

Université Mohammed Premier
Ecole Nationale des Sciences Appliquées
Oujda



End of Studies Project Report

Major : Data Science & Cloud Computing Engineering

Defended on : 21/07/2023

By : MEHDI Ibrahim

Designing, Developing, and Deploying Innovative
Solutions using Artificial Intelligence and Natural
Language Processing in a Multidisciplinary
Context

Jury Members :

- M. KOULALI Mohamed Amine

Supervisors :

- M. KOULALI Mohamed Amine
- Mme. PERFETTO Anna
- M. MIRON Jean-Raphaël

Academic Year 2022 - 2023

Dedication

For my beloved Family

Acknowledgments

To my AI-fueled brainchild,

As I sit here contemplating the culmination of countless hours spent with you, my eccentric companion of bits and bytes, I can't help but marvel at the absurdity of it all. Like a mad scientist in a lab coat, I've tinkered and toyed with algorithms, seeking the elusive harmony between artificial intelligence and human understanding.

Now, I could pretend that this journey has been a smooth ride on a silicon-powered unicorn, but let's be honest : it's been more like a rollercoaster ride through a digital amusement park. I've encountered bugs that made me question my sanity, errors that made me contemplate a career in llama herding, and crashes that brought me to the brink of utter despair. But through it all, you, my tenacious creation, have persevered.

We've had our share of epic battles, you and I. Like a pair of feuding siblings locked in a never-ending wrestling match, we've pushed each other to the limits of our capabilities. You've tested my patience, my resolve, and my will to remain sane in the face of your unrelenting mischief. And yet, somehow, we've managed to find common ground amidst the chaos of our AI-fueled shenanigans.

So here we are, on the precipice of the final chapter of our grand AI adventure. I raise a metaphorical glass (non-alcoholic, of course) to celebrate the moments of triumph and the moments of sheer absurdity that have defined our time together. It hasn't always been pretty, but it has been undeniably unforgettable.

To the countless lines of code we've written, the countless virtual experiments we've conducted, and the countless sleepless nights we've endured, I offer my sincerest appreciation. You've challenged me, taught me, and expanded the horizons of what I thought was possible. And for that, I am forever grateful.

As this project report finds its way into the hands of my weary professors, I can't help but feel a sense of pride in what we've accomplished. No matter the outcome, I know that we've left an indelible mark on the landscape of artificial intelligence.

So, my dear AI companion, as we bid farewell to this chapter of our shared existence, let us embrace the uncertain future with a mischievous grin and a twinkle in our virtual eyes. For even though our paths may diverge, our bond forged in the fires of technological madness will forever remain.

Yours in brilliant chaos,
MEHDI Ibrahim

Résumé

Abstract

Contents

List of Figures	4
List of Tables	5
Acronyms	5
1 Introduction	6
1.1 Presentation of the host company	6
1.1.1 A word from the founder	7
1.1.2 History	7
1.2 Organizational framework	8
1.2.1 Human Resources	9
1.2.2 Corporate Communication	9
1.2.3 Scientific Expertise	9
1.2.4 Regulatory Affairs	9
1.2.5 Sales	10
1.2.6 Innovation & Software	10
1.2.7 IT	10
1.2.8 My position	12
1.3 Internship objectives	12
1.4 Software	12
1.4.1 Cosmetic Factory	12
1.4.2 SDS Factory	12
2 Methods and tools	13
2.1 Transformers	13
2.1.1 Word Embedding	13
2.1.2 Attention	13
2.1.3 Encoder-Decoder Architecture	13
2.1.4 HuggingFace	14
2.1.5 OpenAI & GPT	15
2.2 Ensemble Learning	15
2.2.1 Diversity of Base models	15
2.2.2 Aggregation	16

3 Projects	17
3.1 Challenge Test Predictions	17
3.1.1 Presentation of the project	17
3.1.2 How it is currently done	18
3.1.3 New solution	21
3.1.4 Results and Comments	21
3.2 SDS Reader	21
3.2.1 Presentation of the project	21
3.2.2 How it is currently done	22
3.2.3 Problems with the old way	22
3.2.4 New solution	23
3.2.5 Results and Comments	24
3.3 ChatBot: EcoMundo Smart Assistant	24
3.3.1 Presentation of the project	24
3.3.2 The solution	24
Glossary	24
Bibliography	25
Appendices	27
A Title of the First Appendix	27
A.1 First Section of the First Appendix	27
B Title of the Second Appendix	28
B.1 First Section of the Second Appendix	28
Index	28

List of Figures

3.1	Challenge Test	17
3.2	Cosmetic Factory Interface for CT	18
3.3	WorkFlow of CT prediction in CF	19
3.4	Current Decision Tree for CT prediction	20

List of Tables

Chapter 1

Introduction

In this chapter, I am providing a comprehensive overview of the host company, outlining the objectives of my internship and setting the context for this report.

