UQAC
Université du Québec
à Chicoutimi

PRÉDICTION DE LA CONCENTRATION DE MATÉRIAUX DANS LES FORMULATIONS COSMÉTIQUES À L'AIDE DE MODÈLES D'APPRENTISSAGE AUTOMATIQUE

PAR CHRISTIAN GONZALO FRANTZ SEGOVIA

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI EN VUE DE L'OBTENTION DU GRADE DE MAÎTRE ÈS SCIENCES (M.SC.) EN INFORMATIQUE

QUÉBEC, CANADA

RÉSUMÉ

L'industrie des produits de beauté détient une position considérable sur la scène mondiale, générant des ventes dépassant des centaines de milliards de dollars américains à l'échelle mondiale. Ce marché est hautement compétitif, caractérisé par la domination des principaux acteurs mondiaux. La détermination des concentrations idéales d'ingrédients est une procédure importante dans le secteur des cosmétiques pour les formulations chimiques, dans le but de garantir la qualité, l'efficacité et l'économie liées aux produits développés. Dans ce contexte, l'interdépendance complexe entre les matériaux représente un défi significatif, exigeant une attention particulière pour la prédiction des concentrations de matériaux, afin d'éviter les inefficacités et les réactions indésirables éventuelles déclenchées par le produit final. La capacité à prédire avec plus de justesse les concentrations d'ingrédients permet la sécurité et l'efficacité, bien que la détermination des concentrations appropriées pour chaque matériau nécessite une évaluation prudente et pondérée. La détermination de la concentration des matériaux dans les formulations cosmétiques est un processus complexe qui commence par la formulation de la recette, en tenant compte des réglementations du secteur, qui établissent des limites de sécurité et d'efficacité pour les matériaux utilisés. La concentration des matériaux dans les formulations cosmétiques impacte directement leur stabilité et leur efficacité. Par conséquent, suggérer des concentrations de matériaux pour les formulations chimiques doit tenir compte de ces défis inhérents à la détermination des concentrations de matériaux chimiques et cosmétiques. Le domaine de l'apprentissage automatique est inséré dans le contexte de l'*Intelligence Artificielle* (IA) et implique l'application d'algorithmes informatiques pour transformer des données empiriques en modèles utilisables. Ces algorithmes permettent de comprendre les propriétés des ensembles de données analysés, en abstrayant les motifs sous-jacents à travers un modèle, en prédisant les valeurs inconnues basées sur le modèle généré et en détectant les comportements anormaux observés. Son objectif principal est de développer un modèle qui présente de hautes performances non seulement pendant l'entraînement, mais aussi lors de son application à un ensemble de tests ou à de nouvelles données. Jusqu'à présent, nous ne connaissons pas dans la littérature des recherches ayant utilisé des modèles d'apprentissage automatique pour suggérer la concentration de matériaux dans les formulations cosmétiques. Notre objectif est d'identifier la méthodologie optimale pour prédire la concentration de matériaux, de manière à ce qu'elle puisse être appliquée comme recommandation dans la production de formulations cosmétiques. Pour atteindre nos objectifs, nous avons utilisé quatre algorithmes d'apprentissage automatique : *Random Forest Regressor* (RFR), *Extreme Gradient Boosting* (XGBoost), *k-Nearest Neighbors* (k-NN) et *Multi-Layer Perceptron* (MLP). Nous avons sélectionné des mesures de performance, telles que *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE) et Coefficient de détermination (R^2), pour garantir une évaluation robuste et fiable des modèles proposés. Nous avons extrait et modélisé un total de 1679522 enregistrements, avec lesquels nous avons entraîné et testé les quatre modèles d'apprentissage automatique. La troisième approche a été utilisée pour effectuer les prédictions, car parmi les approches développées, c'était celle qui obtenait les meilleurs indicateurs de performance. Nous avons

également développé une application où l'utilisateur final pourra effectuer les prédictions à travers une *Interface Graphique Utilisateur* (IGU). Les résultats obtenus indiquent que le modèle RFR a présenté les meilleurs résultats parmi les modèles testés, avec une valeur de R^2 de 0,66892, démontrant que le modèle est capable d'expliquer environ 66,89% de la variabilité des données. Cette étude représente une étape importante vers le développement de modèles prédictifs pour l'industrie chimique et cosmétique, mettant en évidence l'importance de l'application de techniques d'apprentissage automatique et de validation croisée dans la résolution de problèmes dans ce domaine.

