they will be rescheduled according to the needs of the moment. In that case they would be included in the table after a few months of work.

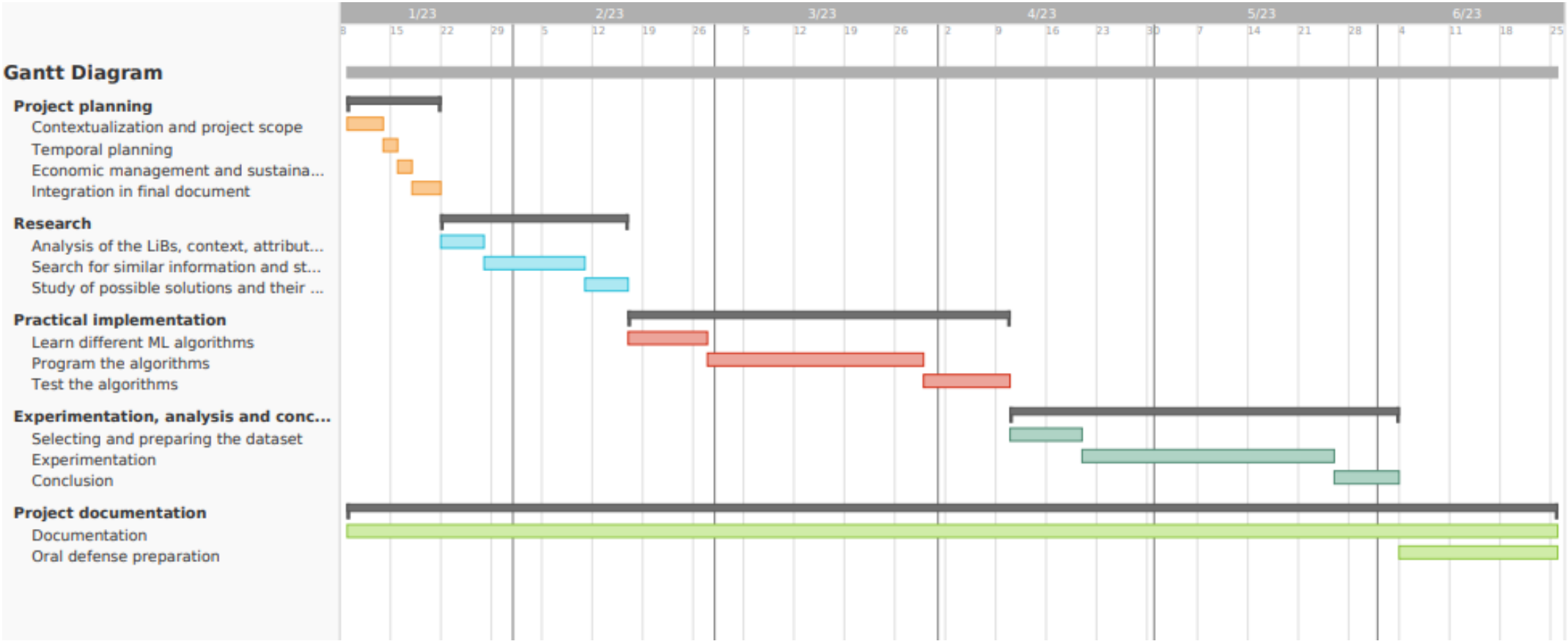

Figure 2.2: Gantt diagram

2.3 Risk management

As we have seen in section 1.3.3, problems may arise during the course of the project that negatively influence the progress of the project. In this section we will present these risks, to what extent and how severely they affect the realization of the project and how to solve them with alternative tasks.

2.3.1 Lack of experience

As this is the first long and complex study carried out by the principal investigator, this may imply inconveniences for the development of the work. In addition, the lack of experience in the subject and the ML models used may also affect the vision and realization of the work.

- **Impact:** Low
- **Proposed solution:** dedicate time to study the context of the work, both the LiBs and how they work and the ML models, their application and analysis. A first period of time, between 3-4 weeks has been estimated for the contextualisation of the topic and the study of previous works. In this way, it is intended to reduce the impact of the lack of experience in these fields. On the other hand, the support of the different tutors for the organization and structuring of the project helps to reduce the lack of experience in the realization of a thesis such as this one.

2.3.2 Limitations dataset

The research utilizes a dataset produced by NASA, containing information on the charge and discharge of various batteries. While this dataset is comprehensive, it is possible that additional attributes may be required to make the desired predictions or that some datasets may not contain sufficient data entries for a thorough analysis.

- **Impact:** Low
- **Proposed solution:** attributes can be calculated from existing attributes, as well as SOH from capacity and cycles. On the other hand, the dataset is organized by similar batteries which can be concatenated respecting the values of each data entry to make a larger data entry to be trained and tested.

2.3.3 Computational risk

In this paper we study different models and prediction techniques using ML. The computational cost of some of these techniques can be a problem if the resources available in the Google Colaboratory environment are exceeded, making it impossible to run some models. Although the resources provided by this tool are very extensive, some models that train with many layers of depth or with large data structures may not be finalized due to the lack of computational power of this tool.

- **Impact:** Medium
- **Proposed solution:** if required, the runtime environment can be programmed and computational power can be extended in the Google Colab environment by extending resources provided by the computer itself. If necessary, this extension can be carried out for the execution of a prediction.

2.3.4 Deadline

One of the main problems in the realization of this project is the estimation of the hours spent. I may encounter underestimation of the hours to be spent or problems of an external nature beyond the control of the actors involved in the work that delay the time needed for the realization of the project.

- **Impact:** Medium
- **Proposed solution:** adapt the project plan to the new situation. For a correct approach in each meeting with the tutor, a review of the point of the work where we are with respect to the planning will be made and the short-term objectives will be adjusted in case of problems or delays. A diagnosis of the problem and rapid action and restructuring is essential. This can be of vital importance for the completion and review of all branches of the work.

Budgets

In the following section the economic cost of the thesis will be studied. For this purpose, a study will be made of the cost of the personnel necessary to carry out a project such as this one, the generic costs and the indirect costs. Finally, the control mechanism used for possible deviations in the initial budget will be analyzed.

3.1 Staff costs

The following is an estimate of the personnel cost required for this project. First of all, in order to make a proper estimation, it is necessary to define in advance the functions
(the necessary resources are described in section 2.1.2).
The salary of the workers will be calculated by establishing an hourly rate and multiplying it by the number of hours they will work on the project. The cost of a task will be calculated as the sum of all the costs necessary to perform it.

We have defined a total of roles to be played in the work to carry out a distribution of the work and responsibilities of the same. First of all, a project manager is required, who is responsible for the planning and supervision of the project. Secondly, a junior researcher is also required, who will perform study tasks in relation to the LiBs and their characteristics. Thirdly, an analyst is required, who will study the data provided by the researcher. Finally, a junior programmer and a tester will be needed, to program the different algorithms and scripts and the tester to check the correct operation of them.

The role played by the project manager is the one developed by the tutors but the rest of the roles will be played by myself. Once the necessary roles have been defined, let us see in the following table the current cost of these profiles.

Role	Cost(€)/h
Project Manager	23 [12]
Junior Researcher	13 [13]
Junior Developer	13 [14]
Tester	14.5 [15]
Analyst	13.5 [16]

Table 3.1: Cost per hour of the different roles.

Then, are defined the costs of each process (see 2.1.2).

Task	Hours	Project Manager	Junior Research	Junior Developer	Tester	Analyst
Project planning	30	30	0	0	0	0
Meetings	10	10	0	0	0	0
Research	100	0	80	20	0	0
Practical implementation	190	0	0	170	20	0
- Learn different ML algorithms	20	0	0	30	0	0
- Program the algorithms	120	0	0	140	0	0
- Test the algorithms	50	0	0	0	20	0
Experimentation, analysis and conclusion	200	0	0	50	50	100
Documentation	40	0	20	0	0	20

Table 3.2: Estimated time per task.

Task	Cost(€)
Project planning	690
Meetings	230
Research	1300
Practical implementation	2500
- Learn different ML algorithms	390
- Program the algorithms	1820
- Test the algorithms	290
Experimentation, analysis and conclusion	2050
Documentation	530

Table 3.3: Estimated cost per task.

Finally we calculate the cost of hiring, known as CPA. In this way we can know the overall cost of the personnel needed to carry out the project, by multiplying the hours of each worker by the hourly price that has been estimated.

Role	Hours	Total Cost(€)
Project Manager	40	920
Junior Researcher	100	1300
Junior Developer	240	3120
Tester	70	1015
Analyst	120	1620
Total	570	7975

Table 3.4: Total cost of the staff.

3.2 Generic costs

Amortization

To calculate the overall economic cost of the whole project, the amortization of the different resources used must also be taken into account. In this case, since no special software has been used and all the research material is open source, the main resource used has been the computer used for the research, writing, programming and analysis of the thesis. We estimate that we worked an average of 167 days, an average of 3.5 hours per day. Taking these data into account and considering the formula for the calculation of the amortization we have the following:

$$Amortization = ResourcePrice * \frac{1}{YearsOfUse} * \frac{1}{DaysOfWork} * \frac{1}{HoursPerDay} * HoursUsed$$

$$Amortization = 600 * \frac{1}{4} * \frac{1}{167} * \frac{1}{3.5} * 585\ hours$$

Hardware	Cost(€)	Life expectancy (years)	Hours used	Amortization (€)
ACER laptop	600	5	585	150,13

Table 3.5: Amortization of the hardware used.

Indirect costs

In this section we will also analyze indirect costs that have an influence on the overall performance and development of the thesis. In the case of this project it is done independently and by remote work. Therefore, the cost related to working from home will be described.

- Internet: The cost of an internet tariff in Spain, where the thesis will be carried out, at the date of the thesis may be around 90€. Taking into account that the project lasts 6 months and that the working hours per day are 3.5 hours, the cost of the internet is 6 months * (90€/month) * (3,5h/24h) = 78,75€.
- The price per kwH [17] is 0.1638 €/kWh. Assuming that the power of the laptop we will use is around 200 W the total cost would be (0.1638 €/kWh) * 0.2kW * 167 days * 3.5h = 19,5€.

Generic cost of the project

Below is a summary of the generic cost of the job, including amortization of resources as well as indirect costs.

Description	Cost (€)
Amortization	150,13
Internet	78,75
Electricity	19,15
Total	248,03

Table 3.6: Generic cost of the project

3.3 Deviations of the budget

Contingency

In any project, unforeseen events may arise that can affect the plan and generate additional expenses. Therefore, it is important to have a contingency fund to deal with these unexpected expenses and avoid delays in project execution. It is important to keep in mind that some project costs are more likely to increase than others. For example, personnel costs may increase if there are delays in planned working hours. To prevent this, a 10% contingency margin is included in the total project cost.

On the other hand, the overall project costs may be more stable due to the time constraint and, therefore, a 5% contingency margin is defined to cover any unforeseen events.

In summary, having a contingency fund and planning adequately is essential for any project, as it allows us to deal with unforeseen events that may arise and to avoid delays and additional costs.

Incidental costs

In section 1.3.3 above, all possible risks that may affect the project have been defined. The following table (3.3) quantifies these risks in order to include the cost of implementing alternative plans in the final budget. alternative plans in the final budget.

Incident	Estimated cost(€)	Risk(%)	Cost(€)
Lack of experience (30h)	390	25	97,5
Limitations dataset (20h)	260	25	65
Computational risk	25	50	12,5
Deadline	0	20	0
Total	675	-	175

Table 3.7: Incidental costs.

Final budget

After having considered all of the above aspects, it is important to calculate the final cost of the project by integrating each of the sections. Table 3.3 shows the calculation of the total cost of the project.

Activity	Cost (€)
PCA	7975
CG	248,03
Contingency margin	797,5
Incidential cost	175
Total	9195,53

Table 3.8: Final budget

3.4 Management control

Although the budget for possible occurrences has already been defined in the previous section, it is still necessary to explain how any deviation from this budget will be managed during the project. In this sense, the procedure to be followed to detect any change in the project cost is presented below.

During the execution of each task described in the planning section, its individual deviation from the estimated cost will be calculated using the formulas presented below.

- Estimated cost: Estimated cost of the task (see 3.1).
- Real cost: To verify if there have been any incidents that may have increased the cost of each task, it is necessary to recalculate the actual cost of the task. To do this, the PCA, the CG, the contingency and the incidences of each task will be recalculated, in order to detect any error in the initial planning and to be able to readjust it accordingly.
- Deviation: Difference between the estimated cost and the real cost of the task.

The variance factor is an indicator that shows how much the actual cost has changed compared to the initial estimate. If the difference is positive, it means that the cost of the task has been overestimated, so the excess money can be reallocated to cover other project issues. On the other hand, if the difference is negative, it is necessary to allocate part of the contingency fund to that task to cover the variance.

Sustainability

4.1 Self assessment

Today, the world is in a critical situation due to the continuous degradation of the environment, aggravated by humanity's lack of awareness and responsibility. Pollution and climate change are just some of the negative effects of our impact on the planet, making it essential to adopt a more sustainable approach. This change not only involves taking measures to reduce our ecological footprint, but also considering the social and economic impact of the projects we undertake.

In this context, new technologies are opening up new possibilities for addressing the challenges we face as a society. These tools are helping to foster social change and empower minorities, which makes it important to consider the social impact of projects such as these. Together with economic evaluation, these elements make up the sustainability of a project. Sustainability assessment is therefore critical to ensure that projects are environmentally, socially and economically viable.

This project is approached from the point of view of climate emergency awareness and technological waste. It proposes the reduction of a highly demanded technological resource with an impact on the planet both in production and in the use of the same and it proposes the maximization of the use of these batteries to reduce the negative effects they have on our planet.

4.2 Environmental impact

Regarding the PPP, have you thought about the environmental impact of your project? Have you considered minimizing this impact, by for example reusing resources?
Given the research focus of my thesis, it is complex to determine its possible environmental impact. The only aspect that could have some impact on the environment is the energy consumption generated by the training process of the different models used. However, it is contemplated to use a grouping in the data structures so that the data sets are smaller than the original ones, thus decreasing the time needed for training and processing them, consuming less energy and decreasing the environmental impact.

Regarding life expectancy, how is it solved the problem you are trying to solve? Does your solution provide any improvement in the environmental impact?
Indeed, the aim of this thesis is precisely to reduce the environmental impact that LiBs generate on the planet. The aim is to obtain a fast and efficient sorting method to greatly extend the life expectancy of an electric battery which consumes resources and affects the planet's footprint both in production, use and recycling. A positive outcome of this project would help to reduce the environmental impact of a particularly polluting and environmentally aggressive industry.

4.3 Economy

Regarding the PPP, have you estimated the impact that your project will have, including both human and material costs?
Section 3 of this project describes the consequences that this project may have, taking into account both human and material costs. In addition, possible deviations that may arise during the development of the project are considered.

Regarding life expectancy, how is it solved the problem you are trying to solve? Does your solution provide any improvement economically?
Yes, the approach of reusing LiBs provides a significant economic improvement compared to the production of new batteries. By reusing batteries, the cost of extracting the materials and processing raw materials needed to manufacture new batteries is avoided, as well as the cost of disposing of old batteries. In addition, the process of reusing batteries is much faster and less expensive than the production of new batteries, resulting in significant cost savings. Not only does it save on the non-production of new resources, but it also saves on the treatment required for discarded batteries.

4.4 Social

Regarding the PPP, how do you think this project will enrich you personally?
Clearly, this is the first time I have done a research project of this style and although it can be overwhelming at first, it ends up being very rewarding. This project gives me the opportunity to learn about all the aspects involved and all the considerations involved in a research thesis and how to approach it. It also gives me the opportunity to put into practice all the knowledge acquired throughout the degree of computer engineering, both the structure, development, etc.

Regarding life expectancy, how is it solved the problem you are trying to solve? How do you think your solution will improve people's quality of life? Is there a real need for developing your solution?
The proposed solution will affect people's lives mainly in an indirect way. More directly they will probably be able to get good quality LiBs with sufficient lifetime ahead for cheaper secondary uses, as well as electric scooters, energy storage systems from renewable sources, etc. Indirectly, we have already justified how this project aims to reduce the environmental impact caused by these batteries. We will all be affected as well as, for example, less air pollution and exploitation of material resources.
It is clear that it is necessary to develop this solution. There is currently no similar alternative in the sector, and we have observed how polluting these manufacturing and waste treatment processes are. This is an industry that is not going to cease production but we must change the way we understand and use resources. It is necessary to implement such a solution for the good of all.

The fundamentals

This section will briefly introduce fundamental concepts of artificial intelligence and machine learning, which are necessary to correctly understand the following parts. Later sections will describe and explain complex structures and architectures based on these concepts.

5.1 Fundamentals of artificial intelligence, machine learning and its applications

We can commonly define artificial intelligence (AI) as a set of systems or combination of algorithms, whose purpose is to create machines that mimic human intelligence to perform tasks and can improve as they gather information. One of the most promising branches of AI is ML, which can be a bit more complex. We can define machine learning as the science (and art) of programming computers to learn from data [18]. We can understand it with a more precise definition: "*A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.*" —Tom Mitchell, 1997 [18].

AI and ML are based on the use of algorithms and computational techniques inspired by the human ability to learn, reason and make decisions. They rely on mathematical, statistical and computational approaches to process large amounts of data, identify patterns and relationships in them, and make predictions or inferences based on these patterns. ML models are trained with input data and expected outputs, and from these data, they adjust their internal parameters to improve their performance on specific tasks without being explicitly programmed for each situation. In this way, AI and ML allow computers to learn and adapt to new data and situations autonomously.

The ability to process large amounts of data is essential in AI and ML. Data is used to train and validate models, and subsequently obtain knowledge and patterns from them. Moreover, mathematical factors such as statistics and probability are fundamental in these techniques. These factors allow analyzing and modeling uncertainty in the data and in the predictions made by the models. These are also based on other aspects such as optimization, used to adjust the parameters of ML models and improve their performance in specific tasks, where the main objective is to found the optimal values of the parameters of a model, the parameters which minimize or maximize an objective function, using various mathematical algorithms (such as gradient descent, genetic algorithms, among others).

AI and ML have numerous utilities and cover a wide range of sectors and application areas. These technologies are already present in numerous fields, from medicine and healthcare, industry, logistics, entertainment, marketing, etc. These techniques are used to automate tasks, improve processes, optimize decisions, personalize experiences, predict behaviors and trends and analyze data among other functionalities. These technologies continue to evolve and expand into new domains, offering unlimited potential to transform and improve the way activities are performed in virtually any sector or industry.

5.2 Supervised and unsupervised learning

Supervised and unsupervised learning are two machine learning techniques used for processing data and obtaining information from them. In the following section we will briefly explain what each one consists of in order to justify why we have decided to use one of them in this research and to be able to understand more complex explanations presented later.

In the first instance, supervised learning is a training method in which labeled input and output data are used to train predictive models [19]. In this approach, data are pre-labeled before being used to train the model. The goal is for the model to learn to predict the correct output from new and previously unseen data. This approach is based on the use of labeled data to learn patterns and relationships between inputs and outputs. Therefore, human supervision is required to prepare and label the input data. In this type of training we know the target variable and we intend to decide (classify) on it.

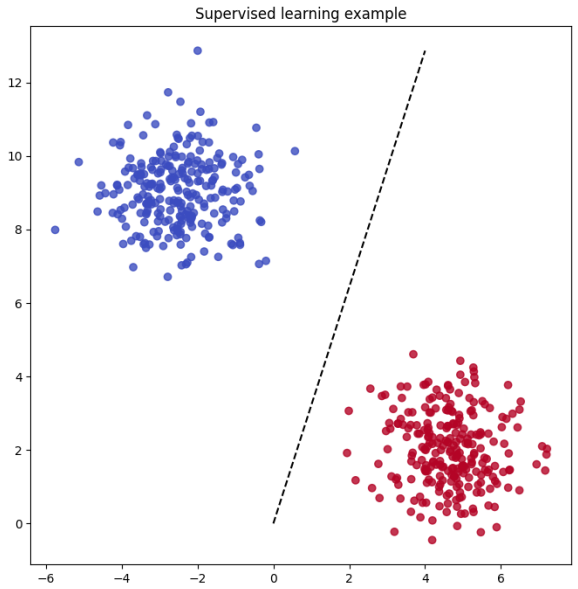

Figure 5.1: Supervised learning plot example (own elaboration)

On the other hand, unsupervised learning is a training method in which unlabeled data is used to learn patterns and relationships within the data [19]. In this approach, the model has to find patterns and structures in the data without any prior supervision. The goal is to identify patterns and structures within the data without knowing the correct answer. The approach is based on the use of unlabeled data to learn patterns and relationships between inputs. This technique, as we can imagine, requires less human supervision than supervised learning, since we don't have to classify the data before to do the study and the researcher don’t need to have a lot of knowledge about the context before training the model.

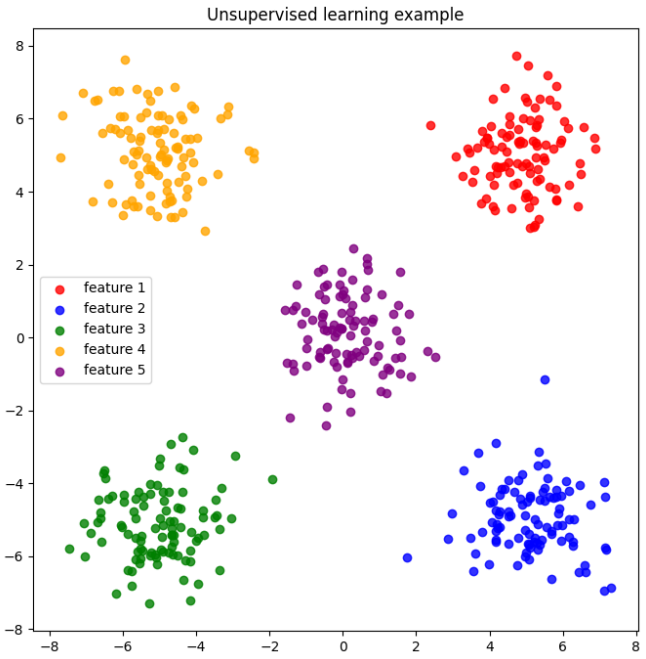

Figure 5.2: Unsupervised learning plot example (own elaboration)

Supervised learning is used in cases where labeled data is available and you want to predict an output (which is limited by a subset of possible values) from new data. For example, it is used in applications such as spam detection, image classification and price prediction. Instead, unsupervised learning is used in cases where you want to explore the data and discover hidden patterns and structures. It is used in applications such as customer segmentation and anomaly detection.