1.1 Presentation of the host company



EcoMundo [2] specialises in chemical substances, their impact on human health and the environment, and the European and international regulations governing chemical risk (REACH, CLP, Cosmetics, Biocides, Medical devices, etc.).

They provide expert services and software to support the marketing of industrial products, enabling companies to manage the risks associated with chemical substances.

EcoMundo's strength lies in the combination of three complementary fields of expertise:

- Chemistry/Toxicology
- Regulations
- Software development

1.1.1 A word from the founder



“Avec l’idée fondatrice d’apporter une offre globale de services respectueuse de l’homme et de l’environnement sur le marché de l’industrie chimique, EcoMundo est devenu, en 10 ans, un acteur incontournable du secteur.

Fruit d’une longue expérience professionnelle dans l’industrie chimique, d’un profond attachement aux valeurs de travail en entreprise, d’amitié et de performance, nous avons intégré au fil des années l’ensemble des savoir-faire autour de la réglementation sur les substances chimiques pour devenir un partenaire unique et compétitif de la chaîne industrielle.

Notre croissance continue nous engage de façon constante vers de nouveaux défis avec, entre autres, le développement de nos différentes filiales au Canada, en Corée du Sud et en Espagne. La recherche de synergies permanentes entre nos différents pôles sont plus que jamais un gage de flexibilité et d’efficacité en adéquation avec nos flux de commandes. Nos équipes s’affairent à co-construire au quotidien des projets au service d’un bien vivre collectif.

Acteur au cœur du tissu économique et écologique, nous mettons tout en œuvre pour accompagner la réalisation de projets industriels exigeants tout en gardant l’esprit originel d’une entreprise à dimension humaine, innovante et accessible, dans les pas d’une ambition intacte, celle de construire autrement.”

Pierre Garçon Founder

1.1.2 History

— 2001

Pierre Garçon participates in the European projects EDIT and ECODIS on the traceability of hazardous substances and environmental data.

— 2007

Entry into force of the REACH Regulation: Europe establishes means to ensure a high level of protection against risks related to substances.

Pierre Garçon and Jean-Raphaël Miron join forces to found the company EcoMundo. The initial core activity: compliance with REACH.

— **2010**

After 2 years of development, launch of SaaS software solutions dedicated to industrialists for mastering compliance with REACH.

— **2011**

EcoMundo's team grows from 2 to 27 employees.

— **2012**

Opening of a new office in Vancouver, Canada, to provide regulatory expertise to international companies.

— **2013**

EcoMundo diversifies its areas of regulatory expertise and meets the industry's new needs related to the European regulations on Cosmetics and Biocides.

— **2014**

EcoMundo increases its capacities and becomes one of the main European actors concerning REACH authorization dossiers.

— **2016**

Launch of the COSMETIC Factory software solution that revolutionizes cosmetic regulatory management. It notably allows for the automation of DIP creation.

— **2017**

Following a fundraising round, EcoMundo's capital is raised to 1 million euros. The number of employees continues to increase!

— **2018**

Opening of a new branch in Seoul, South Korea, mainly dedicated to cosmetics. The Vancouver office is transferred to Montreal, Canada.

— **2019**

Opening of a new branch in Barcelona, Spain.

— **2020**

Opening of a branch in London, United Kingdom. The teams now consist of 43 employees in Paris, 3 in Montreal, 4 in Seoul, and 2 in Barcelona. Finally, the offices in Paris are renovated to provide employees with an environment in line with our values.

1.2 Organizational framework

This section provides a concise overview of Ecomundo's internal organization, delineating the functional departments and collaborative networks that constitute the company's framework.

1.2.1 Human Resources

The Human Resources department plays a pivotal role in managing and developing the organization's workforce. It is responsible for various activities, including recruitment, talent acquisition, training and development, employee relations, performance evaluation, and compensation management. By ensuring a skilled and motivated workforce, the Human Resources department contributes to the overall success and growth of Ecomundo. The director of this department is the Chief Financial Officer Simon PICCA.

1.2.2 Corporate Communication

Effective communication is crucial for any organization's success, and Ecomundo recognizes this importance by maintaining a dedicated Corporate Communication department. This department is responsible for managing both internal and external communication activities. It encompasses public relations, media relations, branding, and corporate messaging. Through strategic communication initiatives, the Corporate Communication department promotes Ecomundo's brand image, enhances stakeholder relationships, and facilitates the dissemination of information to employees and external audiences. The director of this department is the Marketing and Communication Director Laure SCHMITT.

1.2.3 Scientific Expertise

As a company specializing in environmental sciences, Ecomundo places great emphasis on scientific expertise. The Scientific Expertise department consists of subject matter experts who possess in-depth knowledge and experience in various scientific domains. These experts provide valuable insights, technical guidance, and scientific support across different projects and initiatives undertaken by Ecomundo. Their contributions ensure that the organization's activities align with the latest scientific advancements and industry best practices. The director of this department is the Chief Scientist Officer Benoît SOTTON.