ABSTRACT

The beauty products industry holds a considerable position in the global landscape, generating sales that exceed hundreds of billions of US dollars worldwide. This market is highly competitive, characterized by the dominance of major global players. Determining the ideal concentrations of ingredients is an important procedure in the cosmetics sector for chemical formulations, aiming to ensure the quality, efficacy, and economy related to the developed products. In this context, the complex interdependence among materials represents a significant challenge, requiring special attention to predicting material concentrations to avoid inefficiencies and potential undesired reactions triggered by the final product. The ability to predict ingredient concentrations more accurately allows for safety and efficacy, although determining suitable concentrations for each material requires careful and thoughtful assessment. Determining the concentration of materials in cosmetic formulations is an intricate process that begins with recipe formulation, considering industry regulations that establish safety and efficacy limits for the materials used. The concentration of materials in cosmetic formulations directly impacts their stability and efficacy. Therefore, suggesting material concentrations for chemical formulations must consider these inherent challenges in determining chemical and cosmetic material concentrations. The field of machine learning is embedded in the context of Artificial Intelligence (AI) and involves the application of computational algorithms to transform empirical data into usable models. These algorithms enable understanding the properties of analyzed datasets, abstracting underlying patterns through a model, predicting unknown values based on the generated model, and detecting observed anomalous behaviors. Its main objective is to develop a model that exhibits high performance not only during training but also when applied to a test set or new data. To date, research utilizing machine learning models to suggest material concentrations in cosmetic formulations is lacking in the literature. Our goal is to identify the optimal methodology for predicting material concentration, so it can be applied as a recommendation in the production of cosmetic formulations. To achieve our goals, we employed four automated learning algorithms : Random Forest Regressor (RFR), Extreme Gradient Boosting (XGBoost), k-Nearest Neighbors (k-NN), and Multi-Layer Perceptron (MLP). We selected performance measures such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) to ensure a robust and reliable evaluation of the proposed models. We extracted and modeled a total of 1679522 records, with which we trained and tested the four machine learning models. The third approach was used to make predictions, as it obtained the best performance indicators among the developed approaches. We also developed an application where the end user can make predictions through a Graphical User Interface (GUI). The results obtained indicate that the RFR model showed the best results among the tested models, with an R^2 value of 0.66892, demonstrating that the model can explain about 66.89% of the data variability. This study represents an important step towards the development of predictive models for the chemical and cosmetic industry, highlighting the importance of applying machine learning techniques and cross-validation in problem-solving in this field.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

MSE	*Mean Squared Error*
RMSE	*Root Mean Squared Error*
MAE	*Mean Absolute Error*
RFR	*Random Forest Regressor*
CSV	*Comma-Separated Values*
MLP	*Multi-Layer Perceptron*
XGBoost	*Extreme Gradient Boosting*
RF	*Random Forest*
k-NN	*k-Nearest Neighbors*
LR	*Logistic Regression*
SVR	*Support Vector Regression*
SVM	*Support Vector Machine*
CART	*Classification and Regression Tree*
bagging	*Bootstrap Aggregating*
GBM	*Gradient Boosting Machine*
DTR	*Decision Tree Regressor*
DT	*Decision Tree*
R^2	Coefficient de détermination
IPF	Indice de Protection Solaire
UVA	*Ultraviolet A*
UV	*Ultraviolet V*
PA	*Protection Against UVA*
LR	*Logistic Regression*
PO	*Political Optimizer*
AARD	*Average Absolute Relative Deviation*
IA	*Intelligence Artificielle*
RNAs	*Réseaux Neuronaux Artificiels*
CNN	*Convolutional Neural Network*
RNN	*Recurrent Neural Network*
IGU	*Interface Graphique Utilisateur*
API	*Interface de Programmation Applicative*
AUC	*Area Under the Receiver Operating Characteristic Curve*
mRNA	*messenger ribonucleic acid*
LASSO	*Least Absolute Shrinkage and Selection Operator*

DÉDICACE

Je dédie ce travail à ma femme bien-aimée, Luciana Medeiros Paungartner, pour son soutien infatigable et son amour indéfectible.

À ma mère, Gunilda Maria Frantz, dont l'amour inconditionnel et la sagesse ont toujours guidé mes pas,

et à mon regretté père, Antonio Sergio Segovia Segovia, dont les exemples de détermination et d'amour sont éternels.

À mon cher beau-fils, Bernardo Paungartner Marcellino, c'est avec gratitude que je dédie également ce travail à vous pour votre présence significative dans mon parcours.

Que ce travail soit un hommage au soutien et à l'amour de ceux qui m'ont accompagné, me façonnant en direction de mes objectifs.