Each approach has its advantages and disadvantages. Supervised learning has the advantage that it can predict the correct output from new and previously unseen data. In the case of the research that we have done, supervised learning has been implemented. As previously stated, the main idea was the classification of LiBs for second life, the idea is to predict whether a LiB is optimal or not to be used in another context. Therefore the model that fits this idea is supervised learning, from the classified data that have been labeled in the dataset we try to adjust a model able to predict for new input data.

5.3 Regression and classification models

This section will explain classification and regression models, which are two supervised learning techniques used to predict output values from input values. Classification models are used to classify data within a set of different categories, while regression models are used to predict a numerical value.

First, as already introduced, regression models provide a numerical outcome, the prediction of that value is based on a range of possible outcomes. This statistical modeling technique is used to predict numerical values of a variable of interest as a function of one or more explanatory variables. In mathematical terms, regression consists of fitting a mathematical function to the observed data in such a way that the value of the dependent variable can be

predicted for any value of the independent variable [20]. The objective of regression is to find the best function that explains the relationship between the variables, which is achieved by minimizing the distance between the observed values and the values predicted by the model.
To evaluate the performance of a regression model we usually use the prediction error measure, also known as "loss", which measures the discrepancy between the observed values and the values predicted by the model. It is based on observing how much the predictions deviated from the actual values. The choice of error function depends on the type of variable being predicted and the type of regression model being used.

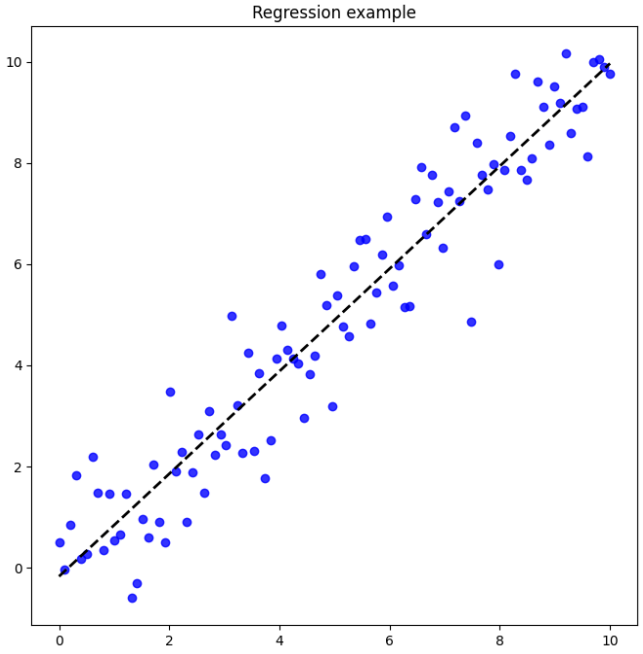

Figure 5.3: Regression plot example (own elaboration)

On the other hand, classification models seek to fit each sample into a discrete class, within a finite set of possible classes. These models are based on predicting the probability that each sample belongs to each of the available classes. In mathematical terms, classification consists of fitting a statistical model to the observed data in such a way that the class of an unknown sample can be predicted based on its characteristics [20]. The goal of classification is to find the best function that explains the relationship between the characteristics of the samples and the corresponding classes, which is achieved by using machine learning algorithms and statistical techniques.
To evaluate the performance of a classification model, various evaluation measures are used, such as precision, recall or F1-score, which measure the model's ability to correctly classify samples. Accuracy measures the proportion of samples that have been correctly classified out of the total number of classified samples, while recall measures the proportion of positive samples that have been correctly classified as such. The F1-score is a measure that combines precision and recall to provide an overall measure of model quality. The choice of evaluation measure depends on the type of classification problem being addressed and the characteristics of the data used.

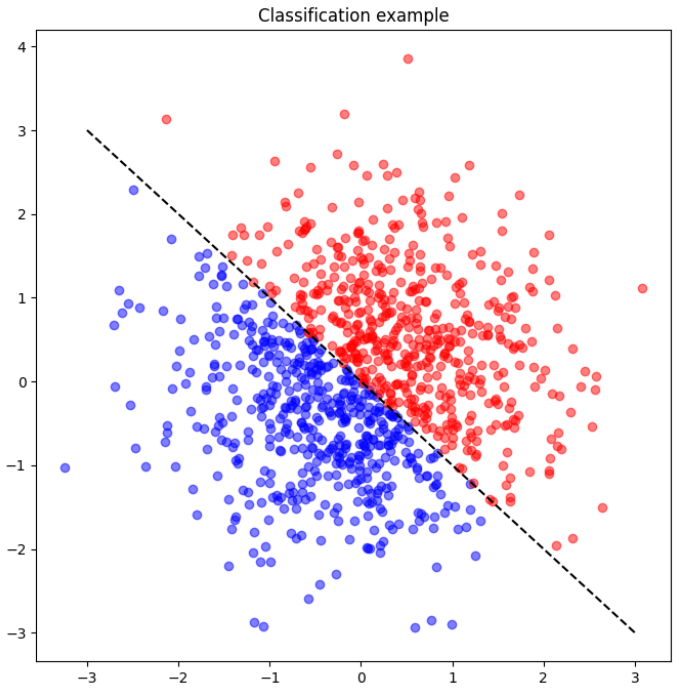

Figure 5.4: Classification plot example (own elaboration)

This project is intended for a classification model. The main objective is to classify in different classes, defined in a finite set, such as, for example, "not reusable", "human review", "short term reusable", "long term reusable". It is not intended to predict a value or to magnify how reusable the LiB is, but to classify whether it really is or not.

5.4 Model evaluation

The evaluation of the classification model is a crucial stage of the supervised learning process. Its main objective is to determine how well the model can classify new instances after being trained with a training dataset. The evaluation process is based on comparing the predictions made by the model with the real values of instances in a test data set. For this purpose, several evaluation measures are used, such as those already discussed above: precision, recall, r-squared, etc. that measure the ability of the model to correctly classify the samples and adapt to the data.

In addition to these evaluation measures, it is also important to take into account other aspects that can affect model performance, such as overfitting and underfitting. The overfitting occurs when the model fits the training data too well and isn't able to generalize to new data. If we consider the overfitting in mathematical terms, it can be measured by the difference between the training error and the validation error. If the difference is significantly high, then the model may be suffering from overfitting [21].
On the other hand, underfitting occurs when the model is too simple to capture and understand the complexity of the problem and, therefore, fails to adequately fit either the training or test data. In other words, underfitting occurs when the model is unable to explain the behavior of the data, resulting in poor performance on both training and validation data. In this case, the model underfits the data and fails to generalize adequately to new data not seen during training [21]. From a mathematical point of view and similar to overfitting,

underfitting can also be measured by comparing the training error to the validation error, if both errors are high, then the model may be suffering from underfitting.

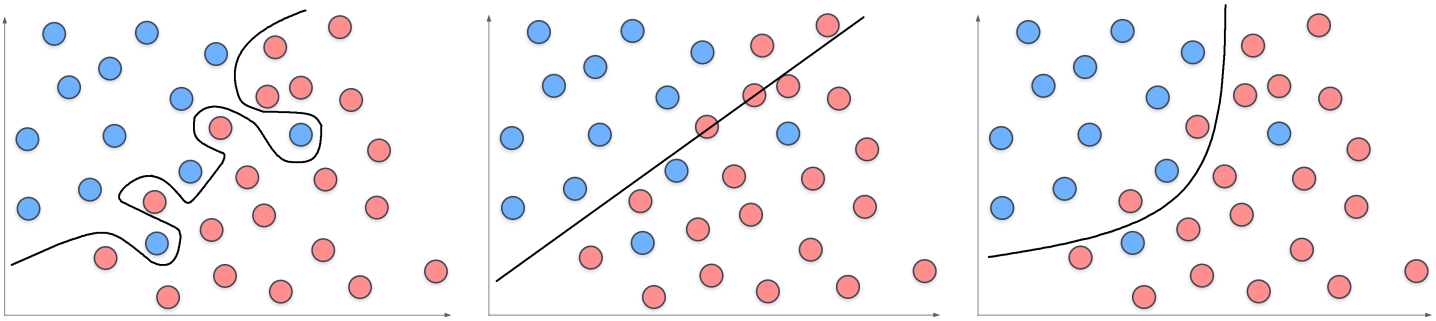

Figure 5.5, 5.6, 5.7: Example of overfitting (too good to be true), underfitting (too simple to explain the variance) and proper fit, respectively (own elaboration)

To avoid overfitting and underfitting, we can use techniques such as cross-validation and hyperparameter fitting. Cross-validation is a technique that divides the data set into k subsets and performs k iterations of training and testing, where in each iteration it is trained with k-1 subsets and tested with the remaining subset. This provides a more trusted assessment of model performance. In the second instance, hyperparameter adjustment refers to the search for the optimal values of the model parameters, which can significantly affect to the model performance, adaptability and understanding. These two techniques have been used at different points of the research to observe the behavior of the model as well as the different results obtained [21].

It is important to keep in mind that model evaluation is a process that requires supervision, it is not a process that can be automated and the context of the study must be known since it is necessary to make informed decisions on the selection of evaluation measures and validation techniques appropriate for each specific problem. In addition, it should be emphasized that model evaluation is an iterative process, where the model can be retrained and adjusted with new data or techniques to improve its performance.

Experimental models used

This project is based on the research carried out, as any research, several algorithms have been tested in order to observe the results, analyze them and compare them in order to choose the best one. Throughout this section we will briefly explain the algorithms used, as well as their adaptation to the model.

6.1 Support Vector Machine

The SVM (Support Vector Machine) algorithm is a supervised machine learning method used for both classification and regression techniques. It is a technique that seeks to find the hyperplane that best separates the data in a high dimensional space [22].

In its simplest form, SVM is used to find a hyperplane, i.e. a separation surface used to classify the data that optimally separates the dataset under study into two classes. This optimal margin is found by means of a mathematical optimization function that seeks to minimize the classification error and maximize the distance between the closest observations of each class.

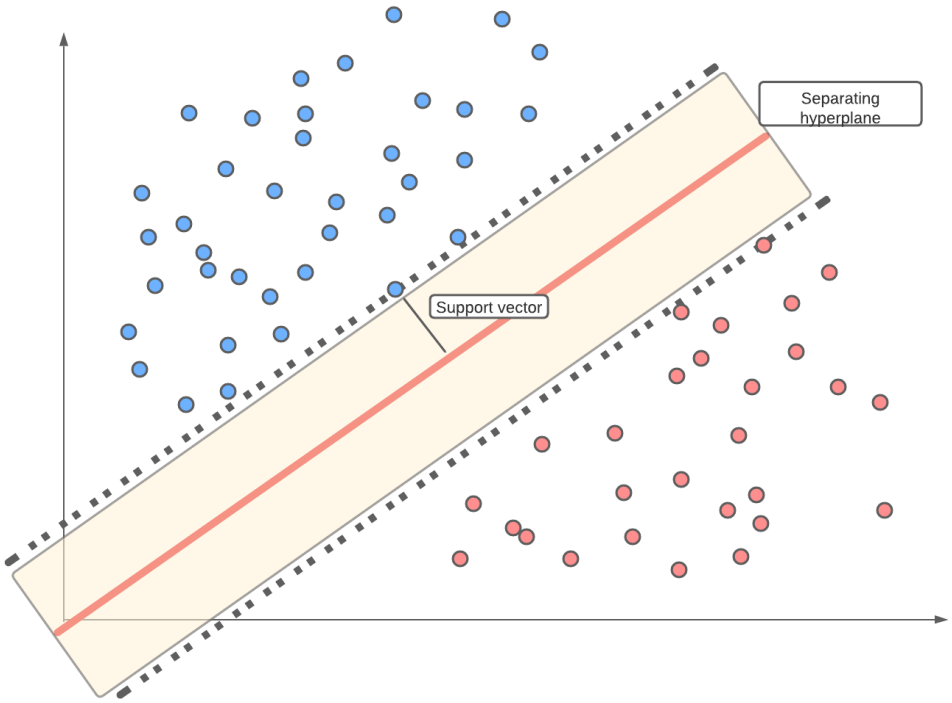

Figure 6.1: Example of SVM algorithm works (own elaboration)

As far as the project is concerned, we have to make special emphasis on the SVM algorithm focused on regression, or also called SVR. This has been the main algorithm used to perform the prediction. It is based on the idea of finding an optimal regression function that fits the input data, for which SVR uses a set of support vectors that define the optimal margin of the regression function. The main objective is still to find the function that has the largest possible margin, but still fits the input data adequately.

In mathematical terms, the goal of the SVR function is to find a hyperplane in a high-dimensional space that has the maximum possible margin while minimizing the number of samples that fall outside the margin. This hyperplane can be defined by the equation:

$$f(x) = w^T \cdot x + b$$

where *x* is a feature vector, *w* is a vector of weights and *b* is a bias or compensation term. The goal of the SVR loss function is to minimize the difference between the predicted output f(x) and the desired output y(x), while keeping the margin between the samples and the hyperplane as large as possible [23].

We also find other important aspects in this algorithm, such as the kernel. This is a function used to measure the similarity between two points in the feature space. The goal of the kernel is to map the input data into a higher dimensional feature space, which can make the data more separable.
On the other hand, the hyperplane is a key concept in SVR. It is a high-dimensional subspace that is used to separate the data points into two groups, one for data points that fit the regression function and one for points that do not fit. The goal of the algorithm is to find the hyperplane that maximizes the distance between the hyperplane and the nearest data points, known as support vectors, also known as the optimal hyperplane.

6.2 Random Forest

The Random Forest (RF) algorithm is a supervised machine learning method used for both classification and regression techniques. It is a technique that combines multiple decision trees and combines their estimates to produce a final prediction [24].

A decision tree is an ML model based on the idea of dividing the data set into smaller subsets recursively, until the data in each subset is as homogeneous as possible. Each subset is divided according to the variable that best separates the classes or explains the variance of the target variable.
The decision tree starts with a root node that represents the entire data set. The data set is then divided into two or more subsets, depending on the most important variable, the target of our study, into which we want to divide it. This division is done in such a way that homogeneity within each subset is maximized. Then, this splitting process is continued in each subset created, recursively, until a stopping criterion is reached, such as a maximum depth of the tree or the minimum number of examples per leaf.

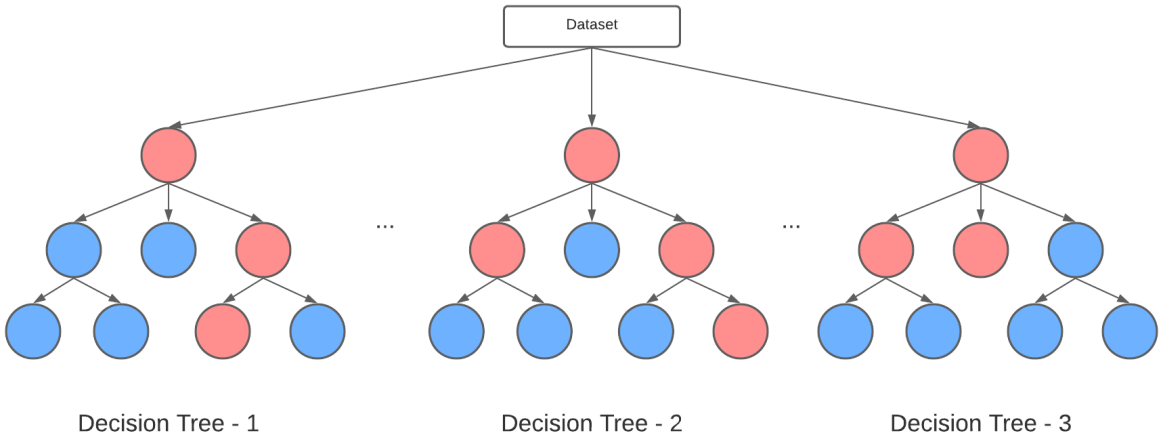

Figure 6.2: Example of RF algorithm distribution (own elaboration)

This algorithm works by constructing a set of decision trees in which, instead of using a single decision tree, many different trees are used. Each tree is constructed using a random sample of the training data and a random selection of features. To construct each tree, the algorithm uses a technique known as bagging (bootstrap aggregating). Bagging is a sampling technique used to construct multiple decision trees from random subsets of the original training data set. Each decision tree is constructed from a different subset of the training data set, randomly selected with replacement. The term "bootstrap" refers to the fact that a training data set is created for each decision tree by randomly sampling with replacement from the original data set.

Once all the trees have been constructed, the Random Forest algorithm uses the "voting" technique to produce a final prediction. In the case of regression, which is the one used for this study, we find how each tree in the algorithm set predicts a numerical value for a data instance. Once all the trees have been constructed, the mean of all the numerical values predicted by the trees for a specific data instance is calculated and it is this mean that becomes the final prediction value of the model for this data instance. This technique aims to improve the accuracy and robustness of the model by combining the results to reduce the effects of outliers and the variability inherent in the prediction using a single decision tree. In addition, the diversity in the trees generated by the random selection of training samples and features during tree construction helps to avoid overfitting and improves the model's ability to generalize to new data.

6.3 Neural networks

The Neural Network (NN) is a supervised machine learning algorithm used for both classification and regression techniques. It is based on the simulation of the human brain, as it is composed of layers of interconnected neurons that work together to process information.

Each neuron in a neural network takes an input and produces an output. The inputs are multiplied by synaptic weights, which are numerical values associated with each connection between neurons that determine the relative importance of the inputs to the neuron's output.

Initially, these weights are set randomly and then adjusted by the neural network training algorithm to improve prediction accuracy. During the training process, the neural network receives a set of input data along with their corresponding desired outputs, and the synaptic weights are adjusted to minimize the difference between the outputs predicted by the neural network and the desired outputs [25].

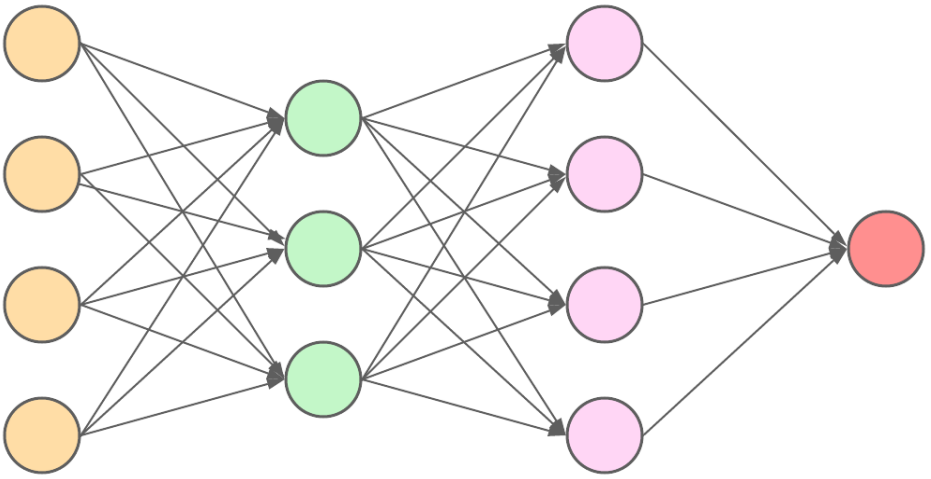

Figure 6.3: Example of NN algorithm distribution (own elaboration)

The neural network begins with an input layer that represents this input data followed by a series of hidden layers that process the data and an output layer that produces the final prediction. During training, the weights of each connection are iteratively adjusted to minimize the cost function, which measures the difference between the network's prediction and the desired output. A neural network can have millions of neurons and millions of connections between them, making it a powerful and complex machine learning model. Neural networks are able to learn patterns in the data and can be used for a wide variety of tasks and adjust the hyperparameters of the network as well as its depth and connections.

SOH and RUL

Throughout this section the concepts of SOH and RUL and their characteristics will be explained. The main objective of this section is to study the origin of these attributes and to argue their importance and weight in the study.

7.1 Definition and concept

In the context of LiBs, SOH (State of Health) refers to the current condition of the battery relative to its original storage capacity. For a more technical definition of SOH in our context, we can consider it as: "the current capacity of the battery relative to its rated capacity after having gone through a certain number of charge and discharge cycles, as well as state-of-charge (SoC) storage and age" [26].

From a chemical point of view, the SOH of a LiB is related to its ability to store and release electrical charges through electrochemical reactions within it [8]. With time and use, batteries undergo changes in their chemical and physical structure, which can result in a decrease in their storage capacity. This is due to the formation of deposits on the electrodes, accumulation of degradation products, loss of active surface area, ionic migration and other processes that can reduce battery efficiency and capacity. In addition, SoC storage, for example, the level of charge at which a battery is stored when not in use, can also affect SOH [26]. For example, keeping a battery in a high or low state of charge for extended periods of time can lead to undesirable chemical reactions that contribute to battery capacity degradation.

On the other hand, RUL in the context of LiBs refers to the estimated life remaining in the battery before its capacity drops below an acceptable level, implying that the battery is no longer suitable for use. Analogous to SOH, for a more technical definition of RUL in our context, we can consider it as: "the time remaining until the battery capacity drops below a predefined threshold, as a function of the number of charge and discharge cycles, SoC storage and age".

Analogously the concept of RUL also has a chemical reason, the RUL is calculated with the last cycle for which the battery still has the capacity to store energy, when a LiB reaches a very low charge threshold, it means that the amount of lithium ions available to store in the electrodes is insufficient. This is because the electrodes may be covered with lithium dendrites (lithium metal deposits on the negative electrode), which decrease the active surface area available for electrochemical reactions, or because chemical degradation products and side reactions have decreased the storage capacity of lithium ions in the electrodes and electrolyte [27]. At this point, the battery cannot be recharged again because there are not enough ions available to participate in the chemical reactions necessary for charging. This marks the end of the effective lifetime of the battery, as its energy storage capacity has been significantly reduced due to the chemical processes that have taken place inside it.