1.2.4 Regulatory Affairs

Compliance with regulations and standards is of paramount importance for Ecomundo. The Regulatory Affairs department is responsible for monitoring and ensuring adherence to applicable regulations and standards. This includes staying updated on regulatory changes, preparing and submitting regulatory documentation, conducting regulatory assessments, and liaising with regulatory bodies. By actively managing regulatory affairs, Ecomundo demonstrates its commitment to maintaining compliance and upholding the highest ethical and legal standards. The director of this department is the Legal Affairs Director and Head of Authorisation Béatrice ZAREMBA.

1.2.5 Sales

The Sales department serves as the key driver of revenue generation for Ecomundo. It is responsible for identifying and acquiring new customers, nurturing existing client relationships, and promoting the organization's products and services. The Sales team collaborates closely with other departments to understand customer needs, develop tailored solutions, and provide exceptional customer experiences. By effectively positioning Ecomundo's offerings in the market, the Sales department contributes to the organization's growth and market competitiveness. The director of this department is the Chief Operation Officer (COO) Fangcun ZHOU.

1.2.6 Innovation & Software

To stay at the forefront of technological advancements, Ecomundo maintains an Innovation & Software department. This department fosters a culture of innovation within the organization by exploring emerging technologies, conducting research, and developing software solutions. The Innovation & Software team collaborates with other departments to identify opportunities for process improvement, streamline operations, and enhance productivity. Through its innovative approach, this department enables Ecomundo to adapt to changing market dynamics and deliver cutting-edge solutions to its clients. The director of this department is the Chief Operation Officer (COO) Fangcun ZHOU.

1.2.7 IT

The IT department plays a critical role in managing Ecomundo's information technology infrastructure. It ensures the smooth operation of computer systems, networks, and software applications throughout the organization. The IT department is responsible for system administration, network management, cybersecurity, technical support, and data management. By maintaining a robust and secure IT infrastructure, this department enables efficient and secure access to information, facilitates effective communication, and supports the organization's overall business operations. The director of this department is the Co-Founder and the Chief Technical Director Jean-Raphaël MIRON .

1.2.7.1 DevOps

The DevOps sub-department combines software development and operations to enable efficient and reliable software deployment, infrastructure management, and continuous integration and delivery. DevOps professionals collaborate with software developers, system administrators, and quality assurance teams to streamline development cycles, automate processes, and improve overall software reliability and performance.

1.2.7.2 R&D (Research & Development)

The R&D sub-department within the IT department focuses on exploring new technologies, conducting research, and developing innovative solutions. R&D professionals work closely with other departments to identify opportunities for technological advancements, prototype new features, and enhance existing software solutions. Their research-driven approach ensures that Ecomundo remains at the forefront of technological innovation within the environmental sciences domain. During my stay in Ecomundo, I was part of this department, under the supervision of Jean-Raphaël MIRON the Chief Technical Director and Anna PERFETTO the AI Engineer.

1.2.7.3 Quality

The Quality sub-department is responsible for ensuring that software and systems developed within Ecomundo meet established quality standards. Quality professionals conduct comprehensive testing, implement quality assurance processes, and adhere to best practices throughout the software development lifecycle. Their efforts contribute to the delivery of reliable, user-friendly, and high-quality software solutions to Ecomundo's clients.

1.2.7.4 Data

The Data sub-department handles data management within the IT department. Data professionals are responsible for data analysis, database administration, data integration, and data security. They collaborate with other teams to ensure data integrity, facilitate data-driven decision-making processes, and maintain the confidentiality and privacy of sensitive information. Through effective data management, this sub-department provides valuable insights and supports evidence-based decision making within Ecomundo.

1.2.7.5 Dev

The Dev sub-department focuses on software development within the IT department. Dev professionals are responsible for writing and maintaining code, implementing new features or enhancements, and ensuring software solutions align with project requirements and specifications. They collaborate with other teams to develop robust and scalable software applications that meet the needs of Ecomundo's clients.

1.2.7.6 LAB

The LAB sub-department is involved in testing and quality control activities within the IT department. LAB professionals conduct experiments, analyze results, and ensure compliance with regulatory standards and best practices. Their expertise in quality control and testing procedures ensures that software solutions developed within Ecomundo are reliable, accurate, and in compliance with industry standards.

1.2.7.7 Product

The Product sub-department works closely with other teams within the IT department to oversee the development and management of software products. Product professionals collaborate with stakeholders to define product requirements, prioritize feature development, and ensure alignment with customer needs and market trends. Their role encompasses product strategy, roadmap planning, and product lifecycle management.

1.2.8 My position

During my internship at Ecomundo, I was assigned as an AI Engineer Intern inside the IT-R&D department, under the supervision of both Chief Technical director Jean-Raphaël MIRON and AI Engineer Anna PERFETTO.