REMERCIEMENTS

Je tiens à remercier tous ceux qui m'ont soutenu pendant la réalisation de ma maîtrise et pendant la rédaction de ce mémoire. Je tiens à remercier également toute l'équipe du Département d'informatique et de mathématique (DIM) de l'UQAC pour tout l'attention dans les moments de doutes au long du projet. Plus particulièrement, j'aimerais remercier mon directeur de recherche M. Kévin Bouchard pour son accompagnement et pour son expertise précieuse offerte tout au long de ma maîtrise.

Je suis également reconnaissant envers le programme MITACS Accelerate et envers M. Lionel Ripoll pour son soutien au fonds de recherche IT29352. Finalement, j'aimerais remercier ma famille et mes amis pour leurs encouragements tout au long de mes études.

CHAPITRE I
INTRODUCTION

1.1 CONTEXTE

L'industrie des produits de beauté occupe une position significative sur la scène mondiale, générant des ventes dépassant 155 milliards de dollars américains à l'échelle mondiale. D'ici 2025, le marché mondial des produits de soins de la peau devrait atteindre environ 189,3 milliards de dollars américains, selon les données collectées de 2012 à 2025 (selon un rapport de la Statista de 2022) [1].

L'industrie cosmétique au Canada est un secteur significatif, avec le Québec jouant un rôle prépondérant. Selon le dernier rapport du Ministère de l'Économie, de l'Innovation et des Exportations du Québec, en 2014, la province représentait environ 32% (soit 623 millions de dollars) de la production totale de produits de beauté dans le pays. En 2021, le marché des cosmétiques au Canada a enregistré un chiffre d'affaires d'environ 1,24 milliard de dollars, avec des projections de croissance annuelle de 1,45%, espérant atteindre la barre des 1,8 milliard de dollars d'ici 2024. Ce marché est hautement compétitif, caractérisé par la domination des principaux acteurs mondiaux. L'Ontario et le Québec se distinguent en tant que principaux producteurs de cosmétiques et de soins de la peau, tout en servant de marchés de consommation les plus actifs pour ces produits dans le pays. [2, 3].

La détermination des concentrations idéales d'ingrédients est un processus important dans le secteur cosmétique pour les formulations chimiques, visant à garantir la qualité, l'efficacité et l'économie liées aux produits développés. Ce besoin ne se limite pas exclusivement au domaine de la cosmétique, mais s'étend à divers autres secteurs où la prévision des concentrations joue un rôle central dans l'optimisation des processus et le développement de produits, comme c'est le cas des concentrations prévues dans le processus de fabrication de

médicaments. L'établissement des concentrations idéales de matériaux dans les formulations chimiques est influencé par l'interaction entre les matériaux et la variation des possibilités de concentrations de produits chimiques. Dans le contexte des formulations cosmétiques, l'interdépendance complexe entre les matériaux représente un défi significatif, nécessitant une attention particulière à la prévision des concentrations de matériaux, afin d'éviter les inefficacités et les réactions indésirables potentielles pouvant être causées par le produit final [4, 5, 6, 7, 8, 9].

L'utilisation de plus en plus fondamentale de la technologie, en particulier des systèmes automatisés, est intrinsèquement liée à sa capacité à simplifier et améliorer les processus quotidiens. À mesure que nous avançons dans le temps, la technologie devient un composant indispensable, apportant avec elle une pertinence croissante [10]. Dans ce contexte, l'apprentissage automatique se distingue comme un outil puissant pour identifier des modèles et effectuer des prévisions sur de grands ensembles de données. Plus précisément, les modèles d'apprentissage automatique émergent comme un choix éminent en raison de leur capacité à traiter la complexité et l'imprévisibilité inhérentes à ces défis. L'efficacité de ces modèles est reconnue pour leur capacité à analyser et interpréter des volumes massifs d'informations, contribuant à la prise de décisions et à des prévisions judicieuses. Ainsi, face à l'évolution constante de l'environnement technologique, l'adoption croissante de l'apprentissage automatique, en particulier à travers des modèles d'apprentissage automatique, répond à la nécessité de faire face à la complexité et à l'ampleur des données actuelles, permettant une compréhension plus approfondie et une meilleure prévision des événements et des modèles [11].

La capacité à prédire avec plus de justesse les concentrations des ingrédients garantit leur sécurité et leur efficacité, bien que la détermination des concentrations appropriées pour chaque matériau nécessite une évaluation soigneuse et pondérée [4, 5, 6, 7], et peut améliorer la conception et le développement accéléré des matériaux [12]. Dans un contexte de plus en plus

axé sur les données et axé sur l'efficacité à l'échelle mondiale, l'apprentissage automatique joue un rôle pertinent, par exemple dans l'application de modèles d'apprentissage automatique pour la justesse dans la prédiction de maladies génétiques multifactorielles, le traitement du langage naturel et les solutions de détection de fraudes financières [13, 14, 15].