7.2 Methods for SOH and RUL calculation

The SOH calculation is very important to get an idea of the battery's state of health in each cycle and to observe its evolution. There are several methods to calculate the SOH; however, there are two main methods. The first one, based on the battery impedance [28] and the second one based on the battery capacity [29]. For this work, we have relied on the same argumentation as in the scientific study "*Battery State-of-Health Estimation Using Machine Learning and Preprocessing with Relative State-of-Charge*" [21] where they argue that the method defined by battery impedance is not suitable for a project like the one underway because it requires instruments such as electrochemical impedance spectroscopy, which can be costly and complex to implement in certain contexts. In addition, this method may also require expert knowledge in electrochemistry and battery opening, which may be invasive and not suitable for some practical applications. Therefore, in this study, we use the SOH of a battery based on its useful capacity. This method can be expressed as follows:

$$SOH = \frac{C_{usable}}{C_{rated}}$$

where C_{usable} is the usable capacity which represents the maximal releasable capacity when it completely discharged, while C_{rated} is the rated capacity, which is provided by the manufacturer. The usable capacity declines over time. We can also express the SOH as the percentage of the current capacity of the battery with respect to its initial capacity, as shown in the following formula:

$$SOH(\%) = \frac{C_{usable}}{C_{rated}} \cdot 100$$

On the other hand, the numerical calculation of the RUL can also be conceived in different ways. For this work it has been based in turn on another scientific paper *"A novel multistage Support Vector Machine based approach for Li ion battery remaining useful life estimation"*. We can see the formula below:

$$RUL = \frac{N_{EOL} - N_i}{N_{EOL}}$$

where N_{EOL} represents the last cycle where the battery can still be recharged, the threshold cycle where the battery can no longer be used again, while N_i is each of the different charge cycles. Analogous to the SOH, we can express the RUL as a percentage which indicates the remaining life with the following formula:

$$RUL(\%) = \frac{N_{EOL} - N_i}{N_{EOL}} \cdot 100$$

7.3 SOH Importance

The state of health is a fundamental parameter in lithium-ion battery monitoring, as it allows, in part, to determine the aging level of the battery and, therefore, its ability to store energy. SOH prediction is a critical issue in many applications, as battery degradation can have a significant impact on system lifetime and performance.

The paper *"Battery State-of-Health Estimation Using Machine Learning and Preprocessing with Relative State-of-Charge"* [30] highlights the importance of SOH prediction in energy storage systems. According to the authors, accurate SOH prediction can help maximize the efficiency and lifetime of batteries, reduce maintenance costs, and improve the safety of power systems.
Another important aspect of SOH in lithium-ion batteries is its relationship to battery performance. As seen in other studies [8], SOH degradation can have a very significant impact on battery performance by reducing its capacity and energy efficiency. Therefore, SOH prediction and maintenance are critical to ensure the optimal performance of lithium-ion batteries.

7.4 RUL Importance

RUL is a fundamental parameter in power system monitoring, it indicates how much time a LiB has left before it becomes unusable. The prediction of this characteristic is important in many fields as it allows to anticipate failures, reduce maintenance costs and increase the safety and efficiency of the systems.

From a theoretical point of view, the prediction of the RUL is determinant given the limited lifetime of LiBs and the fact that their performance degrades over time due to the factors discussed above. Knowing the RUL of a component can help engineers make informed decisions about its maintenance and replacement, which in turn improves system reliability and safety. In addition, predicting RUL can also be very important because it allows users to plan the life of a system more efficiently and effectively.

Being able to know this factor can help to plan or redistribute LiB discarded for some functions but still suitable for others, which after all, this is the main objective of the project. To be able to give another function to a LiB that is no longer suitable for its use in a context for security reasons but that can be used for other purposes or functionalities. This fact is what they highlight in the paper "*Review of Prognostic Methods for Battery Health Management*" [31] where they discuss the importance of RUL prediction in the health management of LiBs, as it allows users to know in advance when a battery may fail and how much time they have before it fails.

Experimentation

Throughout this section we will contextualize the experimental tools used as well as the nature of the data, the data preprocessing, the training and the models obtained. The analysis made and all the decisions taken that affect the development of the work will also be justified.

8.1 Description of the data set

For this study, the data set collected by NASA has been used to study the behavior of LiBs for further studies. According to NASA's web page [32] this data set has been collected from a custom built battery prognostics testbed at the NASA Ames Prognostics Center of Excellence (PCoE). Li-ion batteries were run through three different operational profiles (charge, discharge and Electrochemical Impedance Spectroscopy) at different temperatures. Discharges were carried out at different current load levels until the battery voltage fell to preset voltage thresholds. Some of these thresholds were lower than that recommended by the OEM (Original Equipment Manufacturer) (2.7 V) in order to induce deep discharge aging effects. Repeated charge and discharge cycles result in accelerated aging of the batteries. The experiments were stopped when the batteries reached the end-of-life (EOL) criteria of 30% fade in rated capacity (from 2 Ah to 1.4 Ah).

The choice of this data set is due to the fact that this is the most complete and comprehensive data collection in the context of LiBs that exists. In addition, the data set has numerous battery types and different behaviors throughout the aging period that greatly enrich the possible study.
Likewise, the purpose of the data collection is in great part to carry out a study of the characteristics of the one we have carried out. All the scientific articles cited and referenced throughout the work that have made predictions about LiBs have done so based on this data set and NASA itself, in the final section of its website [32] states the main objective of this collection: “The data sets can serve for a variety of purposes. Because these are essentially a large number of Run-to-Failure time series, the data can be set for development of prognostic algorithms. In particular, due to the differences in depth-of-discharge (DOD), the duration of rest periods and intrinsic variability, no two cells have the same state-of-life (SOL) at the same cycle index. The aim is to be able to manage this uncertainty, which is representative of actual usage, and make reliable predictions of Remaining Useful Life (RUL) in both the End-of-Discharge (EOD) and End-of-Life (EOL) contexts.”

The data are structured according to the type of information measured in each cycle, which can be load, discharge or impedance. Each type of analysis has its own characteristics. The figure 8.1 corresponds to the structure of the data for any cycle, the data of a battery represents an array of these cycles, with as many cycles as have been measured to reach the EOL.

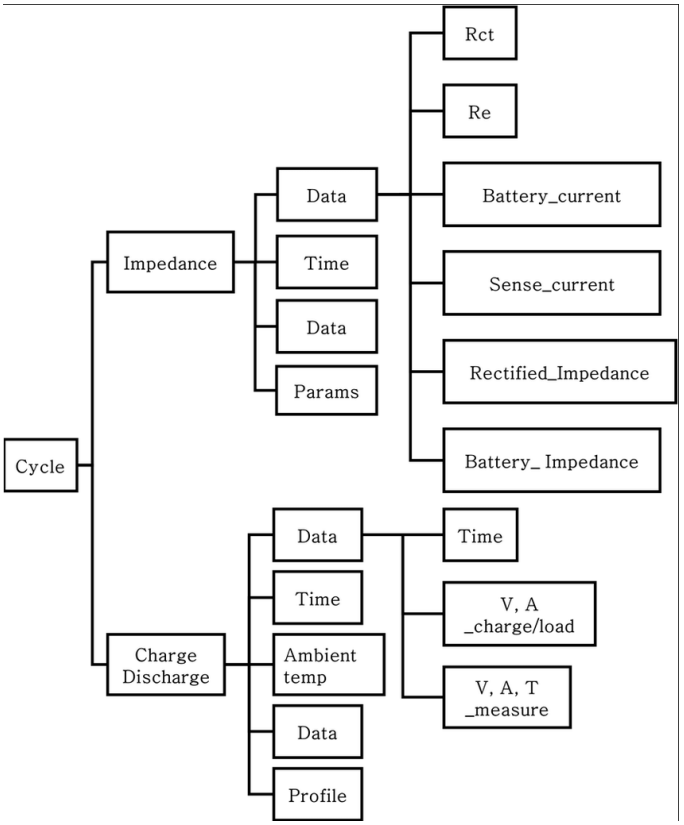

Figure 8.1: NASA Li-ion Battery Aging Datasets structure [32]

For the development of this study it has been decided to use the discharge cycles. The main reason behind this decision is the fact that the recorded cycles of this type contain the concept of "capacity" this attribute shows the capacity value of the LiB for that particular cycle. As we have already argued above, we base the calculation of important attributes for this study such as SOH on its capacity and not on other characteristics such as electrochemical impedance. That is why we have discarded other data inputs such as charging (which provide the same information as discharging but without the capacity attribute) and impedance (which do not provide a differential value to the main objective of the research).

8.2 Data preprocessing and feature exploration

Once the data set to be used has been defined, it is time to preprocess the data and visualize the relationships between them.

First of all, before studying any type of prediction or training, we must perform a series of processes on the data in order to be able to treat them and train the models with them. We observe how there are variables in the dataset of a temporal nature, attributes such as *'time'* and *'datetime'* which provide us with information about when the sample was taken. This information is shown in order to provide transparency and to complement the sampling of the batteries, but it does not add value to our research. In addition to not adding value to the research, since the date and time of the sample measurement do not have a direct relationship with the internal resistance or any other element of the battery and do not

influence its behavior, they also hinder the research since they are indicated with a data format that is incompatible with the prediction models used. For example, in the case of the SVM, datetime variables cannot be used directly in a model of this type because this algorithm is based on the creation of hyperplanes and separation of data in a feature space and datetime variables are temporal values and do not have a clear numerical interpretation that can be used by the model.

On the other hand we find another similar attribute such as *'ambient_temperature'*. This attribute does not provide any type of value to the data since it remains constant in all the captured samples (24ºC). The value provided by this variable is to guarantee that all the samples have been taken under the same conditions and that the outside temperature could not have affected the degradation or conservation of the battery; we are simply studying the behavior of the battery after a certain number of cycles of use.
Therefore, it has been decided to discard these three attributes from the study, to simplify the study and the creation and training of the models, since their absence does not affect the prediction at all, as it is not possible to find linear relationships between these attributes and the target variable, due to their characteristics, they are discarded and the study is simplified.

Secondly, entering the treatment of relevant attributes, we find the variable *'rectified_impedance'*. In this variable we find how, for each cycle of the impedance type, where the different values of the measured variables of this type are recorded, the number of samples for each cycle of the *'rectified_impedance'* attribute is about 15% less than the rest of the attributes, as shown in the following figure:

1×1 struct with 7 fields

Field	Value	Size	Class
Sense_current	*1×48 complex double*	1×48	double (complex)
Battery_current	*1×48 complex double*	1×48	double (complex)
Current_ratio	*1×48 complex double*	1×48	double (complex)
Battery_impedance	*48×1 complex double*	48×1	double (complex)
Rectified_Impedance	*39×1 complex double*	39×1	double (complex)
Re	0.0467	1×1	double
Rct	0.0763	1×1	double

Figure 8.2: Example of an impedance cycle measurement with number of samples per cycle (own elaboration)

Understanding the value of this attribute, defined as the impedance of the calibrated and smoothed battery [32], the fact of finding smaller samples in each cycle measurement may be due to an inability to measure at some moments of the cycle or to the definition of the variable itself, where some values have been smoothed. The drawback arises at the moment of making a prediction with these data, where we find how the difference of samples within each cycle represents a problem for the formalization of the model. For this reason, and given the high number of samples of this attribute with respect to the total (85% in all cases), it has been decided to replace the missing values by the weighted average of the remaining values.

Following the thread, there have been other variables related to impedance that have had to be adapted to carry out the study. As we see in Figure 8.2 above, the *'Class'* column indicates that there are complex type variables. This data format is not compatible with the models used in this study. It has therefore been necessary to convert these values to real values in order to be able to use them.

1×48 complex double

	7	8	9	10	
1	3.3048e+02 - 6.1829e+01i	8.2960e+02 - 6.2428e+01i	8.2713e+02 - 6.5093e+01i	8.2734e+02 - 6.7767e+01i	8.2673
2					
3					

Figure 8.3: Example of imaginary values of the variable *'sense_current'* (own elaboration)

Then, based on previous studies and especially on the *'SOH and RUL'* section of this project, the *'SOH'* variable is added to the dataset and calculated as explained in the aforementioned section.

So far we have a dataset where we have eliminated irrelevant attributes and transformed those attributes whose original data type was incompatible with the creation of the models for the predictions of the study. However, the dataset is very large, this is because for each cycle we store each of the measurements of each attribute for the entire cycle. As can be seen in Figure 8.2 and as explained above, for each cycle we have up to 48 measurements (in that case) of the variables. These values represent the measurement of that attribute throughout the cycle, as we can see in Figure 8.3, it is the evolution of the variable throughout that particular cycle. We observe in the following figure how some attributes have many values in the same cycle as is the case of *'voltage_measured'*, which has up to 940 samples per cycle.

1×940 double

	1	2	3	4	5	6	7	8
1	3.3251	3.0020	3.4346	3.4549	3.4688	3.4810	3.4919	3.5019
2								
3								

Figure 8.4: Example of values per cycle of the variable *'voltage_measured'* (own elaboration)

This fact poses a problem for the study, if the dataset is maintained we find that there can be up to almost 1000 samples per cycle which can make the dataset too large. In addition to slowing down the prediction time, the fact of not having the same number of samples per cycle (not even between the same type of process) and being so disparate between processes (we see how there are almost 20 times more samples per cycle between impedance process Figure 8.3 and load processes 8.4) can cause a bias in the creation of the prediction model. In addition to all this, the different samples taken in the same cycle of any value do not differ greatly from each other, they show the evolution of that attribute, this trend is also observable with respect to different subsequent and previous cycles of the same type.

For these reasons it has been decided to group all the attributes of the dataset by cycles, determining the value of that variable in a particular cycle as the arithmetic mean of all the samples obtained for that variable in that cycle. In this way, the trend of the data and the information it provides is preserved while reducing the volume of total data handled and maintaining the information for the study and prediction. Below is a comparison of the dataset before and after clustering:

Sample	Cycle	Capacity	Voltage Measured	Current Measured	Temperature Measured	Current Load	Voltage Load	SOH
0	2	1.856487	4.191492	-0.004902	24.330034	-0.0006	0.000	1.000000
1	2	1.856487	4.190749	-0.001478	24.325993	-0.0006	4.206	1.000000
2	2	1.856487	3.974871	-2.012528	24.389085	-1.9982	3.062	1.000000
3	2	1.856487	3.951717	-2.013979	24.544752	-1.9982	3.030	1.000000
4	2	1.856487	3.934352	-2.011144	24.731385	-1.9982	3.011	1.000000
…	…	…	…	…	…	…	…	…
50280	614	1.325079	3.579262	-0.001569	34.864823	0.0006	0.000	0.713756
50281	614	1.325079	3.581964	-0.003067	34.814770	0.0006	0.000	0.713756
50282	614	1.325079	3.584484	-0.003079	34.676258	0.0006	0.000	0.713756
50283	614	1.325079	3.587336	0.001219	34.565580	0.0006	0.000	0.713756
50284	614	1.325079	3.589937	-0.000583	34.405920	0.0006	0.000	0.713756

Table 8.1: Example of dataset of discharge cycles before cycle grouping (own elaboration)

Sample	Cycle	Capacity	Voltage Measured	Current Measured	Temperature Measured	Current Load	Voltage Load	SOH
0	2	1.856487	3.529829	-1.818702	32.572328	-1.805570	2.404944	1.000000
1	4	1.846327	3.537320	-1.817560	32.725235	-1.804583	2.399260	0.994527
2	6	1.835349	3.543737	-1.816487	32.642862	-1.803575	2.397969	0.988614
3	8	1.835263	3.543666	-1.825589	32.514876	-1.812863	2.408289	0.988567
4	10	1.834646	3.542343	-1.826114	32.382349	-1.812876	2.408505	0.988235
…	…	…	…	…	…	…	…	…
50280	600	1.293464	3.466462	-1.674488	33.275688	1.661799	2.073168	0.696726
50281	604	1.288003	3.468509	-1.667447	33.320678	1.655086	2.064189	0.693785
50282	608	1.287453	3.466806	-1.667470	33.373150	1.655103	2.062717	0.693488
50283	612	1.309015	3.471071	-1.688898	33.713519	1.676430	2.107460	0.705103
50284	614	1.325079	3.475472	-1.697928	33.865318	1.685264	2.120230	0.713756

Table 8.2: Example of dataset of discharge cycles after cycle grouping (own elaboration)

These tables have been extracted from the results obtained in the different datasets before and after this process. The table has been adapted from the results obtained, see figure 1 and 2 in the annexes where the image of the results is shown. We can visualize the following tables corresponding to the comparison for the load type cycles corresponding to figures 4 and 5 of the annexes.

Sample	Cycle	Voltage Measured	Current Measured	Temperature Measured	Current Load	Voltage Load
0	1	3.873017	-0.001201	24.655358	0.000	0.003
1	1	3.479394	-4.030268	24.666480	-4.036	1.570
2	1	4.000588	1.512731	24.675394	1.500	4.726
3	1	4.012395	1.509063	24.693865	1.500	4.742
4	1	4.019708	1.511318	24.705069	1.500	4.753
…	…	…	…	…	…	…
50280	616	0.236356	-0.003484	23.372048	0.000	0.003
50281	616	0.003365	-0.001496	23.369434	0.000	0.003
50282	616	4.985137	0.000506	23.386535	0.000	5.002
50283	616	4.984720	0.000442	23.386983	-0.002	5.002
50284	616	4.213440	-0.000734	23.385061	-0.002	4.229

Table 8.3: Example of dataset of charge cycles before cycle grouping (own elaboration)

Sample	Cycle	Voltage Measured	Current Measured	Temperature Measured	Current Load	Voltage Load
0	1	4.187420	0.643455	25.324079	0.638452	4.359487
1	3	4.058826	0.949043	26.635623	0.941762	4.430904
2	5	4.058139	0.950529	26.778176	0.943114	4.402619
3	7	4.058905	0.952312	26.703204	0.944735	4.418979
4	9	4.058330	0.947728	26.617004	0.940361	4.364055
…	…	…	…	…	…	…
50280	602	4.180892	0.476511	25.506487	0.472509	4.333942
50281	606	4.181592	0.463218	25.517453	0.459319	4.252485
50282	610	4.180125	0.493932	25.664855	0.489729	4.423386
50283	613	4.180702	0.489857	25.433647	0.486064	4.431494
50284	616	2.884604	-0.000953	23.380012	-0.000800	2.847800

Table 8.4: Example of dataset of charge cycles after cycle grouping (own elaboration)

Finally, we look at the tables corresponding to the impedance cycles, where the data have been extracted from Figures 5 and 6 of the annexes.

Sample	Cycle	Re	Rct	Sense current	Battery Current	Current Ratio	Battery Impedance	Rectified Impedance
0	41	0.044669	0.069456	-1.000000	-1.000000	1.000000	-0.438926	0.070069
1	41	0.044669	0.069456	820.609497	337.091461	2.320415	0.130088	0.068179
2	41	0.044669	0.069456	827.242188	330.631561	2.424193	0.058771	0.067933
3	41	0.044669	0.069456	827.193481	330.808624	2.447002	0.005814	0.066918
4	41	0.044669	0.069456	824.929504	332.682678	2.434305	0.126081	0.068071
…	…	…	…	…	…	…	…	…
50280	615	0.050036	0.074792	915.489014	230.149506	3.334835	0.245024	0.067925
50281	615	0.050036	0.074792	916.725525	212.188858	3.440393	0.264594	0.067925
50282	615	0.050036	0.074792	914.619629	176.598038	3.670656	0.288571	0.067925
50283	615	0.050036	0.074792	880.340820	136.847626	4.060164	0.317700	0.067925
50284	615	0.050036	0.074792	801.361816	97.058853	4.550338	0.352680	0.067925

Table 8.5: Example of dataset of charge cycles before cycle grouping (own elaboration)

Sample	Cycle	Re	Rct	Sense current	Battery Current	Current Ratio	Battery Impedance	Rectified Impedance
0	41	0.044669	0.069456	811.176657	315.905164	2.493579	0.171894	0.060823
1	43	0.044669	0.069456	809.823676	316.504352	2.503652	0.171267	0.061974
2	45	0.044669	0.069456	810.495946	316.617010	2.489316	0.171488	0.060362
3	47	0.044669	0.069456	809.475062	317.225532	2.491553	0.170585	0.061330
4	49	0.044669	0.069456	810.923232	315.977955	2.495459	0.172002	0.060840
…	…	…	…	…	…	…	…	…
50280	605	0.050036	0.074792	842.370181	308.431960	2.694519	0.220480	0.071815
50281	607	0.050036	0.074792	845.925238	305.007640	2.731819	0.224576	0.074615
50282	609	0.050036	0.074792	842.150261	308.301196	2.692602	0.220251	0.071730
50283	611	0.050036	0.074792	845.754082	304.957511	2.734526	0.224682	0.074486
50284	615	0.050036	0.074792	833.888655	316.960652	2.591606	0.208417	0.064925

Table 8.6: Example of dataset of charge cycles after cycle grouping (own elaboration)

As can be seen, the total samples have been greatly reduced by performing this grouping by cycles and a complete and compact dataset has been created.
Finally, we have the final dataset ready to work with it and to carry out the appropriate studies and predictions. However, before going into the creation of the models, it is important to visualize the data in order to detect patterns, similarities, disparities and differences between the different attributes.

The visualization will be performed by focusing on each possible type of cycle, observing the distribution of the data in relation to the attributes of the same type and how they evolve through aging. First we will study the unloading process and visualize the behavior of its attributes.

Figure 8.5 shows us the distribution of the data and their relationship with each other by means of a Scatter Matrix (SM), which provides information on how the variables are related to each other and allows us to identify possible patterns or correlations. When analyzing the SM, we can observe that some variables show a positive linear relationship, indicating that as one variable increases, the other also tends to increase. On the other hand, we can also identify variables that show a negative linear relationship, where the increase of one variable is associated with the decrease of the other. In addition, in the SM we can visualize the distribution of the data for each individual variable. We can observe if the data are scattered or if there are clusters or concentrations in certain ranges. This information is useful to understand the variability of the data and to detect possible outliers.