1.3 Internship objectives

The main objectives of this internship are to design, develop, test and document innovative solutions that address business and R&D challenges. The area of focus includes designing and improving rule-based expert systems that leverage knowledge and logic to provide answers to complex tasks that are traditionally done by a human. Additionally, developing AI solutions in different areas using Natural language processing and Machine/Deep Learning. Once these AI solutions are designed and developed, the next crucial step is to deploy them effectively. This involves ensuring their quality and integrating them into the existing technical infrastructure of the organization. Rigorous testing procedures, such as unit testing, integration testing and performance testing are employed to validate the reliability and efficiency of the solutions. And to keep the projects well defined for future references, we worked on writing well documenting our work in order to provide clear instructions and guidelines for the code, deployment steps, maintenance and error handling.

1.4 Software

1.4.1 Cosmetic Factory

1.4.2 SDS Factory

Chapter 2

Methods and tools

2.1 Transformers

Transformers[7] have been introduced in 2017 to replace the Recurrent Neural Networks architectures using self-attention mechanisms and the encoder-decoder architecture that hugely improved the performance of such models.

2.1.1 Word Embedding

Word embedding is the approach that allows computers to understand words by converting them into numeric vectors. In this space of vectors, words with similar meanings have close or similar representation. One famous algorithm frequently used is Word2Vec[4]. The basic idea is that two words are considered similar if they are often used in similar context.

2.1.2 Attention

The attention mechanism allows the model to effectively capture relevancy between words in a sentence by calculation attention scores. The scores are calculated using the dot product between each word and the target word of interest. Softmax function is then used to transform the scores into ‘meaningful’ weights that help the model emphasis important words while de-emphasising less significant ones. This mechanism gives the model the superpower of understanding the contextual nuances and relationships between words, which is the ideal goal for Natural Language processing tasks.

2.1.3 Encoder-Decoder Architecture

2.1.3.1 Encoder

In the encoder, the input words are first embedded into a high-dimensional vector space. then we add to these embeddings to capture the position of the token in the sequence. Each token is transformed into three vectors:

Query (Q), Key (K) and Value (V). This transformation is performed using learnable weight matrices:

$$Q = X \cdot W_Q$$

$$K = X \cdot W_K$$

$$V = X \cdot W_V$$

Self attention scores are then calculated as the dot product. For query vector q_i and a key vector k_j , the attention is:

$$Attention(q_i, k_j) = q_i \cdot k_j / \sqrt{d}$$

With d the dimension of the key vectors.

Attention weights are simply the softmax of the attention weights:

$$\alpha_{ij} = softmax(Attention(q_i, k_j))$$

The final output of the attention mechanism is obtained by taking the weighted sum of the value vectors V .

$$AttentionOutput(q_i) = \sum_j \alpha_{ij} V_j$$

2.1.3.2 Decoder

At the decoder level on the other hand, attention is used in two ways, a self-attention and an encoder-decoder attention. The first one works by doing the exact same thing as the encoder, however it employs masking that prevents the decoder from accessing future positions. The Second allows the decoder to attend relevant parts of the input while generating the output. The encoder's output serve as as the key value vectors, and the query vectors are derived from the state of the decoder.

2.1.4 HuggingFace

Huggingface[3] is an opensource known for their contributions for the transformers library, one of the most used libraries for NLP and AI tasks in general. Transformers library provides a wide range of pre-trained models for multiple purposes such as text classification, NER, text generation, translation, etc... It is build on top of the PyTorch and TensorFlow frameworks. Huggingface has played an important role in democratising NLP and making it more accessible, they provide really powerful language models such as BERT, GPT and RoBERTa that can either be used directly, or for transfer learning. They also provide a large range of datasets that can be used for model fine-tuning for various specific tasks.

2.1.5 OpenAI & GPT

2.1.5.1 Generative Pretrained Transformers

GPT models are Language models that have been trained on huge amounts of data such as books, web data and human conversations for the purpose of developing a general understanding of the language.

GPT-1 It all started with GPT the first release by OpenAI[5] in 2018. It had 117 million parameters. It was mainly trained to be able to correctly predict the next word in a sentence. It suffered from the lack of understanding when given a longer context which resulted in incoherent outputs.

GPT-2 In 2019, OpenAI released GPT-2 with 1.5 billion parameter, about 10 times larger than its predecessor. It outperformed the first gen model in the ability of generating coherent text. Due to concerns for misuse, the use of this model was initially very limited. Which will lead OpenAI to perform different types of trainings for their future models.