Cependant, déterminer les concentrations idéales pour chaque composant peut être un défi en raison des propriétés physico-chimiques des matériaux chimiques qui rendent difficile la justesse de leurs concentrations. Par exemple, les matériaux avec différents états physiques tels que solides, liquides et gaz peuvent être difficiles à mélanger uniformément. Les matériaux avec différentes solubilités peuvent se séparer du mélange, et les matériaux avec différentes densités peuvent se sédimenter ou flotter. De plus, il y a des moments où les formulations chimiques peuvent être complexes, contenant des matériaux variés. Cela peut rendre difficile le contrôle de la concentration de chaque matériau individuel. Enfin, de nombreuses formulations chimiques impliquent des substances aux propriétés dangereuses. Manipuler ces matériaux nécessite une formation et des protocoles de sécurité appropriés [6]. De plus, il existe des défis dus aux interactions complexes entre des composés peu communs et la chimie naturelle des organismes vivants [11]. En résumé, la justesse dans la formulation des produits cosmétiques est importante pour garantir leur sécurité et leur efficacité, bien que la détermination des concentrations adéquates pour chaque matériau nécessite une évaluation minutieuse et pondérée [4, 5, 6, 7].

1.2 MOTIVATION

La détermination de la concentration des matériaux dans les formulations cosmétiques est un processus complexe. Il commence par la formulation de la recette, en tenant compte des réglementations du secteur, qui établissent des limites de sécurité et d'efficacité pour les matériaux utilisés. Cela implique la sélection et la proportion relative de chaque composant dans le

mélange, où l'évaluation de la concentration des matériaux nécessite d'abord l'estimation de la perméabilité des composés à travers la peau [16].

La concentration des matériaux dans les formulations cosmétiques affecte directement leur stabilité et leur efficacité. Les tests de stabilité sont importants pour garantir la sécurité et l'efficacité des produits au fil du temps, comme c'est le cas du test de stabilité accélérée, qui soumet des échantillons d'un produit à des conditions contrôlées de température et d'humidité, plus extrêmes que les conditions habituelles, pendant une période spécifique dans le but d'anticiper les changements physiques, chimiques et microbiologiques, aidant à déterminer la durée de conservation du produit et à garantir sa qualité avant sa mise sur le marché.

La concentration de la formulation, qui représente la quantité de matériau chimique par unité de surface de la peau, est essentielle pour déterminer les performances de l'absorption cutanée [17, 18]. Les variations de la concentration chimique sur la peau affectent l'absorption cutanée, bien que cette relation soit complexe et varie selon la nature spécifique de la formulation chimique [19].

De plus, les formulations cosmétiques intègrent souvent des matériaux avec des concentrations différentes dans chaque composition et ont ces limites de concentration déjà établies pour éviter les effets indésirables sur la santé. La variation des concentrations de composants spécifiques peut entraîner des produits finaux distincts, offrant des propriétés et des efficacités diverses [20]. Par conséquent, suggérer des concentrations de matériaux pour les formulations chimiques doit prendre en compte ces défis qui imprègnent la détermination des concentrations de matériaux chimiques et cosmétiques.

1.3 PROBLÉMATIQUE

Le domaine d'étude de l'apprentissage automatique est inclus dans le contexte de l'IA, et englobe l'application d'algorithmes informatiques pour transformer des données empiriques en

modèles utilisables [21]. Ces algorithmes permettent de comprendre les propriétés d'ensembles de données analysés, d'abstraire les motifs sous-jacents par le biais d'un modèle, de prédire les valeurs inconnues sur la base du modèle généré et de détecter les comportements anormaux observés. Son objectif principal est de développer un modèle qui présente des performances élevées non seulement pendant l'entraînement, mais également lorsqu'il est appliqué à un ensemble de test ou à de nouvelles données [22].