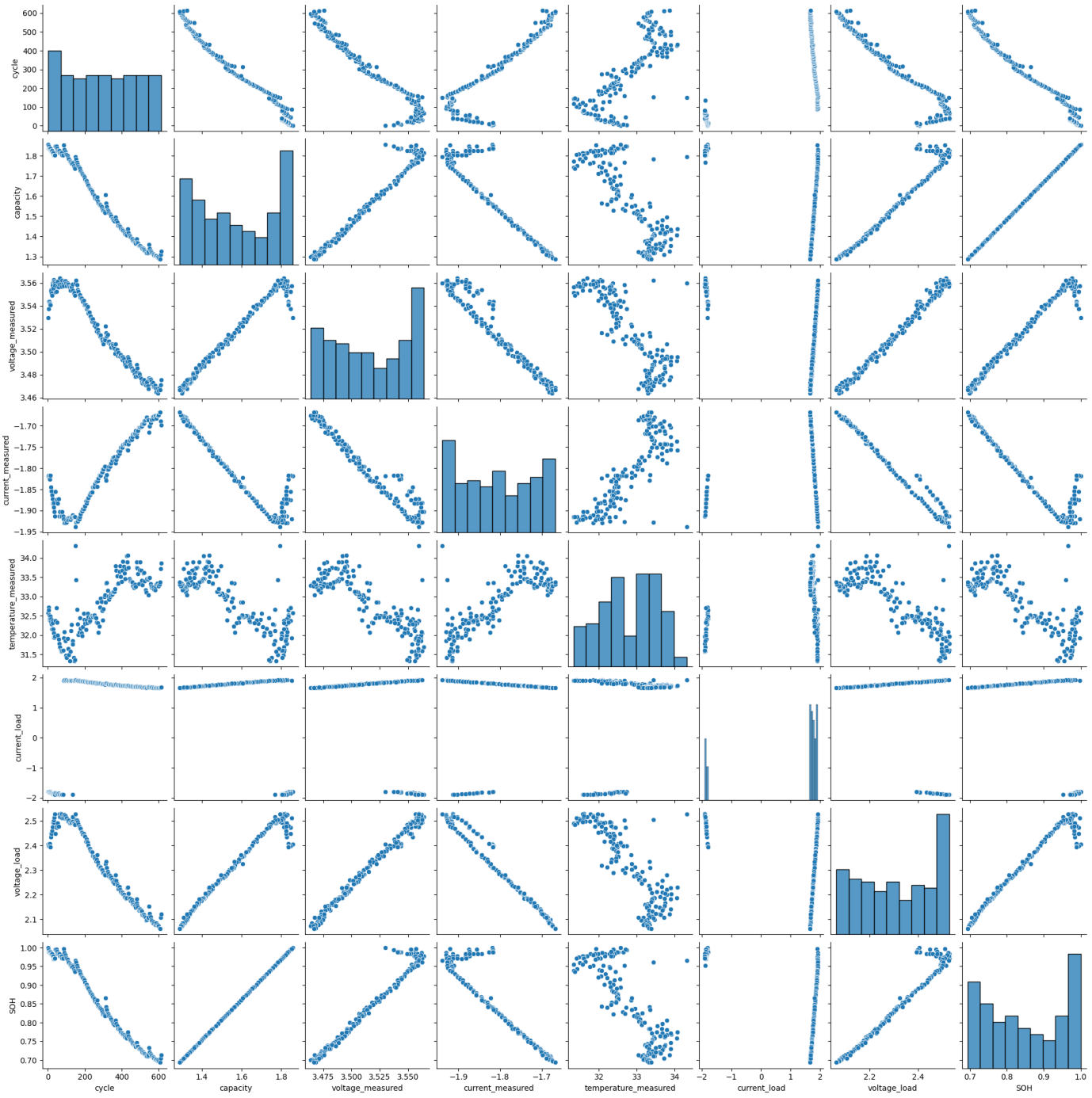

Figure 8.5: SM after grouping on discharge cycles (own elaboration)

We observe how most of the relationships between data respond to a linear behavior or, at the very least, to an ordered grouping of data. When a linear relationship between data is observed in a SM, it means that there is a clear trend in the way two variables are related to each other. The presence of a linear relationship in the SM is important because it allows us to infer and predict the behavior of one variable based on the other variable. In addition, it provides us with information about the dependence or independence between variables and is useful in exploratory data analysis and in the construction of predictive models.

It is interesting to observe in this case, the SM that we obtain without performing the data grouping process, as we can see in Figure 8.6 to compare the visualization and distribution of the data.

We observe how it results in an almost indecipherable matrix, where only some attributes, such as SOH, where we only had one value per cycle, are decipherable or similar to the figure observed previously. Similar figures are drawn but with this distribution and these data it is very difficult to perform a study since no linearity or relationship is shown between the different attributes of the discharge cycles. However, we can also observe how there are trends, tendencies or drawings that develop along the figures that resemble the same figures of the grouped study, therefore it shows us how the grouping allows us to observe the data and perform the study correctly.

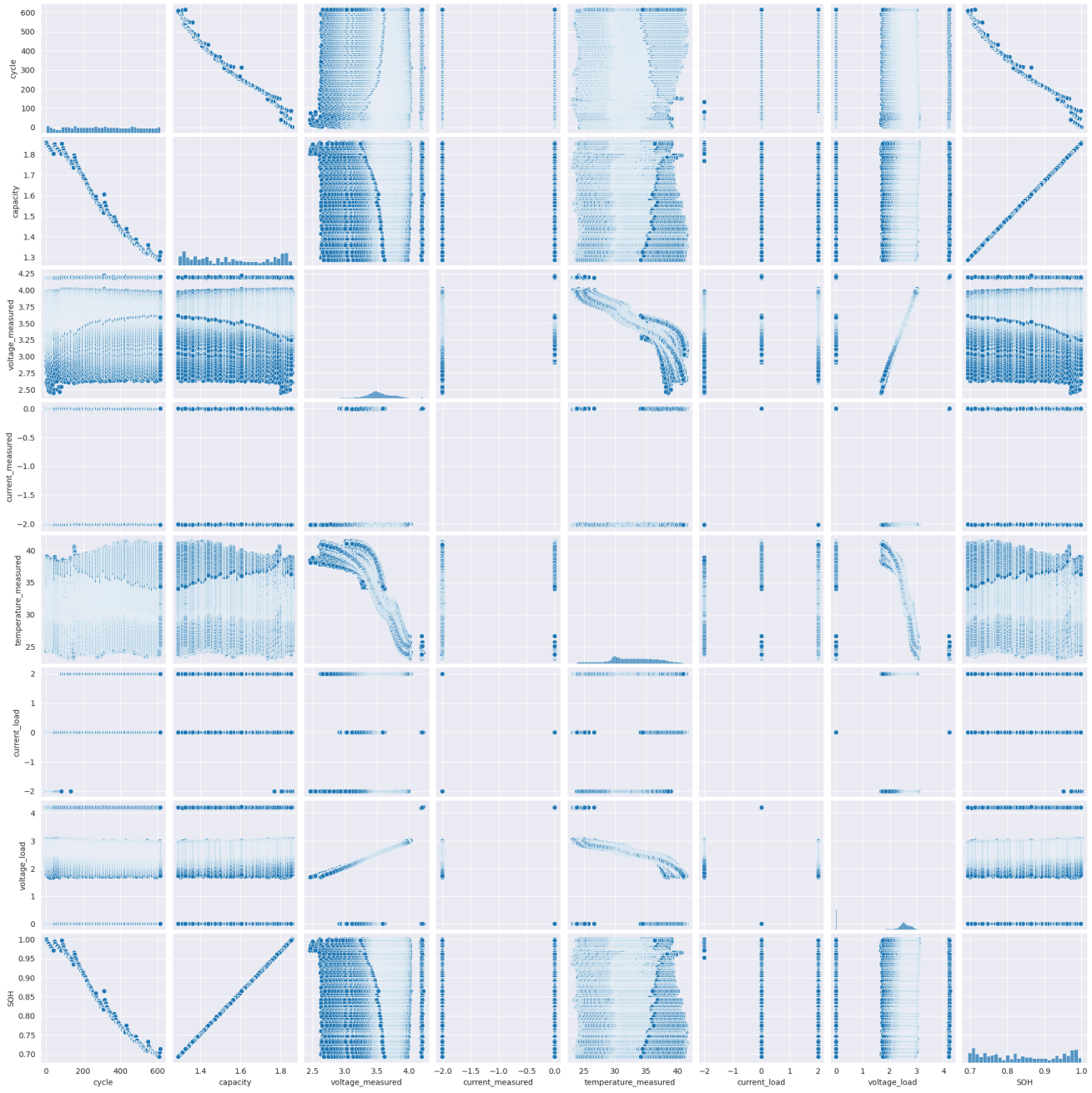

Figure 8.6: SM before grouping on dicharge cycles (own elaboration)

Secondly, we also observe the data through a heatmap diagram, which is a graphical representation that uses colors to show the relationship between two variables in a data matrix. It is a useful tool to identify patterns, trends and correlations in the data. To interpret a heatmap correctly, it is important to understand some key aspects such as colors, which are used to represent the values of the cells in the data matrix. A color scale is typically used, where darker or more intense shades represent high values and lighter shades represent low values. Darker shades represent an inverse linear relationship and lighter shades represent a direct linear relationship and the higher the color intensity the greater the degree of linear relationship between the two variables.
When observing a heatmap, attention should be paid to the patterns and trends that form. Blocks of similar or recurring colors may indicate strong relationships or consistent patterns between variables.

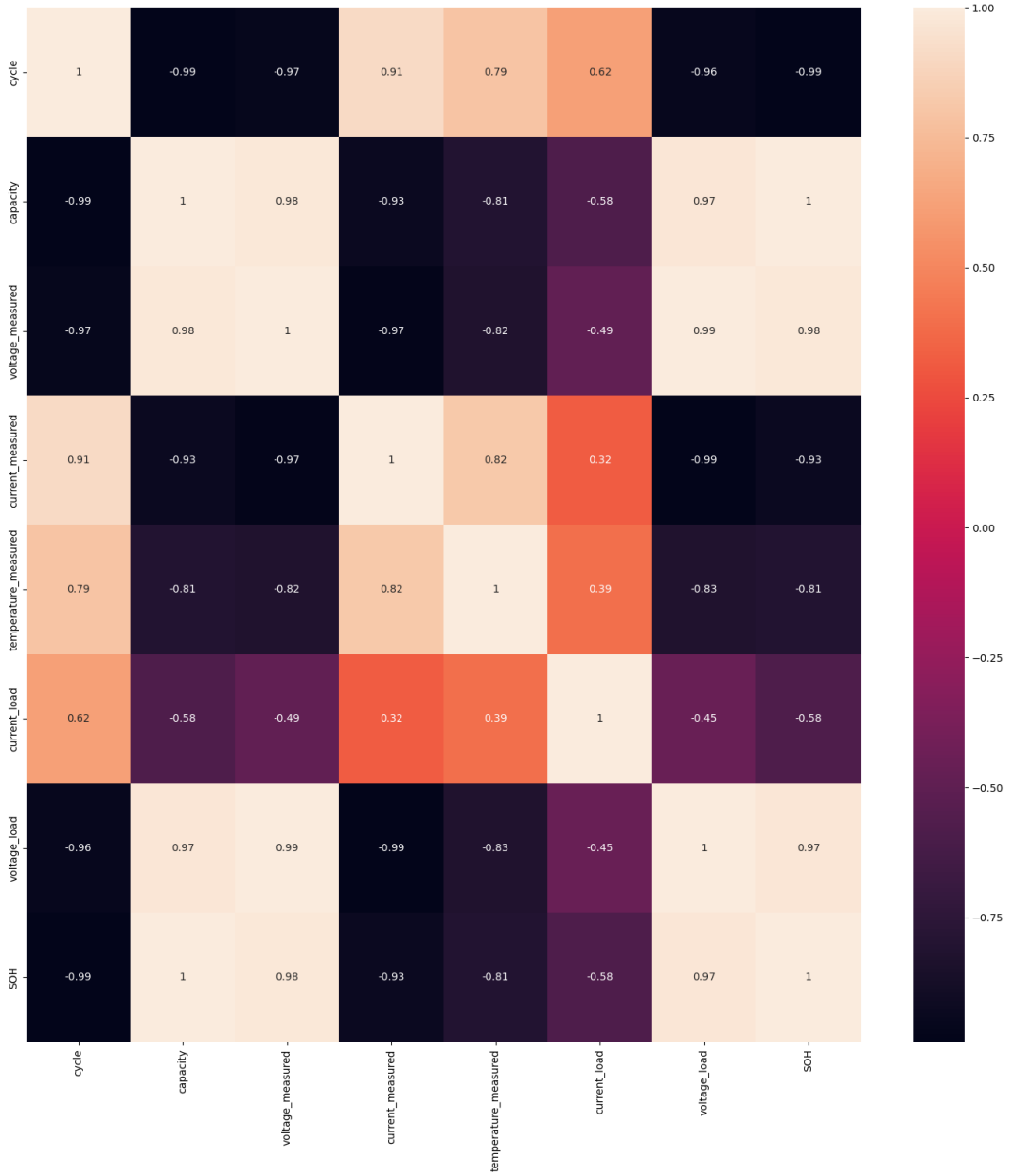

Figure 8.7: Heatmap of discharge cycles (own elaboration)

Figure 8.7 shows the heat map corresponding to the discharge cycles, where we can see some patterns and attributes that are closely related to the other variables. We take special mention to the SOH attribute, which we observe that it has a very significant relationship with almost all the other variables. Visualizing plots like these allows us to make an approximation of which attributes may be more interesting to consider as a target variable for the study. In this visualization supports the initial idea of this attribute and its importance, in addition to its characteristics and implications we observe how it is also closely related to the rest of the variables.

In the following figures 8.8 and 8.9 we observe the evolution of the variables '*voltage_measured*' and '*current_measured*' to observe the trend of these attributes show how the measurements have behaved throughout the aging. It is interesting to observe how as the cycles increase the voltage decreases and the current increases, this fact tells us how the battery behaves as it is used and the clear linear trend is tangible with respect to the SOH graph in Figure 8.10.

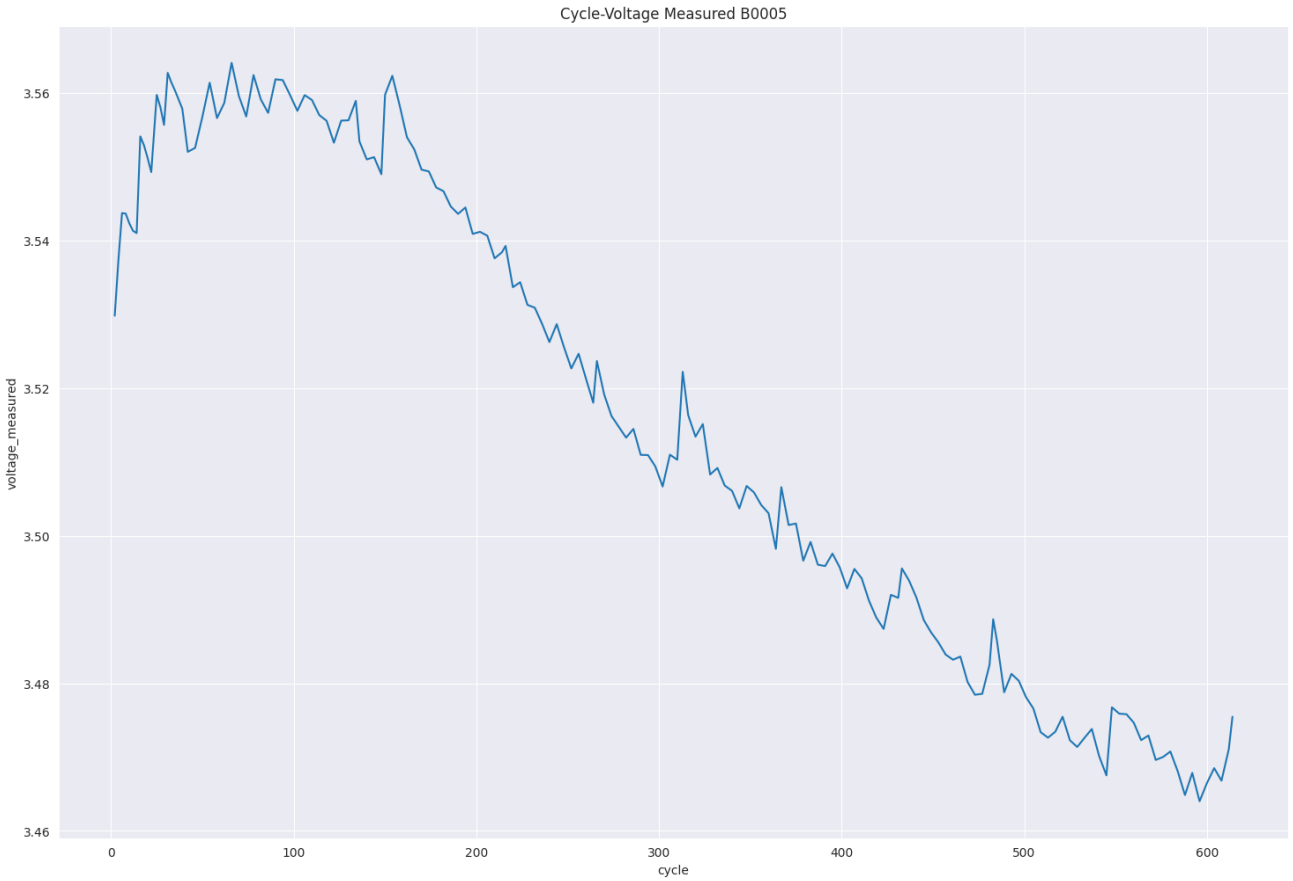

Figure 8.8: Cycle-Voltage Measured graph (own elaboration)

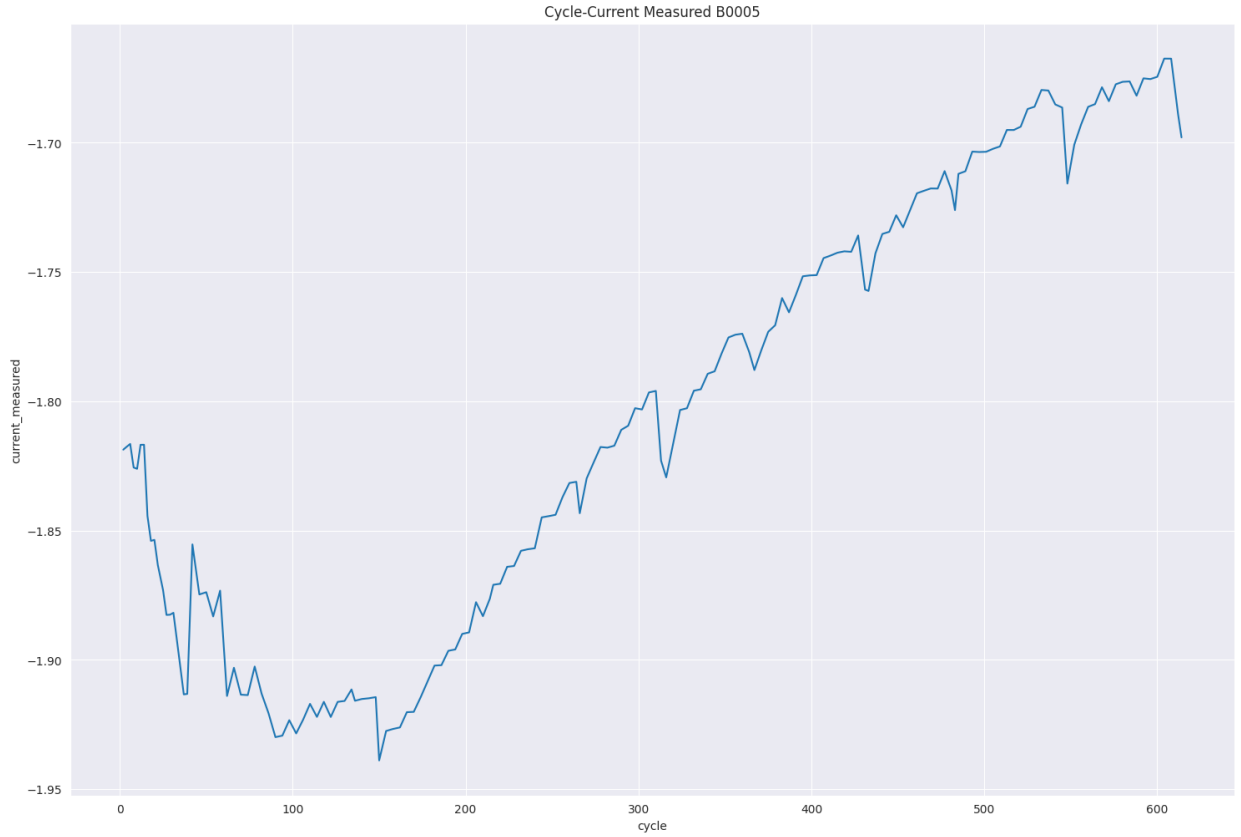

Figure 8.9: Cycle-Current Measured graph (own elaboration)