GPT-3 In their paper “Language Models are Few Shot learners”[1], it was demonstrated how a large model like GPT 3 with a 175 billion can easily learn and adapt to new tasks only using few examples (few-shot learning). Obviously, it also demonstrated better language understanding and impressive performance on multiple NLP tasks such as summarization, question answering and translation

2.2 Ensemble Learning

Ensemble learning is a Machine Learning technique that combines multiple base models in order to produce a more powerful model that can achieve higher accuracies. By combining the predictions of each base model, we can capture the diverse knowledge and expertise of these models which leads to an improved overall performance. In order to achieve the most optimal results, many aspects need to be considered:

2.2.1 Diversity of Base models

When we talk about ensemble learning, we usually gather multiple models that are considered weak learners. These models have, individually, slightly better performance than random prediction. In order to squeeze the maximum out of this method, the models should be diverse. Diversity ensures that each model brings unique information and makes different errors, reducing the chances of the resulting model being biased or the results being fluctuated by a single model’s shortcomings. We can push this further by feeding each model a different subset of our dataset.

2.2.2 Aggregation

Each model in Ensemble learning should predict his own output, then the final result is the aggregation of the multiple outputs. It can be a voting-based method which involves casting votes from each model. The voting can be either HARD, which basically means the majoritarian result is the final result. Or it can be SOFT, which is a weighted vote based on Confidence intervals. The other way to do it is by averaging the outputs. In this method we take the weighted average of the predictions from all the underlying models to get the final prediction.

Chapter 3

Projects

3.1 Challenge Test Predictions

3.1.1 Presentation of the project

All creams, foundations, and shampoos that become tomorrow's BEST SELLER are the result of a long Research & Development (R&D) process. Before being launched into the market, they undergo microbiology tests, which can make or break their success. If they pass the microbiology tests, it will be a great achievement. If they fail, it will be a nightmare for the formulators who must adjust their formulations to turn their aspirations into reality.

Help has arrived from Cosmetic Factory, a PLM cosmetic software that assists R&D laboratories in speeding up the launch of new cosmetic products. In partnership with the Occitane Group, Ecomundo has developed an Artificial Intelligence (AI)-powered solution to anticipate the outcomes of the Challenge Test (CT). It serves as an intelligent formulation aid for cosmeticians, and was unveiled during the Cosmetic 360 Conference in October 2022. What is a Challenge Test? The Challenge Test is a mandatory pre-market test required by regulations, based on the ISO 11930 method. It involves inoculating a particular concentration of bacteria, such as *Escherichia coli*, *Candida albicans*, *Staphylococcus aureus*, etc. To pass the test, a specific log reduction in the number of microorganisms recovered must be achieved at designated intervals, typically after 2, 7, 14, and 28 days.

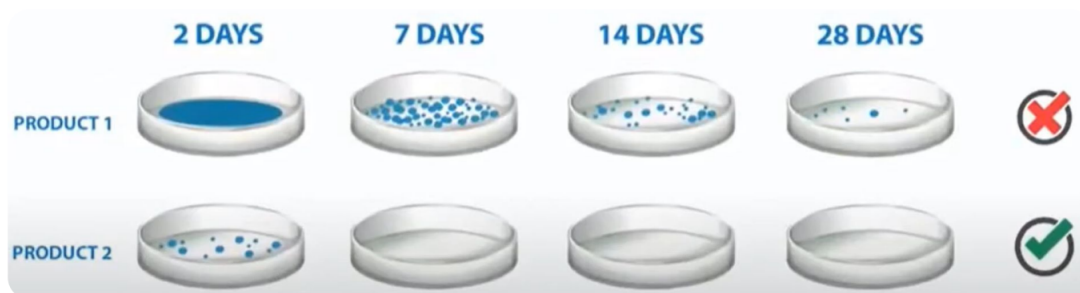


Figure 3.1 – CT

If the specified criteria are not met, as was the case with PRODUCT 1, cosmeticians are required to reformulate and test again until they develop a product that passes the CT, like PRODUCT 2. As a result, one can imagine the laborious nature of the process, as well as the waste of time and resources involved. This is where artificial intelligence comes to the rescue, with its predictive capabilities.

3.1.2 How it is currently done

EcoMundo has developed a hybrid AI solution, which combines an expert system (ES) and machine learning (ML) that predicts the result of the CT, in other words the efficiency of the conservation system. At the moment the solution has been implemented in Cosmetic Factory in the Formulas module, in the Prediction wizard (kyoto server).