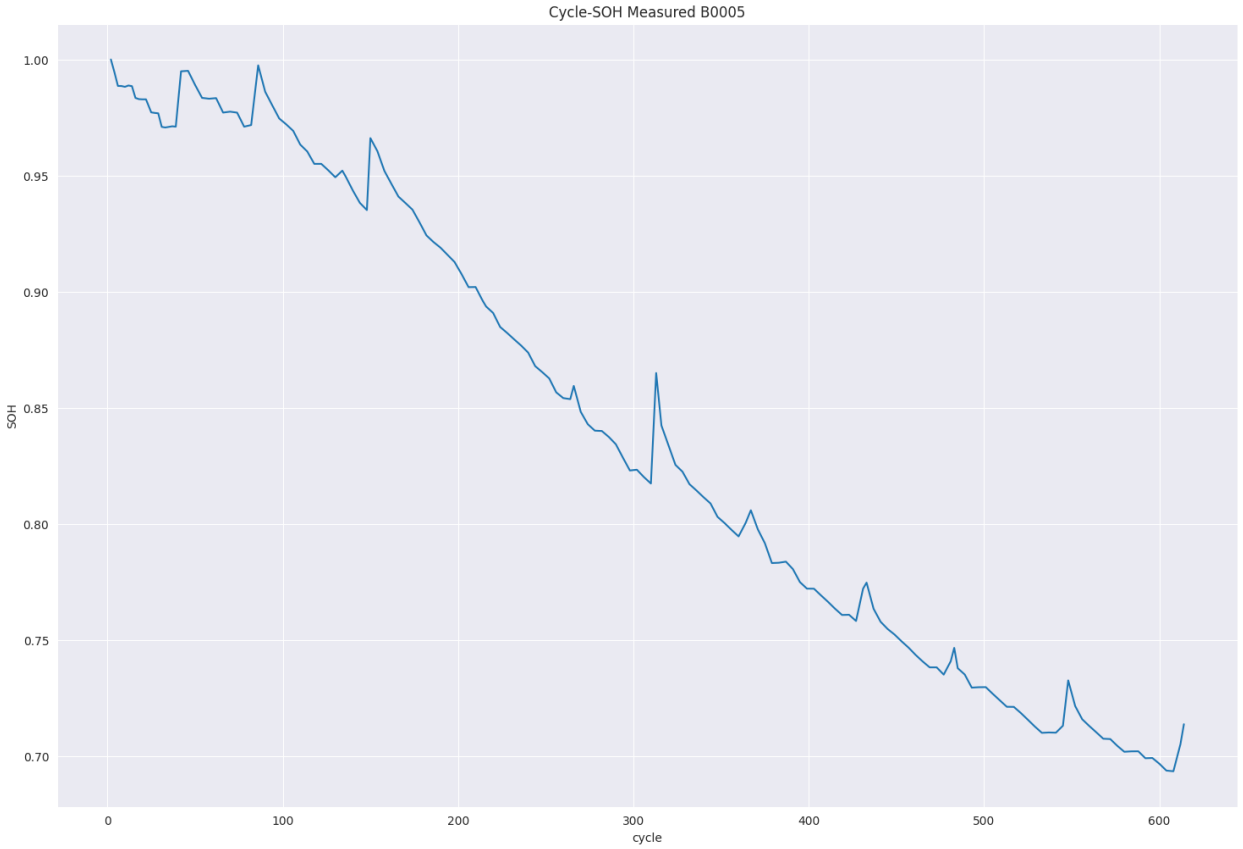

Figure 8.10: Cycle-SOH Measured graph (own elaboration)

Secondly, we looked at the visualization of the data with respect to those loading cycles. Analogously to the unloading cycles, we can see in Figure 8.11 the relationships between the different variables. In this case we can see how the relationships between most of the variables are not as linear as in the unloading cycles, but there are some attributes that show more linear similarity between them than others. It seems clear that this type of cycle is not as optimal a candidate for the study as were the discharge cycles because of this lack of linearity between variables. Also, compare it with Figure 8.12 where clear differences are observed, as with the previous type of cycle, where without performing the grouping the samples are indecipherable and dispersed.

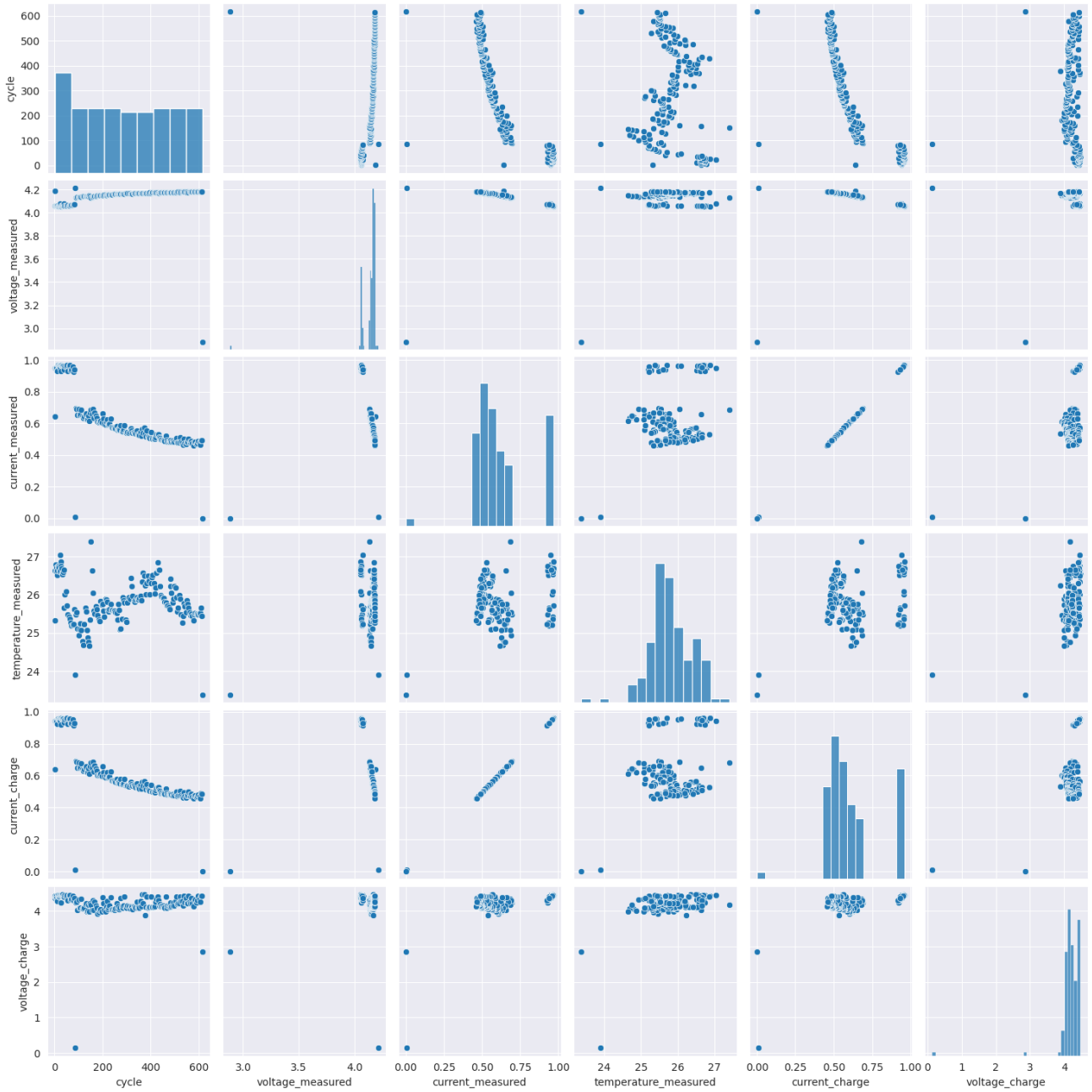

Figure 8.11: SM after grouping on charge cycles (own elaboration)

Figure 8.12: SM before grouping on charge cycles (own elaboration)

Figure 8.13 confirms what has been previously commented, with this heatmap we observe how the relationships between variables, the linearity and correlation between the different attributes is not at all favorable. Except for a couple of specific cases where the correlation is implicit for chemical reasons, all the other attributes show no relationship or linearity between them at all.

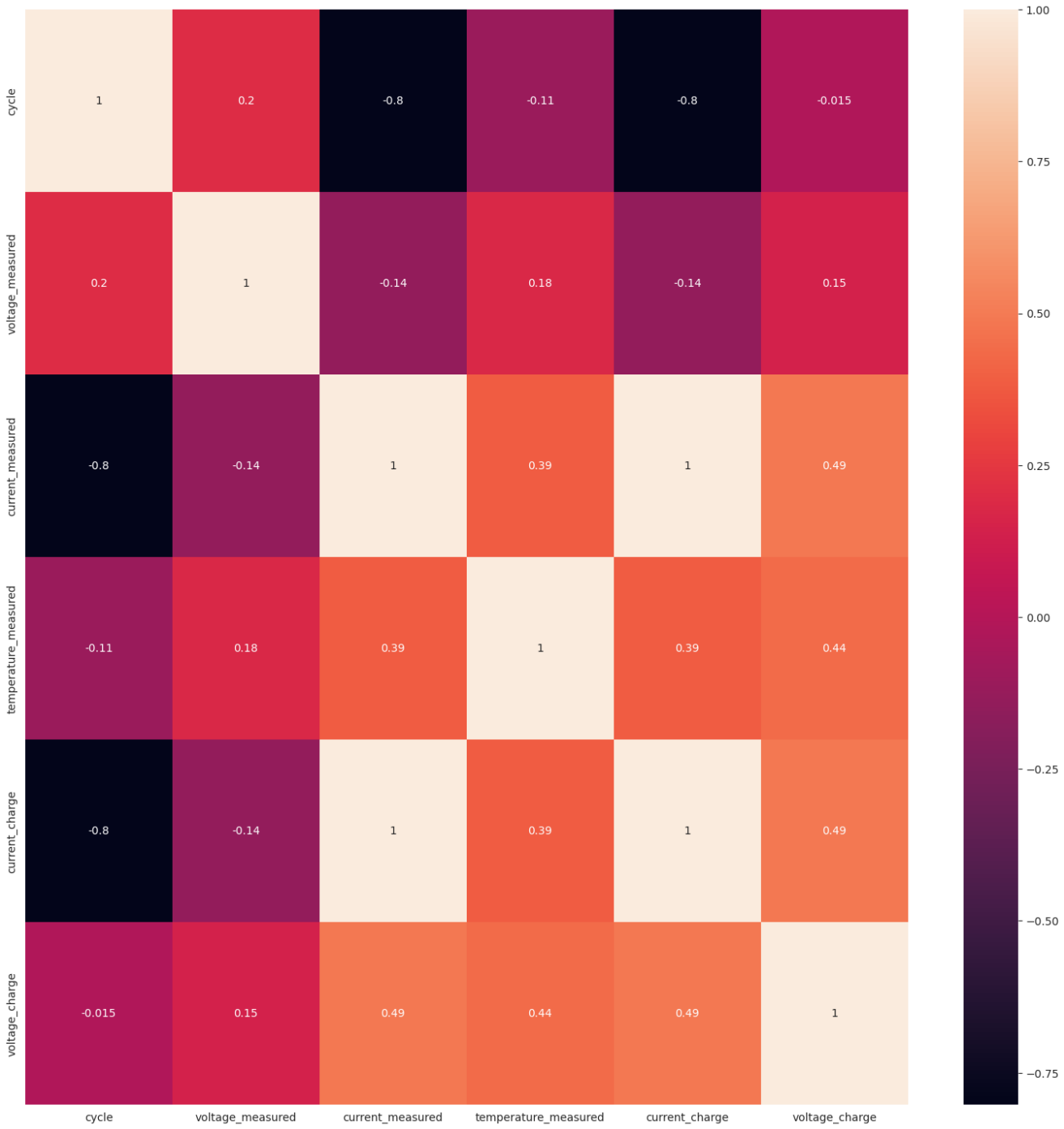

Figure 8.13: Heatmap of charge cycles (own elaboration)

Figure 8.14 shows the graph of the current charge throughout the cycles until degradation. This graph is interesting because it shows the chemical aspects of the degradation of the battery. We observe how there are two main peaks, one in the first instance and another at the end of the cycle, which concludes when the value reaches 0, corresponding to the moment when the battery stops working. This behavior marked by two clear trends explains a little of the character of the attributes of this type of cycle, marked by the chemical factors governing the battery.

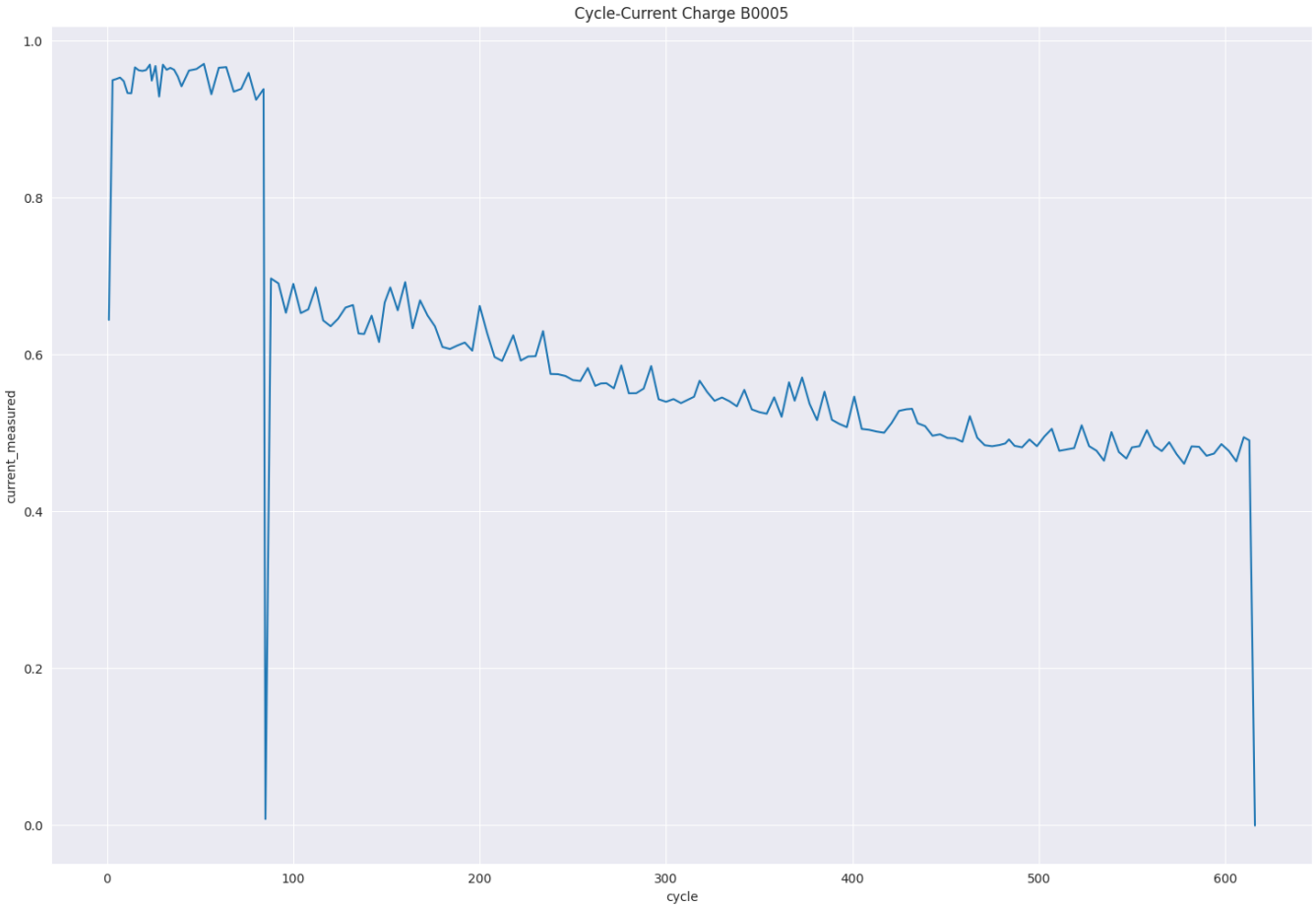

Figure 8.14: Cycle-Current Charge graph (own elaboration)

Finally, we also consider the study corresponding to the impedance type cycles. We begin by observing the SM as we have done in the previous cycles in Figure 8.15. This time it is more similar to the first SM observed, the one corresponding to the discharge cycle than to the charge cycle, in the sense that linear relationships between multiple variables are observed. However, these relationships follow a somewhat dispersed distribution in more attributes compared to the unloading cycles. Even so, we found quite a lot of linearity and relationship between the variables. To close the different analyses of the cycles we also observe Figure 8.16 the result of this same SM without the process of grouping by cycles and we reaffirm the main argument, without this initial processing on the data the study would not have been possible nor would it have been possible to conclude a result or configure models such as those obtained. The use of the data without prior processing would have resulted in a failure of prediction.

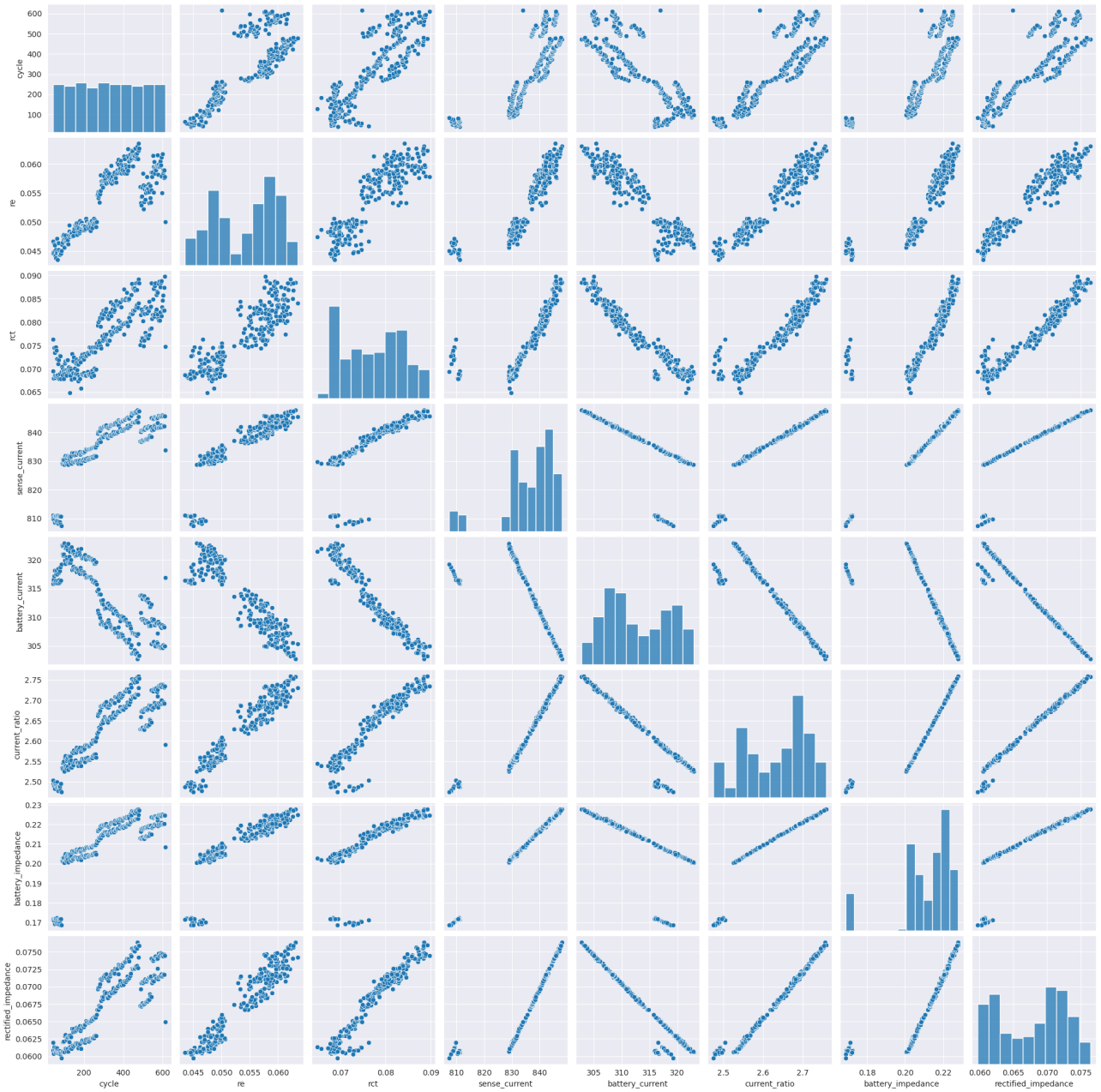

Figure 8.15: SM after grouping on impedance cycles (own elaboration)

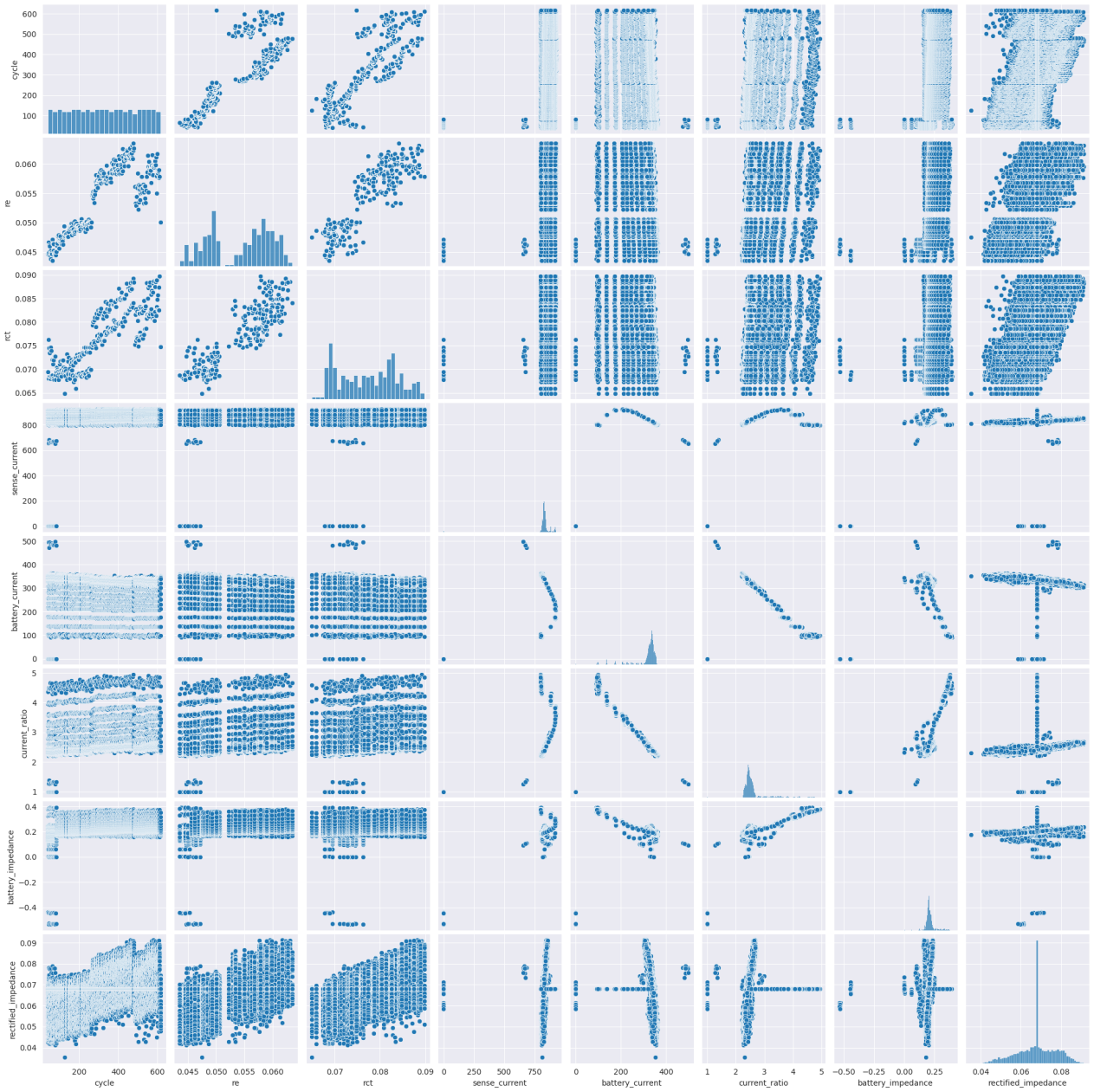

Figure 8.16: SM before grouping on charge cycles (own elaboration)

Figure 8.17 shows us the linearity again with this heatmap, we can appreciate the relationships between variables, the presence of a high linearity and general correlation between the different attributes. All the relationships have more than 70% linearity between both attributes which reaffirms the previous visualizations. Special mention should be made of the variable *'battery_curret'* which stands out for its high but negative correlation, unlike all the other variables. This fact, that all the variables follow the same direction and pattern of relationship between them except the battery current, is also due to chemical and physical issues.

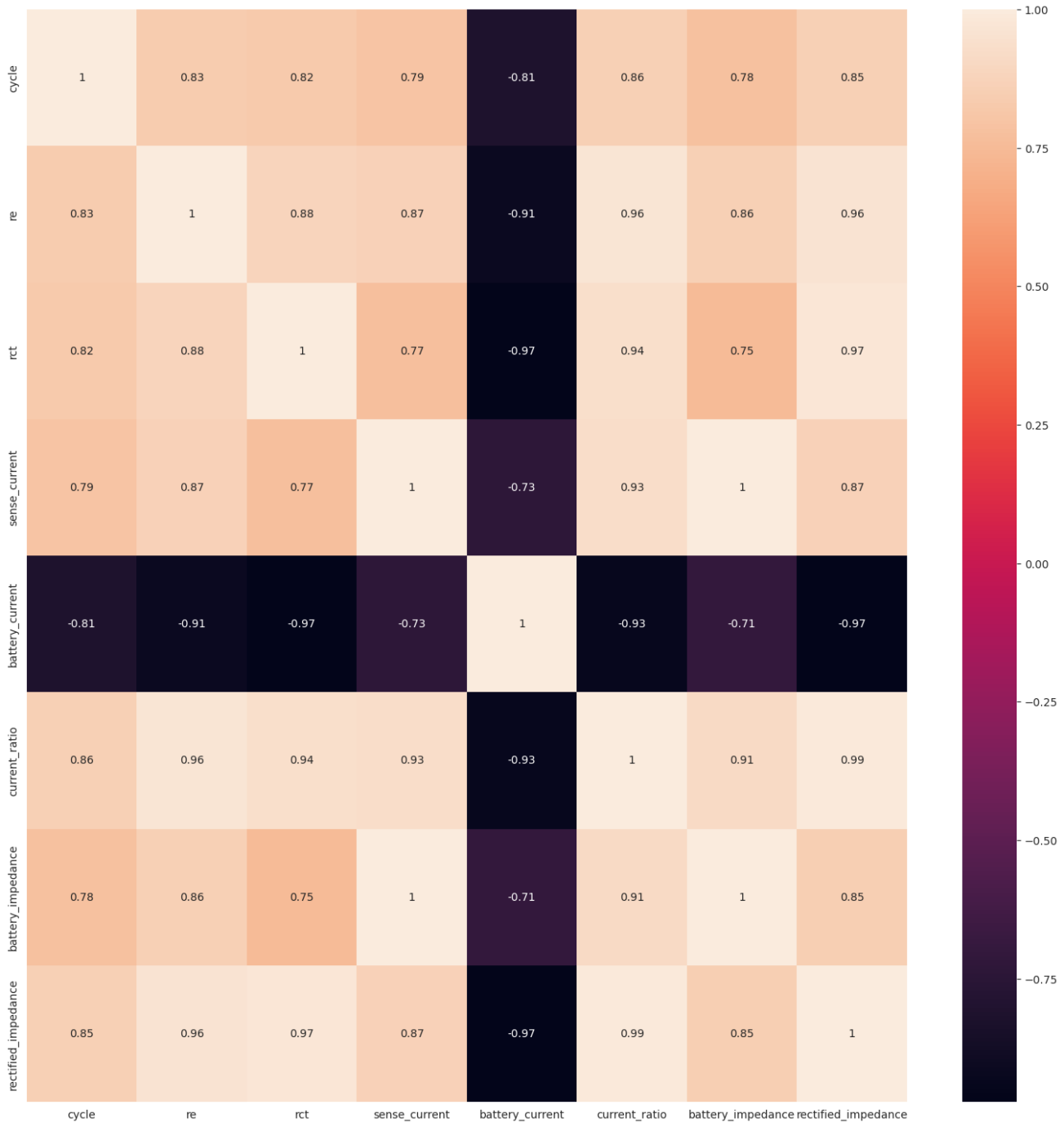

Figure 8.17: Heatmap of impedance cycles (own elaboration)

Figure 8.18 shows us the evolution of this variable that we were commenting on, the battery current, along the degradation of the LiB lifetime. Three different phases are observed until an upturn at the end of life due to several specific factors. In the first phase, the battery current remains relatively stable and constant. This may indicate an initial stage of good battery performance and stability, where the current is maintained at acceptable levels.
In the second phase, a gradual decrease in battery current is observed. This decrease may be due to internal degradation of the battery, where the energy storage and delivery capacity begins to decrease. Deposits may be forming or some internal components may be losing their ability to function efficiently.

In the third phase, there is a more pronounced decrease in battery current. This may indicate further degradation of battery capacity, resulting in limited power delivery. The battery may be experiencing increased internal resistance, which affects its ability to deliver current efficiently. In the spike observed at the end of battery life, there may be several factors at play. One of these could be the formation of degradation products which, paradoxically, can lead to a temporary increase in current. It is also possible that the battery is reaching its maximum charge capacity limit, resulting in unstable and erratic current delivery.

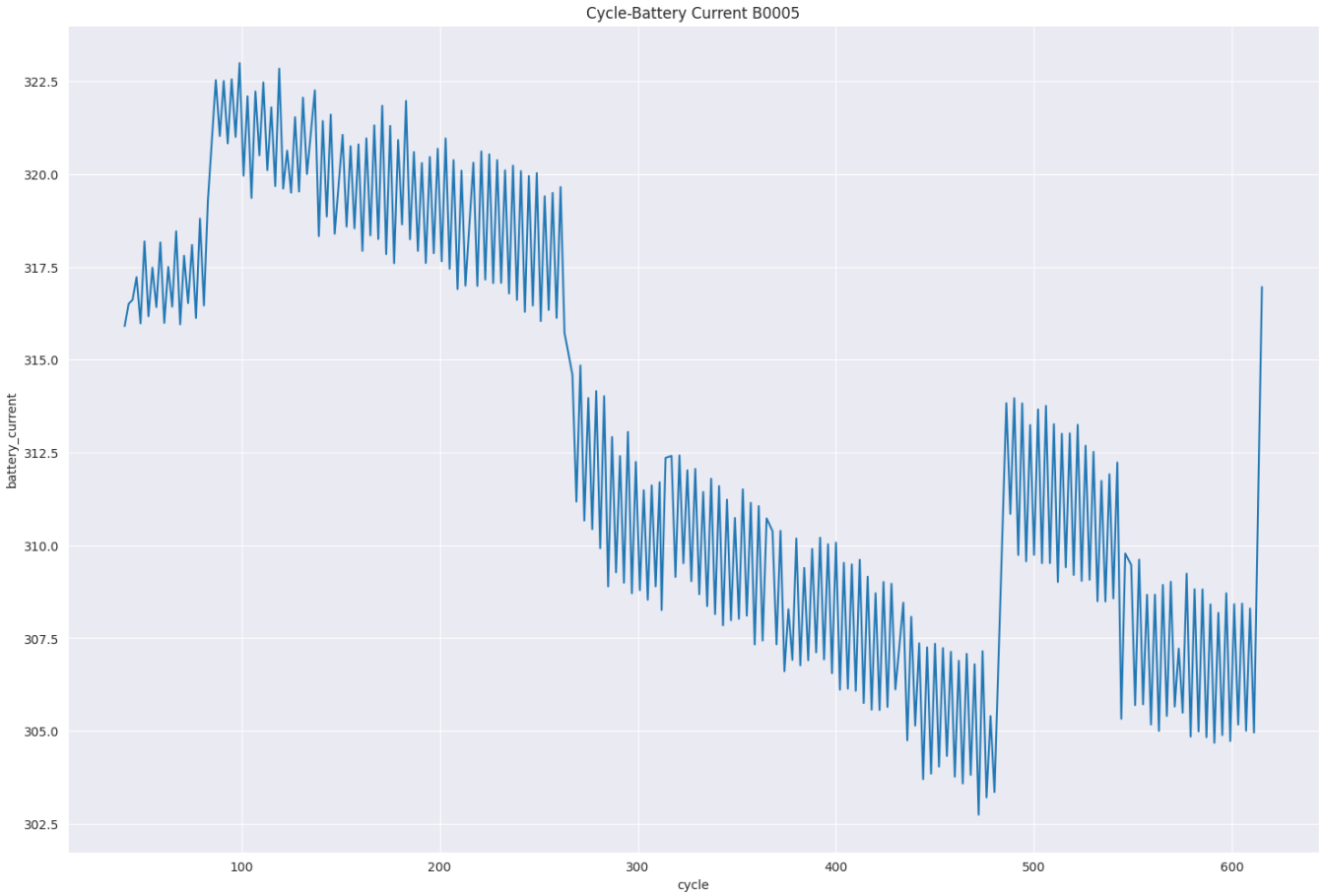

Figure 8.18: Cycle-Battery Current graph (own elaboration)

On the other hand, and to finalize the study in terms of visualization and prediction, it was decided to use the Google tool called "Embedding Projector" [33] to visualize and analyze the results obtained in the study in a complementary way to the work. This is an interactive visualization tool that allows us to explore and understand the structure and relationships between the data. Using the "Embedding Projector", the results of the study can be represented in a multidimensional space, where each point represents an instance or an observation. The position of each point in the space is calculated using dimensionality reduction techniques, in our study we have mainly used the t-SNE (t-Distributed Stochastic Stochastic Neighbor Embedding) algorithm. With this tool we can observe different patterns and groupings in a graphical way and at the same time it feeds the investigation from another point.

The t-SNE is an unsupervised learning algorithm used for visualization of high dimensional data. The main objective of this algorithm is to represent complex, high-dimensional data in a lower-dimensional space, usually 2D or 3D, so that the underlying structure of the data can be better visualized and understood [34]. A probability distribution is used to map high-dimensional points to low-dimensional points. During the mapping process, t-SNE focuses on preserving local similarity between points, i.e., points that are close in the original space tend to be close in the lower dimensional space. This allows revealing interesting patterns, clusters and structures in the data that may not be apparent in the original dimensions.
The key to the success of t-SNE lies in its ability to preserve local similarity. Points that are close to each other in the original space will be represented closely in the lower dimensional space. This means that structures and clusters present in the data will be preserved and will be more evident in the resulting visualization. This is especially useful for discovering complex patterns, identifying clusters of similar data and revealing non-linear relationships between observations.

The viewer tool allows you to interact with the data in various ways, such as zooming, rotating, filtering and selecting individual points. In addition, points can be assigned different colors or labels based on certain characteristics or variables of interest. It also provides a graphical representation of the data in a lower dimensional space, which facilitates the identification of patterns, groupings and relationships between instances. It also allows for the detection of anomalies or outliers in the data. By using this tool in the study, it is intended to obtain a deeper understanding of the data and facilitate the interpretation of the results obtained in the analysis. In addition, being an interactive tool, it provides the possibility to explore the data from different perspectives and perform more detailed analysis.

The objective of this visualization is to be able to observe patterns in the data using innovative techniques that propose a different way of understanding the data. With this visualization we can get an idea of whether classification is possible on the data set. This is especially useful in exploratory data analysis tasks, where we seek to gain an initial understanding of the structure and distribution of the data before applying more advanced modeling or classification techniques.

In this case we now look at the figure 16 in the appendices, we see the projection of the data without performing any operation on them or subjected by any algorithm.

Figure 8.19 shows us the distribution in a 3D plot of the data after having run the t-SNE algorithm. For the execution of the algorithm we find different editable parameters to adjust the model, the most appropriate perplexity value depends on the density of the data. Loosely speaking, a larger / denser dataset requires a larger perplexity. Typical values for perplexity range between 5 and 50; the learning rate, the ideal learning rate often depends on the size of the data, with smaller datasets requiring smaller learning rates; supervision indicating the importance used for supervision, from 0 (disabled) to 100 (full importance).
In our study, we have managed to converge the model to a fixed structure with perplexity values equivalent to 20 and a learning rate of 1, the supervision is fixed at 0, as we have commented above, this is an unsupervised algorithm.

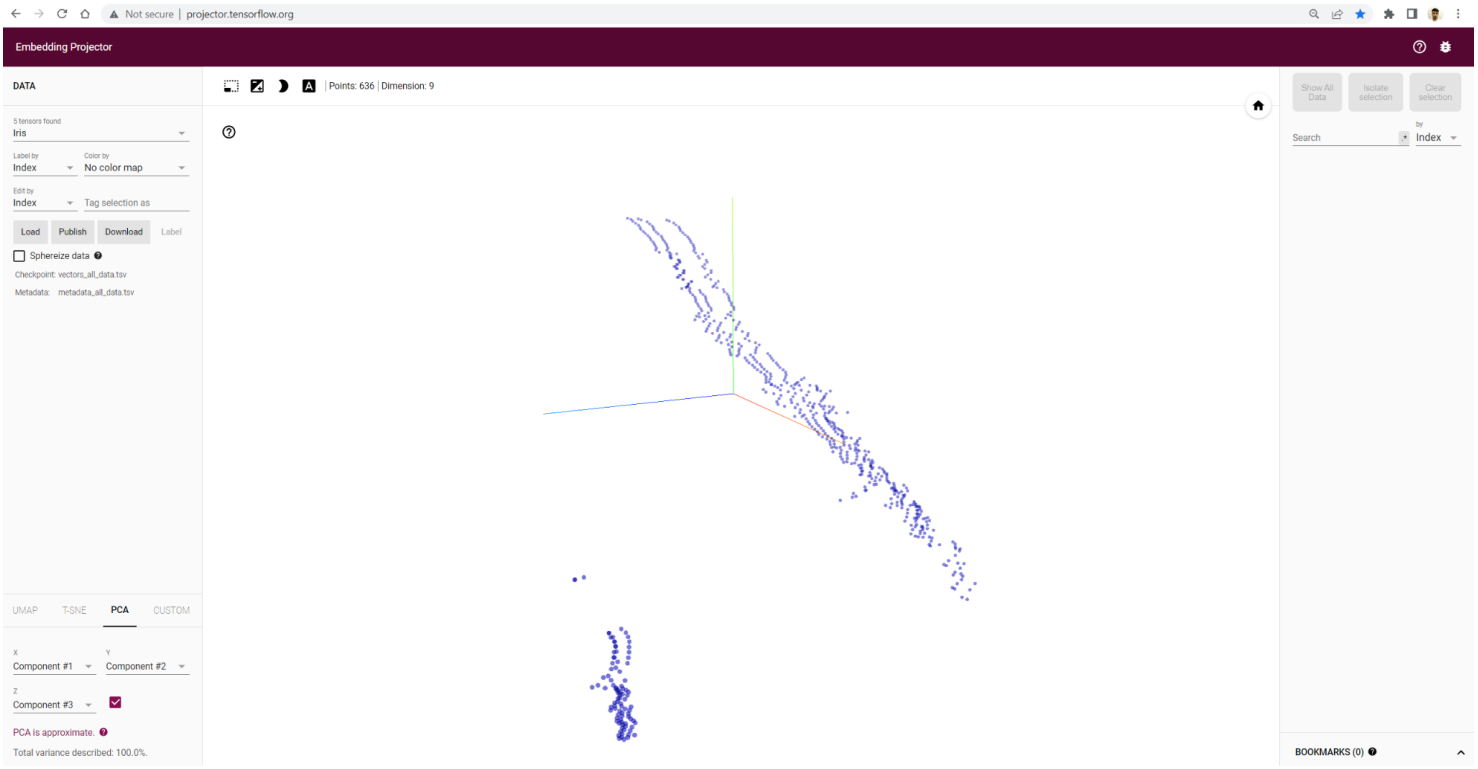

Figure 8.19: Initial distribution of data in the 3D visualization (own elaboration)

From the images we consider several aspects to be noteworthy. Firstly, we must take into account that they have been colored according to the SOH value of that cycle, i.e., the darker and more intense the sample, the higher the SOH of that cycle, and the lighter the color and less intense the lower the SOH value. On the other hand, the label that appears on each sample also shows the SOH value of that cycle.

Considering these aspects, Figures 8.20 and 8.21 show us two large groupings are drawn, the first, more extensive in space and numerous, and the second with a smaller number of samples and more concentrated in space. It is curious to observe how Figure 8.22 can be seen that the lowest SOH values are distributed in this small grouping, where the most discolored samples are observed, including one of the samples with the lowest SOH of the whole study, highlighted in this figure. On the other hand, in the same figure it can be seen how in the large grouping, there is a small subdivision of samples with a very intense color, these correspond to those with a higher SOH value.

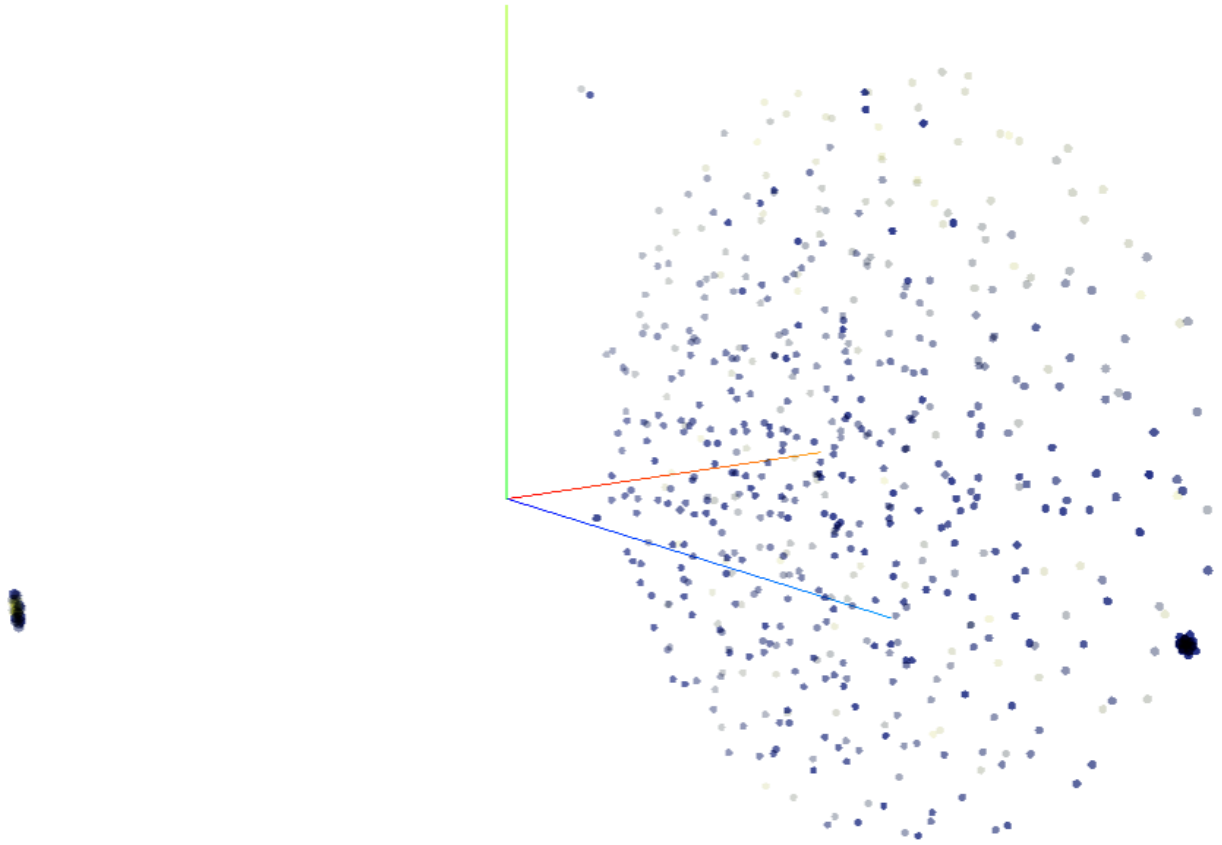

Figure 8.20: Distribution of the data in the three-dimensional space when applying the t-SNE algorithm (own elaboration)