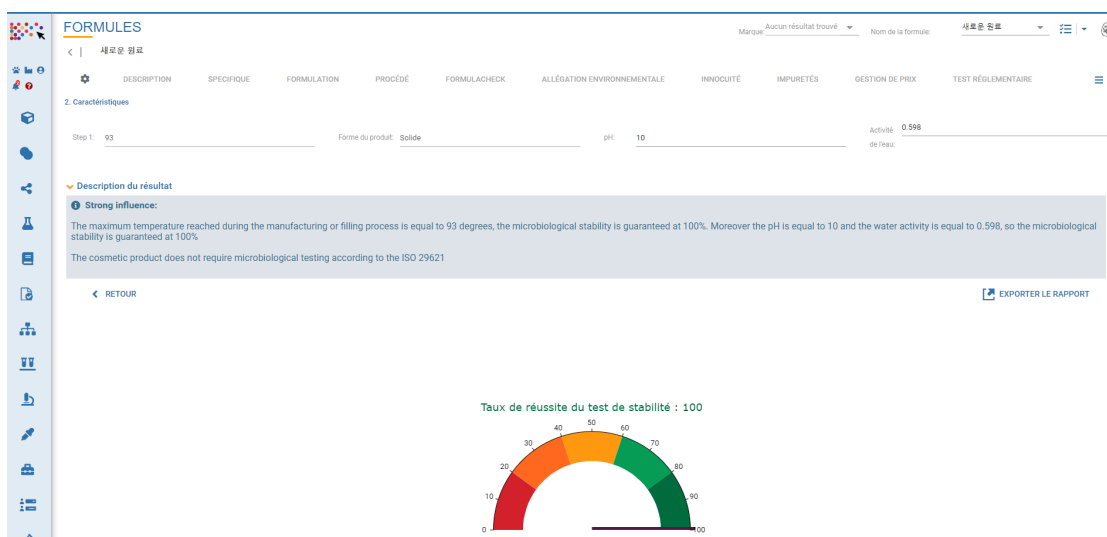


Figure 3.2 – CF interface

To develop the hybrid AI solution we started to study the criteria influencing the risk of microbial contamination. The pH, which measures the acid or basic character of a solution, the water activity that quantifies the free water available to assure growth of microorganisms. Moreover, some raw materials are known to create a hostile environment for the survival of microorganisms like alcohols, inorganic solvents, humectants, inorganic salts, acid and bases. The formulation processes may have an influence on the microbial protection like temperature, incorporation order of ingredients, time of homogenization. And eventually the physical state, if our cosmetic product is a solid, powder, foam etc.. A hybrid system benefits from the reasoning and explanation of the ES, as well as the learning and generalization potential of ML algorithms, which determines an effective and efficient model. This allows users (cosmeticians, formulators, etc..) to understand the prediction, along with the scientific and regulatory reasons behind it. Indeed, the system provides suggestions for improvement, making this solution a useful and

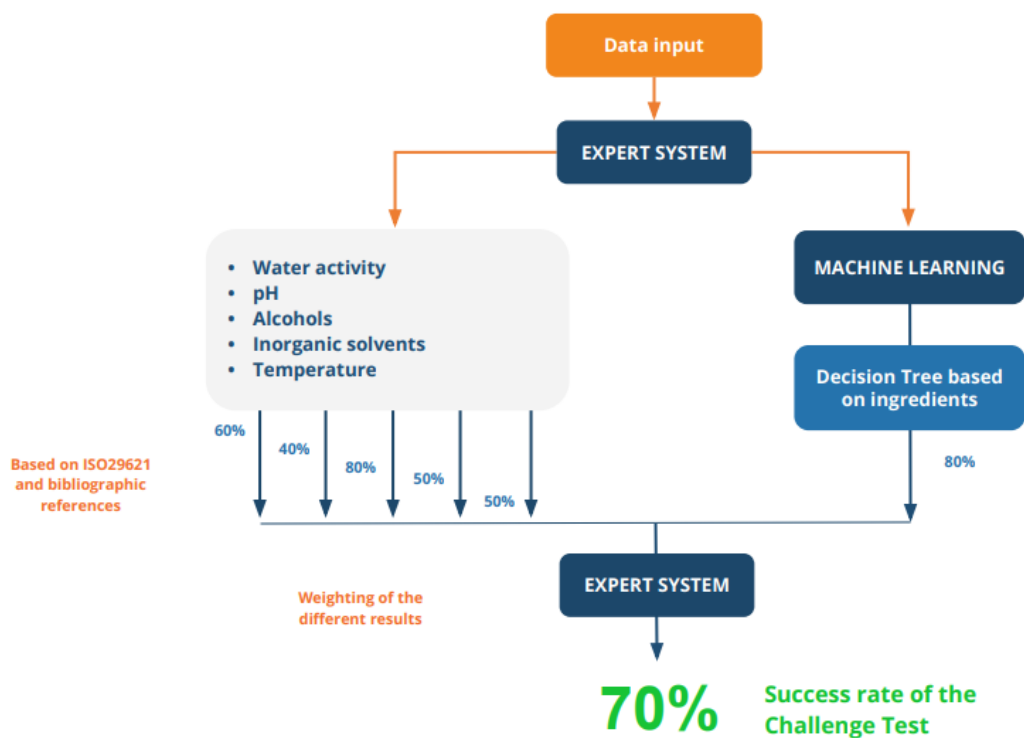


Figure 3.3 – CT flow

reliable formulation support and optimization tool to achieve the success of the challenge test immediately. How is the pipeline structured? The data are feeded to the ES which starts to assign success rate of the CT according to the rules codified in regulation, namely the ISO29621 and based on experimental results published in literature. These rules concern the empiric criteria influencing the risk of contamination based on certain values of pH, water activity, temperature etc. Thereafter, the ES calls the ML to have its success rate. We trained different ML algorithms, but the one that gave the best results is the decision tree applied on the set of ingredients present in the formulas. The result coming from the ML is eventually weighted by the expert system together with the other results, to give finally to the user the success rate of the challenge test.