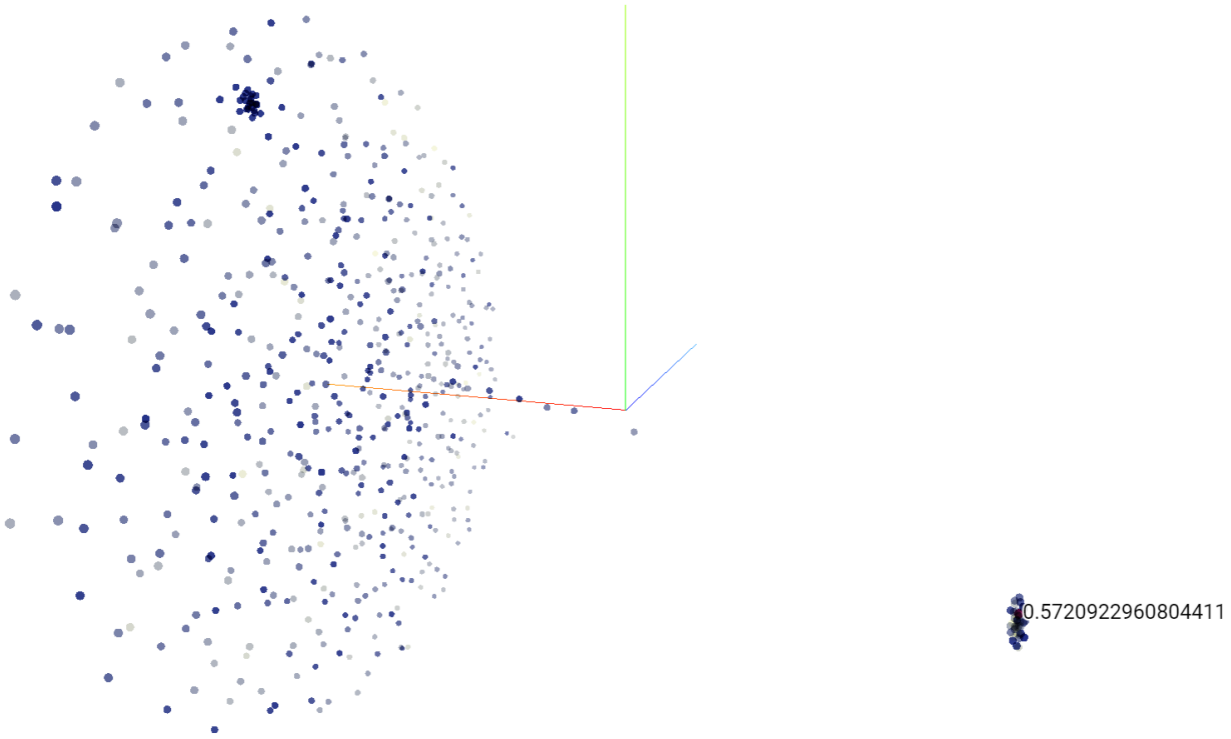

Figure 8.21: Distribution of the data in the three-dimensional space when applying the t-SNE algorithm II (own elaboration)

Figures 8.22 and 8.23 show us these small clusters have been enlarged a little more, it can be seen in Figure 18 the subgrouping of the highest SOH values that we commented previously. Interestingly, the samples with higher SOH values have been distributed and grouped together, and also separated from the rest of the general cluster. On the other hand, we confirm with Figure 19 the small grouping of low SOH values, as it is the samples with this lighter color that are distributed in this small grouping separated from the rest of the more general samples.

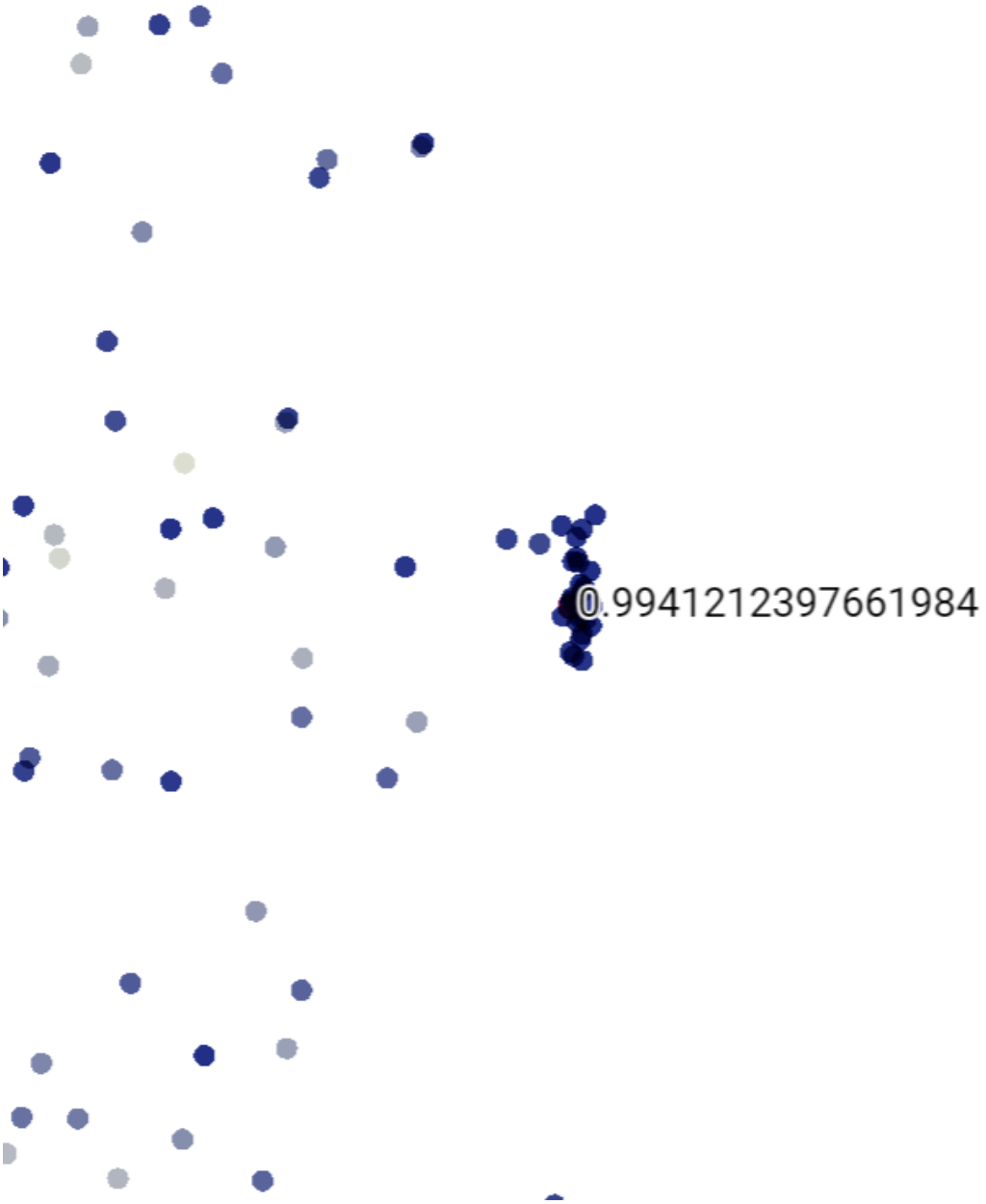

Figure 8.22: High SOH subcluster display (own elaboration)

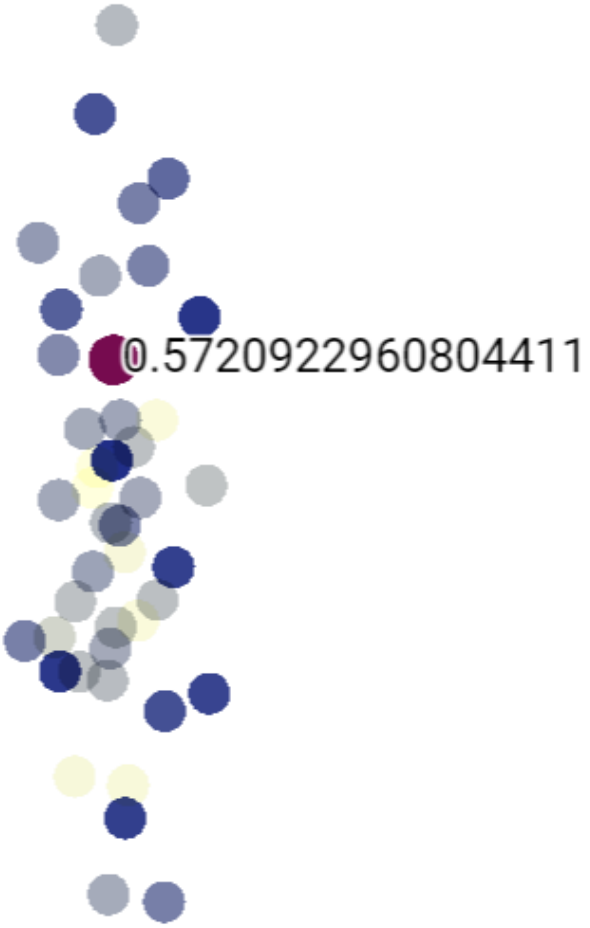

Figure 8.23: Low SOH subcluster display (own elaboration)

8.3 Training methodology and model evaluation

For the study and prediction of the model, the three ML techniques discussed above were performed: SVR, RF and NN. These algorithms have been proposed since they are adapted to the needs of the project as justified in previous sections and have given great results for the data set obtained. For each of them the model has been adapted in order to obtain the most accurate prediction possible.

The study and training has been adapted for each of the ML techniques used, their characteristics and advantages, but they all share some points in common. First, in order to provide the study with greater robustness and a larger and more diverse data set, four different batteries have been used to generate the training and test data sets. The four batteries correspond to the first set of batteries named by NASA as "BatteryAgingARC-FY08Q4": a set of four Li-ion batteries (# 5, 6, 7 and 18). The data have been pre-processed in the same way and through the processes explained in the previous section. These data sets have been chosen because they share the same chemical and physical characteristics of the batteries studied, and observing the evolution of each of them enriches the study and analysis of the degradation of the LiBs attributes.

In all the techniques the definition of the training set and the test set has been done in the same way, the batteries #7 and #18 have been used to form the training set, and the batteries #5 and #6 have been used to form the test set. In this way, two sets are established, one for training and the other for testing, totally independent of each other, where the model is provided with test data that have not been seen before. In addition to this, we provide the model with robustness in training, providing two independent data sets of batteries that represent a significant amount of data to train, while providing the test set with an equal amount of different data. In this way we obtain a balanced evaluation of the model, where we do not find overfitting or underfitting in any of the phases of the model creation giving confidence in the prediction of the model.
Secondly, the target variable is set as the RUL, calculated as defined above, for each of the different cycles of both the training and test set.

The cross-validation technique explained above has been used to train the model. This technique is used to evaluate the performance of the model and to avoid over-fitting or under-fitting during training. As previously discussed, it consists of dividing the data into several partitions (k-folds) and training the model on k-1 partitions and testing it on the remaining partition. This process is repeated k times, so that each partition is used once as a test set. Finally, the results of the k iterations are averaged to obtain a more accurate evaluation of the model. This technique allows obtaining a more accurate estimation of the model performance and avoiding again the overfitting or underfitting that may be caused in the training of the data, resulting in a more robust and generalizable model.

In addition to using cross-validation, we have also utilized the grid search technique to optimize the hyperparameters of the model during training. Grid search is a method that systematically tests different combinations of hyperparameters to find the optimal set of values that maximize the performance of the model. During grid search, we defined a range of values for each hyperparameter that we wanted to tune. The grid search algorithm then exhaustively tested all possible combinations of hyperparameter values within these ranges

to find the optimal set of values that produced the highest accuracy score on the validation set. By combining cross-validation with grid search, we were able to train our model on different subsets of the data, and at the same time, optimize the hyperparameters of the model in a way that ensures the best performance on new, unseen data.

For the study using the SVM technique for the target variable RUL, the hyperparameters chosen for their great result have been the following: {'C': 0.1, 'degree': 2, 'epsilon': 0.01, 'kernel': 'linear'}. The hyperparameter 'C' represents the regularization parameter, which controls the trade-off between data fit and model complexity. In this case, the selected value of 0.1 indicates that more importance has been given to the data fit. The hyperparameter 'degree' refers to the degree of the polynomial in the polynomial kernel. In this case, a degree of 2 has been selected, implying a second degree polynomial kernel. The hyperparameter 'epsilon' represents the acceptable margin of error in the prediction. A value of 0.01 indicates that a small margin of error has been set, implying higher prediction accuracy. Finally, the hyperparameter 'kernel' indicates the type of kernel used in the SVM model. In this case, the linear kernel has been selected, which is suitable for problems with a linear relationship between the input variables and the target variable.
In the case of the study for the SOH variable, these parameters have been similar, we note how they are {'C': 1, 'degree': 2, 'epsilon': 0.01, 'kernel': 'linear'}. We note how this configuration has proven to be effective in both cases.

For the study using the RF technique, the hyperparameters chosen for their great result have been in the following values: {'max_depth': None, 'max_features': None, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 50}. These parameters indicate the optimal settings for the Random Forest model in predicting the target variable RUL. The hyperparameter 'max_depth' represents the maximum depth of the trees in the model ensemble. In this case, the value 'None' has been selected, which means that there is no limit on the depth of the trees. The hyperparameter 'max_features' indicates the maximum number of features considered in each split of a tree. By selecting the value 'None', all available features in each split are used. The hyperparameters 'min_samples_leaf' and 'min_samples_split' determine the minimum number of samples required to form a leaf node and to perform a split, respectively. In this case, a value of 4 has been selected for 'min_samples_leaf' and 10 for 'min_samples_split'. Finally, the hyperparameter 'n_estimators' indicates the number of trees in the Random Forest model ensemble. A value of 50 has been selected, which implies the use of 50 trees in the model.

As for the SOH variable, the best selected hyperparameters are: {'max_depth': None, 'max_features': None, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 100}. We observe how this configuration, as well as the hyperparameters of the SVM model are very similar, which demonstrates the effectiveness in both cases.

Finally, in the study using the NN technique, the hyperparameters chosen for their great result have been in the following values: {'units1: 32, 'units2': 16, 'learning_rate': 0.001, 'batch_size': 16, 'epochs': 100}. These parameters indicate the optimal configuration of the Neural Network model to predict the target variable RUL. The hyperparameter 'units1' represents the number of neurons in the first hidden layer of the neural network. In this case,

32 units or neurons have been chosen. The hyperparameter 'units2' represents the number of neurons in the second hidden layer of the neural network. In this case, 16 units or neurons have been chosen. In this case, we see how we have used two layers for the neural network we have constructed. In this case, we see how we have used two layers for the neural network we have constructed. This is because we have observed that this specific configuration, with 32 neurons in the first hidden layer and 16 neurons in the second hidden layer, has given optimal results in terms of accuracy and performance and are necessary to obtain great prediction results. The hyperparameter 'learning_rate' has been set to 0.001, indicating a low learning rate. This setting has proven to be efficient in controlling the update rate of the neural network weights during training, which has contributed to optimal results. In addition, a batch size ('batch_size') of 16 has been selected, which implies that 16 training examples are used to compute back propagation and update the network weights at each iteration. This choice has been determined to be optimal for the data set and model in question. Finally, a total of 100 epochs ('epochs') have been set for training the neural network. This implies that the training algorithm will run through the entire data set 100 times. This configuration has proven to be effective in achieving optimal results in terms of model accuracy and convergence.

As for the SOH variable, the best selected hyperparameters are: {'units1: 32, 'units2': 32, 'learning_rate': 0.01, 'batch_size': 16, 'epochs': 100}. We observe how this configuration, as well as the hyperparameters of the model are quite similar to the prediction for the RUL variable, which demonstrates the effectiveness in both cases.

In our study, we have used different techniques to evaluate the results of the Machine Learning models. The metrics used were R-squared (R^2), Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

The R-squared is a metric that indicates the proportion of the variation in the target variable that is explained by the model. Its value ranges from 0 to 1, where 0 means that the model explains no variation in the data, and 1 means that the model explains all the variation. This metric is important because it allows us to know what percentage of the variability in the target variable is explained by our model. The calculation of the R-squared is based on the comparison of the variation in the target variable with the variation in the model predictions. The value is calculated as the proportion of the total variance that is explained by the model divided by the total variance in the target variable.

The MSE is the mean of the squared errors between the predictions and the actual values. This metric is important because it allows us to know the average amount of error in our predictions. A high MSE value indicates that the model has a high degree of error in its predictions. The MSE is calculated by adding the square of the differences between the predictions and the actual values, and dividing this result by the number of samples.

The MAE is the mean of the absolute errors between the predictions and the actual values. Unlike the MSE, the MAE does not penalize large errors, so it is useful in cases where large errors are not significantly worse than small errors. This metric is important because it allows us to know the average magnitude of our errors in our predictions. To calculate the MAE, we

take the absolute differences between each predicted value and its corresponding true value, and average these differences.

The RMSE is the square root of the MSE. Like the MSE, the RMSE measures the average amount of error in our predictions, but instead of being on the same scale as our data, it is on the same scale as our target variable. This metric is important because it allows us to know the average amount of error in our predictions in the same unit as our target variable. Also, by taking the square root of the MSE, the RMSE tends to penalize larger errors more, which can be useful in cases where it is important to minimize the impact of significant errors in our predictions.

8.4 Results and discussion

Once the training method, model creation and model evaluation have been defined, it is time to evaluate the results obtained. We will mainly evaluate the results of the RUL prediction but we will also make a special mention to the SOH prediction. Below we can visualize the two different tables where the results of the different models created based on each ML technique explained above are summarized. The "Initial research" row represents the results obtained from the creation of the models with the default parameters, without applying any hyperparameter or previous study. The second row represents the result obtained after applying the grid-research and cross-validation techniques to obtain the best hyperparameters and results.

Index	R^2	MAE	MSE	$RMSE$
Unit	-	%	%	%
SVM				
Initial research	0.774	10.243	160.551	12.671
Hyperparameter	0.999	0.005	3.487 e-05	0.006
RF				
Initial research	0.991	1.619	8.011	2.830
Hyperparameter	0.994	1.301	5.061	2.250
NN				
Initial research	0.995	1.753	4.196	2.048
Hyperparameter	0.997	1.117	2.256	1.502

Table 8.7: Table of model results in RUL prediction

The results obtained in the SVM prediction show that the model is highly accurate, the best performing model with an RMSE of 0.005 and an MSE of 3.487e-05. Furthermore, the R-square obtained is 0.999, indicating that the model explains practically all the variability in the data. These results are very promising as they indicate that the SVM model is highly effective in the prediction task on the data set used. It is important to note that these results have been obtained after a careful and rigorous training process, suggesting that the model is highly robust and generalizable.

The fact that the model has such a low RMSE indicates that the predictions are accurate and deviate very little from the true values. On the other hand, the MSE shows that the model errors are very small and the prediction is highly reliable. It is important to note that a model with a high MSE would not be considered accurate, but in this case, the error measure is extremely low, indicating that the model is accurate. The R-square indicates that the model is highly accurate and effective in the prediction task. These results are very positive and suggest that the SVM model is highly effective in predicting the data. Also, it is important to note that these results are based on a specific data set, and may vary if a different data set is used.

Secondly, the prediction results performed by the RF model present a significantly higher RMSE value than the prediction performed by the SVM model. This means that, on average, the predictions of the RF model have a higher error compared to the predictions of the SVM model. Although the R-square value of the RF model indicates that the model is able to explain approximately the same amount of variation in the target variable as the SVM model, which is a very high value and shows how the model is able to capture a large amount of the variability present in these data, compared to the SVM prediction it does not present such accurate results. On the other hand, the MSE value of the RF model is higher than the MSE value of the SVM model, indicating that the RF model has a larger amount of average error in its predictions.
Despite this slight decrease in model quality, overall, the results suggest that both the SVM model and the RF model are capable of making accurate predictions on the data set used, but with significant differences in their evaluation metrics.

Finally, the NN prediction obtained results quite similar to the random forest (RF) model. Comparing these results with those obtained with SVM prediction, we can see that the SVM model performed better in all attributes, although it is important to keep in mind that SVM and NN models can be more complex and time-consuming to adjust and optimize than the RF model.

A two-layer hidden neural network was used for this prediction, which implies a higher complexity and learning capacity of the model. However, despite having more parameters to fit, the neural network was able to generalize well and obtain results similar to the RF model. The layers have been 32 and 16 units respectively, which means that the neural network had a total of 1216 parameters to fit. Despite having higher complexity, the neural network was able to generalize well and obtain results similar to the RF model, suggesting that the choice of neural network architecture was suitable for the battery capacity prediction problem.

Overall, these results show that all three models are able to make accurate predictions regarding the RUL of LiBs. It has been observed how the best performing one was the model created for the prediction using SVM which was slightly superior to the rest. However, depending on the context and available resources, one or the other model could be chosen. The results have been very positive and it has been possible to prove that it is possible to generate a model capable of generalizing and predicting this objective attribute in an efficient and satisfactory manner.

The table below also shows the result of the prediction made for the SOH estimation. It follows the same format as the previous table.

Index	R^2	MAE	MSE	$RMSE$
Unit	-	%	%	%
SVM				
Initial research	-0.730	0.088	0.013	0.112
Hyperparameter	0.859	0.038	0.002	0.045
RF				
Initial research	0.790	0.040	0.003	0.054
Hyperparameter	0.799	0.039	0.003	0.054
NN				
Initial research	-5.223	0.201	0.092	0.303
Hyperparameter	0.800	0.057	0.003	0.054

Table 8.8: Table of model results in SOH prediction

Analogous to what was analyzed above, the results obtained in the SVM prediction show that the model is very accurate, being the best performing with an RMSE of 0.045, an MSE of 0.002 and an MAE of 0.038, all very low values, very low, indicating that the predictions are accurate and deviate very little from the real values and meaning that the model errors are minimal and the prediction is highly reliable. In addition, the R-squared obtained is 0.859, which indicates that the model explains the variability of the data quite well. We note how in this case the use of hyperparameters has been key to carry out a study to find which ones best fit the predictive model. We observe how the evolution in the study of hyperparameters in this case represents a clear improvement for the creation of an appropriate predictive model. The SOH prediction results are also promising as they indicate that the SVM model is very effective in the prediction task on the data set used. We should also consider that these results have been obtained after a careful and rigorous training process, suggesting that the model is highly robust and generalizable.
Analogously to the RUL prediction, despite this slight decrease in model quality, overall, the results suggest that both the SVM model and the RF model are capable of accurate

predictions on the data set used, in this case without major significant differences in their evaluation metrics.

Finally, the NN prediction obtained results quite similar to both models studied. Comparing these results with those obtained with the other models, we can observe that the results are similar to the others obtained. For this prediction a two-layer hidden neural network was also used, which implies a higher complexity and learning capacity of the model. In this case we observed how the neural network was able to generalize well and obtain results similar to the RF and SVM model. The layers have been 32 and 16 units respectively, which means that the neural network had a total of 1216 parameters to adjust. Despite having higher complexity, the neural network was able to generalize well and obtain similar results to the RF model, suggesting that the choice of neural network architecture was suitable for the battery capacity prediction problem.

In general, these results show that the three models are able to make accurate predictions about the SOH of the LiBs. We observe how, analogously to the prediction of the RUL, the best performing model was the one created for the prediction by SVM, which was slightly superior to the rest but not as superior as the one observed for the prediction of the RUL. In this case we observe how all the models studied have been adapted in a similar way to the data and have offered similar results. The results have been very positive and it has been possible to prove that it is possible to generate a model capable of generalizing and predicting this objective attribute efficiently and satisfactorily.

To conclude, it is interesting to note that these data obtained from the study support the distribution seen in the 3D projection with the "Embedding Projector" tool. We had observed that the algorithm itself had also performed some groupings and subgroupings that correspond to the main objective of the study, to observe the health status of the battery and the remaining life level for possible reuse. In the groupings we observed a group of samples with a very low SOH, we can determine that this grouping corresponds to the samples that we would classify as "not suitable for a second life", while, on the other hand, the samples belonging to the other group could be considered as "suitable for study for second life". Also, under a more extensive study, one can contemplate the possibility of classifying as more or less prone to second life the closer they are to the grouping with the very high SOH value that we saw previously.
The important thing about this tool is that we have been able to observe in a tangible and three-dimensional way the characteristics and distribution in space of the different samples and the clusters that have been made. Under these clusters an approximation of the battery health status and its subsequent classification can be made.

Conclusions

9.1 Summary

This work has focused on the analysis and prediction of the lifetime of LiBs with the aim of contributing to the development of sustainable solutions in the field of energy. Using different machine learning techniques and training different models, we have observed how we have been able to predict the lifetime of batteries with high accuracy. We have also studied the visualization of the data with the idea of using it as a classification tool for further analysis and prediction.

The results obtained demonstrate the effectiveness of the prediction models developed, showing a good fit to the training data and a promising generalization capability. We also observed how in the visualization technique the approach is very promising and helps to understand the distribution of the data and its analysis. These models can be of great use in predicting the lifetime of LiBs in different applications, allowing informed decisions on the replacement, reuse or recycling of these batteries.

In addition, several limitations and areas for improvement have been identified in this study in order to be able to further investigate by the identified branches in future studies. Also, it is proposed to evaluate the feasibility of implementing a circular trade approach in the context of LiBs, promoting the reuse and recycling of components to reduce the environmental impact and foster the circular economy.

In conclusion, this work contributes to the advancement of knowledge in the field of lithium batteries and their lifetime. The predictive models developed and the suggestions put forward for future research offer opportunities to improve the efficiency and sustainability of lithium batteries, paving the way towards a cleaner and more sustainable energy future.

9.2 Implications and contributions of the work

The present work has important implications and contributions in the LiBs industry and in the research of LiBs aging prediction. First, the results obtained with the regression models and neural networks are very promising and can be useful for the industry in predicting the performance of batteries and in making decisions related to their maintenance and replacement. As well as indirectly it also affects the whole sector since giving this new life generates a very competitive second hand market.

In addition, this work contributes to the promotion of the circular economy and the recycling of LiBs components. The growing demand for batteries for electric vehicles and renewable energy storage makes the sustainable and efficient management of used batteries increasingly important. Predicting the lifetime of batteries and identifying the parameters that affect their performance can improve end-of-life planning of batteries, which in turn facilitates their recycling and contributes to a more circular economy. In addition, the identification of factors that influence battery lifetime can also guide the design and manufacture of more

durable and efficient batteries, which in turn can have a positive impact on the sustainability and cost-effectiveness of energy storage systems.

Moreover, this work contributes to research in the field of battery aging prediction, as different modeling techniques have been used and the results obtained have been compared with different regression models and neural networks. This can help researchers to choose the most appropriate modeling technique depending on the data and variables under study. In addition, a large and varied data set has been used, which increases the validity and generalizability of the results obtained.
Also, limitations and possible future steps in the research have been identified, which can guide researchers in improving and expanding this work.

Ultimately, this work can help to obtain another approach in battery industry and research, opening a new space to be considered and can serve as a basis for future studies in this field.

9.3 Limitations and future research

When studying the different limitations that we have found when developing the project, one of the main ones can be considered around the data. The data used have been based on the choice of a main set of batteries, as already mentioned, identified as "BatteryAgingARC-FY08Q4". This fact may condition the study since, although we are relying on multiple batteries, we have only based the study on one set, grouped by their main chemical characteristics. This fact may limit the results since we may be conditioned by the choice of this set and not another with other characteristics and attributes that may lead to different results or predictions.
In addition, following the thread of data limitations, we can find another limitation in the availability of these data. In spite of having used a wide and varied set of data, difficulties have been encountered in obtaining data for some relevant parameters. The two main attributes studied have been calculated from the formulas justified above (see section 7.2) and, although it is true that this is the most commonly used method, there are other methods for calculating these variables that require other more complex attributes.

Another possible limitation that we have encountered in the course of the work is the use of the learning techniques used. Despite being the learning techniques that have offered the best results of all the main techniques used, not each and every one of them has been evaluated. It could be interesting to study and hyperparameterize all of them to find those techniques that best fit the data.

On the other hand, as for future work, the main objective is to adapt the model to be able to classify the new samples into different groups according to the prediction made. The objective would be to be able to separate those batteries that can be used for a second life, those that probably can be used, those that need an overhaul and those that do not or better not to be used for a second life. In this way technical work can be saved and only those in which the model doubts if that battery has sufficient but not optimal conditions to be reused can be evaluated by a professional. In this way a classification would be established with a

higher degree of accuracy that would also allow to use those batteries with better condition for more demanding purposes in terms of the quality of this. In addition, more confidence would be established in the sector, granting degrees of reliability or health of the battery, where the buyer can be sure that the battery he buys at the price he buys it meets the requirements or is within the quality margins, thus strengthening the industry and the sector.

Secondly, it would be interesting to study the modeling of other algorithms that can optimally predict the main attributes of the LiBs. It would be enriching to have multiple alternatives and evaluate the one that gives the best results. On the other hand, and related to this aspect, it would be very beneficial for the study to incorporate new features to the data. New characteristics such as attributes of the LiBs that could be determinant for the calculation and prediction of variables such as RUL and SOH.

Finally, in the future it would be very beneficial to follow the path of visualization and the study of the models by means of three-dimensional projections such as the one already mentioned or other innovative technologies such as these. As could be seen, the visualization in 3D projection has provided a different and enlightening view of the data distribution offering a new starting point for further study. The projection using the T-SNE technique has provided us with a classification view using this model from the different clusters generated. Continuing this thread of research may bring new and innovative insights to this project and this industry.

Future work

As for future lines of work, there are several areas of research and development that could enrich and extend the results obtained in this study. First, an exploration of more advanced learning techniques would be of great interest. Although machine learning techniques have been used in this study, there are other approaches that could be explored, as well as the use of deep learning techniques, for example, given the good results obtained with neural networks, convolutional neural networks or recurrent neural networks could be investigated. These could allow the discovery of more complex patterns and relationships in the battery data. In addition, transfer learning techniques could be applied, where models pre-trained on similar data sets are adapted to improve the accuracy of predictions in this specific domain.

Another aspect to consider is the investigation of new attributes and features. Although various attributes and features have been used in the study conducted, there is always room to explore new indicators that could improve the predictive capability of the model. For example, additional environmental variables could be considered, such as battery temperatures, ambient temperature (measuring at different temperatures throughout the cycles) and humidity, which may have an impact on battery performance and lifetime. Also, the inclusion of additional information on the chemistry and internal structure of the batteries could be investigated, which could provide a more complete understanding of their behavior.

In relation to the data, it would be very beneficial for the project to conduct ongoing data collection with the goal of expanding and volumizing the project. Although a specific data set has been used in this study, it would be valuable to continue to collect information from additional batteries. This would allow the size of the data set to be expanded and improve the generalizability of the predictive models. In addition, by aggregating new data over time, more complex analyses could be performed to better understand the evolution of battery aging and improve long-term predictions. Following this thread, it would be interesting to be able to provide the study with validation and comparison with other approaches. Further validation of the models developed in this study could be performed, using different battery data sets or even testing in real test environments. This would allow evaluation of the robustness and generalization of the models in different contexts. In addition, the results obtained in this study could be compared with other existing approaches in the literature, such as physical models or methods based on specific algorithms for battery aging, in order to determine the relative effectiveness of each approach and identify possible areas for improvement.

Closing the scope of the study, it would be suggestive to perform an assessment of the feasibility of circular trading in this sector. A comprehensive analysis could be carried out on the feasibility of implementing a circular trade model in the context of used batteries. This would involve investigating the possibility of establishing agreements and collaborations with relevant companies and actors in the battery supply chain, such as manufacturers, distributors and recycling service providers. Options such as the reuse of second-life batteries in different applications or the implementation of buy-back and recycling programs by manufacturers could be explored. Assessing the economic, logistical and environmental feasibility of such initiatives would provide a better understanding of the benefits and

challenges of circular trade in batteries and its contribution to sustainability and the circular economy.

In summary, there are several interesting directions for future research in the field of LiBs and lifetime prediction. Exploring new modeling and data analysis techniques could open new doors to better understand battery behavior and improve prediction accuracy. All lines of research contribute significantly to the development of sustainable and efficient LiB solutions, driving the transition towards a circular economy and the mass adoption of clean energy.

References

[1] Intergovernmental Panel on Climate Change (IPCC). (2018). Special Report on Global Warming of 1.5°C. In IPCC, Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (Eds.). World Meteorological Organization.

[2] Jacobson, M. Z., Delucchi, M. A., Cameron, M. A., & Foddy, B. (2015). Renewable energy and energy efficiency: Pathways to sustainable energy. Annual Review of Environment and Resources, 40, 271-299.

[3] Graphite One Inc. (2021). Lithium Supply and Demand: A Strategic Overview. Technical Report, Vancouver, BC.

[4] Field, A. J., Frith, S., & Klaes, J. (2018). Lithium resources and supply for battery production. Nature Energy, 3(4), 314-323.

[5] Escalera, C., Cornejo, P., & Oyarzun, R. (2017). Environmental impacts of lithium production in South America: A review. Science of the Total Environment, 579, 552-568.

[6] Swedish Energy Agency. (2019). Forskningsöversikt om återvinning och återbruk av litiumjonbatterier [Research overview of recycling and reuse of lithium-ion batteries].

[7] Zhou, L., Chen, J., & Chen, Z. (2019). A review of the technical difficulties and solutions in second-life utilization of electric vehicle batteries. Journal of Power Sources, 408, 227234.

[8] Wang, X., Zhang, Y., & Li, J. (2021). Machine learning based state of health prediction for lithium-ion batteries: A review. Applied Energy, 282, 116880.

[9] Sharma, R. (s.f.). Battery Health NASA Dataset [Data set]. Restored from https://www.kaggle.com/code/rajeevsharma993/battery-health-nasa-dataset/notebook

[10] Zhang, J. X., & Wu, Q. H. (2011, December). A review on second life of electric vehicle batteries. In 2011 International Conference on Electric Vehicle Technology and Vehicle Engineering (pp. 538-543). Xiamen, China. doi: 10.1109/EVTE.2011.5710689.

[10] Jiang, C., Wang, B., Cui, L., & Lu, L. (2020). A novel multistage Support Vector Machine based approach for Li ion battery remaining useful life estimation. Journal of Energy Storage, 32, 101911.

[11] Kumar, A., Singh, J., Sharma, A., & Kumar, A. (2018). Artificial neural network based remaining useful life estimation of lithium-ion batteries in electric vehicles. Measurement, 116, 628-638.

[12] Sueldo: Project Manager en Barcelona, Spain 2023. (s. f.). Glassdoor.

[13] Sueldo: Junior Researcher en Barcelona, Spain 2023. (s. f.). Glassdoor.

[14] Sueldo: Junior Developer en Barcelona, Spain 2023. (s. f.). Glassdoor.

[15] Sueldo: Tester en Barcelona, Spain 2023. (s. f.). Glassdoor.

[16] Sueldo: Analyst en Barcelona, Spain 2023. (s. f.). Glassdoor.

[17] Tarifasgasluz. (2023, 24 mayo). Precio del kWh de Endesa.

[18] Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

[19] IBM Cloud. (s.f.). Supervised vs Unsupervised Learning. IBM Cloud Blog.

[20] IBM (n.d.). Classification and regression guidelines. IBM Watson Studio.

[21] IBM Garage. (s.f.). Evaluate and select machine learning algorithm. IBM Garage Method.

[22] IBM. (s.f.). How SVM Works. IBM SPSS Modeler SaaS Documentation.

[23] Ankit, K. (2019). Math behind Support Vector Machine (SVM). Medium.

[24] IBM. (s.f.). Random Forest. IBM.com.

[25] IBM. (s.f.). Neural Network. IBM.com.

[26] Biologic. (s.f.). Battery States: State of Charge (SoC) & State of Health (SoH). Biologic.net.

[27] Chen, L., An, J., Wang, H., Zhang, M., & Pan, H. (2020). Remaining useful life prediction for lithium-ion battery by combining an improved particle filter with sliding-window gray model. Energy Reports, 6, 2086-2093.

[28] Le, D., & Tang, X. (2011). Lithium-Ion Battery State of Health Estimation Using Ah-V Characterization. En Proceedings of the Annual Conference of the PHM Society, Montreal, QC, Canada, 25–29 de septiembre de 2011 (p. 1).

[29] Kim, I.S. (2009). A Technique for Estimating the State of Health of Lithium Batteries through a Dual-Sliding-Mode Observer. IEEE Transactions on Power Electronics, 25, 1013-1022.

[30] Jo, S. (2021). Battery State-of-Health Estimation Using Machine Learning and Preprocessing with Relative State-of-Charge. MDPI.

[31] Saha, S., Goebel, K., & Poll, S. (2011). Review of Prognostic Methods for Battery Health Management. Journal of power sources, 196(18), 7539-7549.

[32] NASA Open Data Portal. (2018). Li-ion Battery Aging Datasets.

[33] Google. (s.f.). Embedding Projector. http://projector.tensorflow.org/

[34] IBM. (s.f.). Graphics t-SNE. IBM.com.

Annexes

	cycle	capacity	voltage_measured	current_measured	\
0	2	1.856487	4.191492	-0.004902	
1	2	1.856487	4.190749	-0.001478	
2	2	1.856487	3.974871	-2.012528	
3	2	1.856487	3.951717	-2.013979	
4	2	1.856487	3.934352	-2.011144	
...	...	...	...	...	
50280	614	1.325079	3.579262	-0.001569	
50281	614	1.325079	3.581964	-0.003067	
50282	614	1.325079	3.584484	-0.003079	
50283	614	1.325079	3.587336	0.001219	
50284	614	1.325079	3.589937	-0.000583	

	temperature_measured	current_load	voltage_load	SOH
0	24.330034	-0.0006	0.000	1.000000
1	24.325993	-0.0006	4.206	1.000000
2	24.389085	-1.9982	3.062	1.000000
3	24.544752	-1.9982	3.030	1.000000
4	24.731385	-1.9982	3.011	1.000000
...	...	...	...	...
50280	34.864823	0.0006	0.000	0.713756
50281	34.814770	0.0006	0.000	0.713756
50282	34.676258	0.0006	0.000	0.713756
50283	34.565580	0.0006	0.000	0.713756
50284	34.405920	0.0006	0.000	0.713756

[50285 rows x 8 columns]

Figure 1

	cycle	capacity	voltage_measured	current_measured	\
0	2	1.856487	3.529829	-1.818702	
1	4	1.846327	3.537320	-1.817560	
2	6	1.835349	3.543737	-1.816487	
3	8	1.835263	3.543666	-1.825589	
4	10	1.834646	3.542343	-1.826114	
..	...	...	...	...	
163	600	1.293464	3.466462	-1.674488	
164	604	1.288003	3.468509	-1.667447	
165	608	1.287453	3.466806	-1.667470	
166	612	1.309015	3.471071	-1.688898	
167	614	1.325079	3.475472	-1.697928	

	temperature_measured	current_load	voltage_load	SOH
0	32.572328	-1.805570	2.404944	1.000000
1	32.725235	-1.804583	2.399260	0.994527
2	32.642862	-1.803575	2.397969	0.988614
3	32.514876	-1.812863	2.408289	0.988567
4	32.382349	-1.812876	2.408505	0.988235
..	...	...	...	...
163	33.275688	1.661799	2.073168	0.696726
164	33.320678	1.655086	2.064189	0.693785
165	33.373150	1.655103	2.062717	0.693488
166	33.713519	1.676430	2.107460	0.705103
167	33.865318	1.685264	2.120230	0.713756

[168 rows x 8 columns]

Figure 2

```
        cycle  voltage_measured  current_measured  temperature_measured  \
0           1          3.873017         -0.001201             24.655358
1           1          3.479394         -4.030268             24.666480
2           1          4.000588          1.512731             24.675394
3           1          4.012395          1.509063             24.693865
4           1          4.019708          1.511318             24.705069
...       ...               ...               ...                   ...
541168    616          0.236356         -0.003484             23.372048
541169    616          0.003365         -0.001496             23.369434
541170    616          4.985137          0.000506             23.386535
541171    616          4.984720          0.000442             23.386983
541172    616          4.213440         -0.000734             23.385061

        current_charge  voltage_charge
0                0.000           0.003
1               -4.036           1.570
2                1.500           4.726
3                1.500           4.742
4                1.500           4.753
...                ...             ...
541168           0.000           0.003
541169           0.000           0.003
541170           0.000           5.002
541171          -0.002           5.002
541172          -0.002           4.229

[541173 rows x 6 columns]
```

Figure 3

```
     cycle  voltage_measured  current_measured  temperature_measured  \
0        1          4.187420          0.643455             25.324079
1        3          4.058826          0.949043             26.635623
2        5          4.058139          0.950529             26.778176
3        7          4.058905          0.952312             26.703204
4        9          4.058330          0.947728             26.617004
..     ...               ...               ...                   ...
165    602          4.180892          0.476511             25.506487
166    606          4.181592          0.463218             25.517453
167    610          4.180125          0.493932             25.664855
168    613          4.180702          0.489857             25.433647
169    616          2.884604         -0.000953             23.380012

     current_charge  voltage_charge
0          0.638452        4.359487
1          0.941762        4.430904
2          0.943114        4.402619
3          0.944735        4.418979
4          0.940361        4.364055
..              ...             ...
165        0.472509        4.333942
166        0.459319        4.252485
167        0.489729        4.423386
168        0.486064        4.431494
169       -0.000800        2.847800

[170 rows x 6 columns]
```

Figure 4

	cycle	re	rct	sense_current	battery_current	\
0	41	0.044669	0.069456	-1.000000	-1.000000	
1	41	0.044669	0.069456	820.609497	337.091461	
2	41	0.044669	0.069456	827.242188	330.631561	
3	41	0.044669	0.069456	827.193481	330.808624	
4	41	0.044669	0.069456	824.929504	332.682678	
...	...	...	...	...	...	
13339	615	0.050036	0.074792	915.489014	230.149506	
13340	615	0.050036	0.074792	916.725525	212.188858	
13341	615	0.050036	0.074792	914.619629	176.598038	
13342	615	0.050036	0.074792	880.340820	136.847626	
13343	615	0.050036	0.074792	801.361816	97.058853	

	current_ratio	battery_impedance	rectified_impedance
0	1.000000	-0.438926	0.070069
1	2.320415	0.130088	0.068179
2	2.424193	0.058771	0.067933
3	2.447002	0.005814	0.066918
4	2.434305	0.126081	0.068071
...	...	...	...
13339	3.334835	0.245024	0.067925
13340	3.440393	0.264594	0.067925
13341	3.670656	0.288571	0.067925
13342	4.060164	0.317700	0.067925
13343	4.550338	0.352680	0.067925

[13344 rows x 8 columns]

Figure 5

```
     cycle        re       rct  sense_current  battery_current  current_ratio  \
0       41  0.044669  0.069456     811.176657       315.905164       2.493579
1       43  0.046687  0.076275     809.823676       316.504352       2.503652
2       45  0.044843  0.067972     810.495946       316.617010       2.489316
3       47  0.046195  0.074534     809.475062       317.225532       2.491553
4       49  0.045101  0.068528     810.923232       315.977955       2.495459
..     ...       ...       ...            ...              ...            ...
273    605  0.058801  0.082370     842.370181       308.431960       2.694519
274    607  0.057630  0.087424     845.925238       305.007640       2.731819
275    609  0.058984  0.082522     842.150261       308.301196       2.692602
276    611  0.057824  0.089757     845.754082       304.957511       2.734526
277    615  0.050036  0.074792     833.888655       316.960652       2.591606

     battery_impedance  rectified_impedance
0             0.171894             0.060823
1             0.171267             0.061974
2             0.171488             0.060362
3             0.170585             0.061330
4             0.172002             0.060840
..                 ...                  ...
273           0.220480             0.071815
274           0.224576             0.074615
275           0.220251             0.071730
276           0.224682             0.074486
277           0.208417             0.064925

[278 rows x 8 columns]
```

Figure 6