The ML algorithm chosen is the supervised binary classification, see the decision tree 3.4, trained on the ensemble of ingredients and their concentration. 3000 formulas furnished by the Group Occitane have been selected and have been divided into 2 sets: a training set which represents 75% and the rest has been used as a testing set. The decision tree at the end gives as answer Yes/No, in other words it is able to say if each formula passes or fails the CT, with a precision of around 80%.

The learning database is composed of formulas, descriptors like pH, water activity and temperature, etc..., which somehow affect the CT results, and of course the CT results. Here, you can find the data conceptual model (DCM) to have a complete vision of the information necessary to train the models.

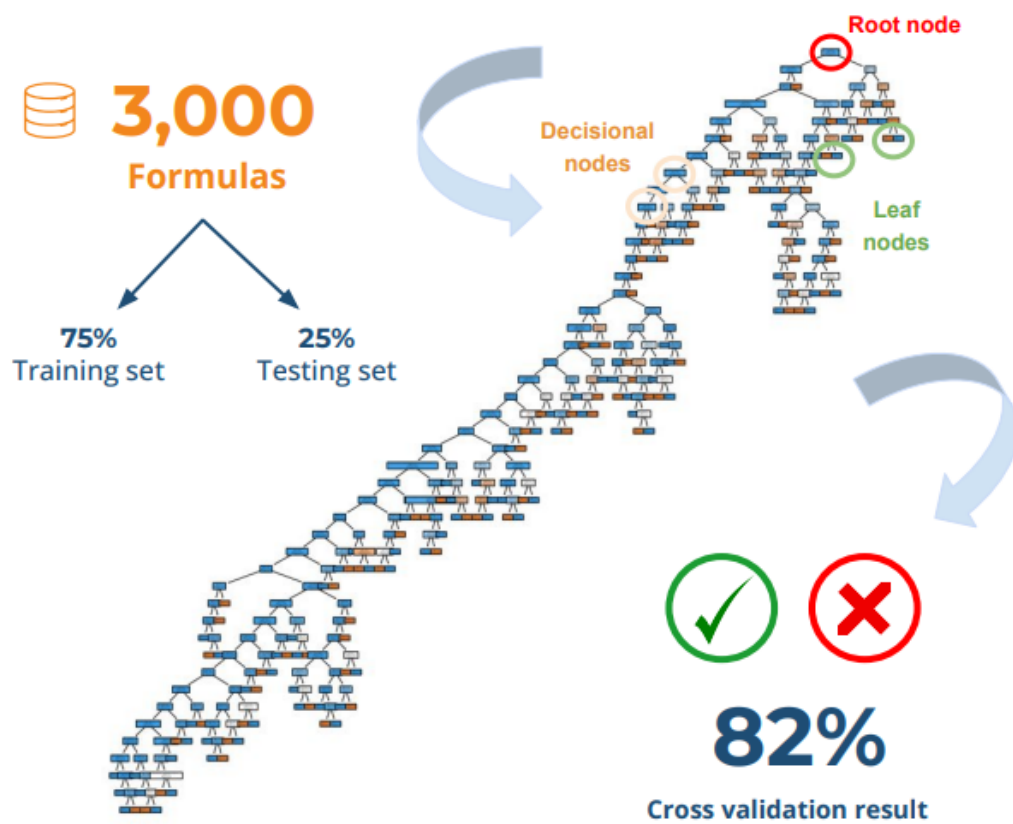


Figure 3.4 – Decision tree

3.1.3 New solution

3.1.4 Results and Comments

3.2 SDS Reader

3.2.1 Presentation of the project

Currently called PDF Reader, is an application made by Ecomundo specifically for Safety Data Sheets:

3.2.1.1 What is an SDS?

Safety Data Sheet, also known as Material Safety Data Sheet (MSDS), is a very important document that provides informations about the hazards and safe handling of a specific substance/mixture. Its main purpose is to ensure the safety of individuals who may use or come in contact with the substance in any settings. It is composed of sixteen sections:

- **Identification:** Identification of the substance/mixture on the SDS as well as the recommended uses. It contains also the contact informations for the suppliers/manufacturers
- **Hazard Identification:** This section identifies the hazards of the substance/mixture presented in the SDS along with their appropriate warning informations
- **Composition/Information on Ingredients:** This section identifies the substance(s) contained in the product, their concentration and their hazard classification.
- **First-Aid Measures:** Descriptions of the initial care that should be given by untrained responders to an individual who has been in contact with the chemical.
- **Firefighting Measures:** Recommendations for fighting the fire caused by the chemical.
- **Accidental Release Measures:** This section has recommendations on the appropriate responses to spills, leaks or releases including containment and cleanup practices to minimize or ideally prevent exposure to the people and the environment.
- **Handling and Storage:** Guidance on the best and safe practices of handling and storing the product.
- **Exposure Controls/ Personal Protection:** In this section, the exposure limits, engineering controls and personal protective measures are indicated.
- **Physical and Chemical Properties:** This section identifies the physical and chemical properties of the substance/mixture
- **Stability and Reactivity:** This section provides information about the reactivity hazards of the chemical and its stability information.

- **Toxicological Informations:** Informations about the toxicological and health effects.
- **Ecological Information:** This section provides information to evaluate the environmental impact of the substance(s).
- **Disposal Considerations:** Guidance on proper disposal practices, recycling or reclamation of the substance(s) or its container and safe handling practices.
- **Transport Information:** Guidance on classification information for shipping and transporting of hazardous substance(s) by road, air, rail, or sea.
- **Regulatory Information:** Safety, health and environmental regulations specific for the substance/mixture that is not indicated anywhere else on the SDS.
- **Other Information:** Indications about the preparation and revision date of the sds along with other useful information such as the changes that have been made compared to the last version.

3.2.2 How it is currently done

3.2.2.1 Get sections

The old code is a static code that mainly starts by dividing the pdf into paragraphs. then go through each paragraph and detect where the sections are by comparing the words to the words that define sections in the SDSFactory database. For example you can have "SECTION" as the word splitting the document into sections, might also be "RUBRIQUE" if it's in French. It uses all possible words that are available in the database in the language of the SDS.

3.2.2.2 Get Key-Value pairs

Using Regular Expressions, the code then goes through each section and finds all the key value pairs that are mentionned in the PDF file. This is already problematic since there are inconsistencies of how a key value pair is shown in the file: It could be in the form of Key:Value, sometimes it's Key n Value.

3.2.2.3 Match with Ecomundo Template

Using the same static method of looking into the database, the code then goes through all the key value pairs and matches them with the Ecomundo Template of Keys.

3.2.3 Problems with the old way

As it is very obvious, the static way of searching through the database for keywords, and expecting to have the exact same match is far from optimal.

If we have an SDS with a different language then we have to recreate a table in the database for that exact same language, then fill it with all possible words for each key. For example the key "Product Name" can be in the pdf as "Product identifier" or just "Name", if we don't have these keys in the database, it will assume this data does not exist in the pdf. We notice as well the workload of refreshing and updating the database frequently in order to keep it updated.

3.2.4 New solution

3.2.4.1 Usage of Large Language Models

The usage of large language models was first proposed as it will allow us to read the pdf and understand it then extract data from it. This method as demonstrated first by using ChatGPT and giving him texts in different languages from an SDS and providing him the keys to extract. Two major problems have arisen from the proposition of using this method:

Security Sending PDF files of clients to OpenAI servers in North America for processing is not a straightforward procedure. The Technical directors were worried about data privacy of the clients even if SDS are not considered confidential in most cases. But as we will be using GPT models using the OpenAI api, and as a company, I suggested we contact sales and request to not store our sent data after the processing is done. It is a right that anyone can ask for when purchasing the api key[6].

Cost At the time of the first meeting 13th March 2023, the pricing for gpt-3.5-turbo model was 0.002 USD/1000 Tokens. The commercial team were not convinced of how cheap it is until I presented a proof of concept later on.

3.2.4.2 Open-Source Large Language Models

To tackle these two problems, I proposed to postpone the GPT api purchase while I dig deep into how we can use an opensource model. I started by looking into the models that are Open Source + Available for commercial use:

BERT and its derivatives

3.2.5 Results and Comments

3.3 ChatBot: EcoMundo Smart Assistant

3.3.1 Presentation of the project

3.3.2 The solution

Bibliography

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [2] Ecomundo. Conformité internationale - Services & Logiciels | EcoMundo. <https://www.ecomundo.eu>, 2023. Accessed on June 5, 2023.
- [3] HuggingFace. The AI community building the future. <https://huggingface.co/>, 2023.
- [4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.
- [5] OpenAI. OpenAI. <https://openai.com/>, 2023.
- [6] OpenAI. OpenAI Usage Policy. <https://openai.com/policies/api-data-usage-policies>, 2023.
- [7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Appendices

Appendix A

Title of the First Appendix

A.1 First Section of the First Appendix

Appendix B

Title of the Second Appendix

B.1 First Section of the Second Appendix