Les algorithmes d'apprentissage automatique sont divisés en quatre catégories principales. L'apprentissage supervisé consiste à entraîner des algorithmes sur des ensembles de données étiquetés, ce qui permet de prédire ou de classer de nouvelles données. Dans l'apprentissage non supervisé, les algorithmes explorent des données non étiquetées à la recherche de modèles ou de relations sous-jacentes. L'apprentissage semi-supervisé combine des éléments des méthodes supervisées et non supervisées. Pendant ce temps, l'apprentissage par renforcement permet aux algorithmes d'apprendre par interaction avec l'environnement, en recevant des récompenses ou des pénalités pour optimiser les décisions futures en fonction des récompenses attendues [23]. De plus, il existe l'apprentissage profond, qui est une sous-branche du domaine de l'apprentissage automatique qui se concentre sur l'entraînement d'algorithmes de réseaux neuronaux artificiels pour apprendre et effectuer des tâches complexes. Il est appelé "profond" en raison de l'utilisation de réseaux neuronaux profonds, qui consistent en de multiples couches de nœuds ou de neurones interconnectés [24]. Le sous-ensemble appelé apprentissage supervisé, inclut de nombreux modèles dont *Decision Tree* (DT), *Random Forest* (RF), XGBoost et k-NN qui figurent parmi les plus connus. Grâce à des techniques statistiques et des algorithmes, ils permettent aux machines d'améliorer leurs performances par l'expérience [24].

Dans ce sens, une étude particulière s'est concentrée sur l'utilisation de modèles d'apprentissage automatique dans le domaine de la science des matériaux, soulignant les similitudes

et les différences entre l'apprentissage automatique et les méthodes traditionnelles de criblage [25]. Dans une autre étude, un modèle in silico a été développé en utilisant RFR pour prédire la pénétration cutanée d'ingrédients actifs dans les produits de protection des plantes [26]. Une autre étude a montré que XGBoost peut être une technique efficace pour prédire les concentrations de composés organiques volatils dans les processus de production chimique avec une haute résolution temporelle [27]. Une étude supplémentaire a utilisé k-NN comme l'un des modèles de base dans une méthode d'ensemble du type *bagging* pour estimer la distribution de concentration d'un soluté médical [28]. De plus, une étude a évalué l'efficacité des approches d'apprentissage automatique, le MLP étant utilisé dans l'industrie pharmaceutique pour prédire la concentration de matériaux chimiques dans le développement de médicaments [29]. Des modèles basés sur l'apprentissage automatique, tels que le MLP, ont été développés pour prédire la stabilité chimique au fil du temps, réduisant ainsi les ressources et le temps nécessaires aux études de stabilité [29]. Ces recherches soulignent le potentiel de l'apprentissage automatique dans le contexte de la détermination de la concentration de matériaux dans divers types de produits chimiques. Dans ce scénario, à ce jour, nous ne connaissons aucune recherche dans la littérature qui ait utilisé des modèles d'apprentissage automatique pour suggérer la concentration de matériaux dans les formulations cosmétiques. Il s'agit donc d'une excellente opportunité de recherche que nous avons choisi de poursuivre dans ce mémoire de maîtrise.

1.4 OBJECTIF

Notre objectif est de prédire la concentration des matériaux dans les formulations cosmétiques en utilisant l'apprentissage automatique. Pour ce faire, nous utiliserons quatre modèles d'apprentissage supervisé : RFR, XGBoost, k-NN et MLP. Pour guider notre approche, nous avons formulé l'hypothèse suivante :

Hypothèse : *Les modèles d'apprentissage automatique permettent de prédire avec précision la concentration des matériaux pour aider à la formulation de produits cosmétiques.*

Dans ce contexte, à travers notre objectif, nous avons élaboré une question de recherche (QR) :

QR : Est-il possible de prédire avec précision la concentration des matériaux dans la formulation des produits cosmétiques à l'aide de modèles d'apprentissage automatique ? Notre objectif en abordant cette question de recherche est d'identifier la méthodologie optimale pour prédire la concentration des matériaux, de manière à ce qu'elle puisse être appliquée en tant que recommandation dans la production de formulations cosmétiques. Cette question vise à déterminer la meilleure approche pour faire ces prédictions en se basant sur des mesures de performance.

1.5 SOLUTION PROPOSÉE

Pour évaluer la justesse de ces prévisions, nous avons utilisé des mesures de performance couramment utilisées pour évaluer les performances des modèles de régression dans des tâches de prévision, complétées par une analyse des graphiques de résidus [30, 31, 32, 33, 34, 35, 36, 37, 38].

Pour atteindre nos objectifs, nous avons utilisé quatre algorithmes d'apprentissage automatique et sélectionné des mesures de performance, telles que MSE, RMSE, MAE et R^2. Pour garantir une évaluation robuste et fiable des modèles proposés, nous avons opté pour l'application de la méthode de validation croisée *k-fold* [39]. De plus, nous avons utilisé différents modèles de graphiques de résidus, tels que des graphiques de dispersion des résidus par rapport aux prévisions, des histogrammes des résidus et des graphiques QQ (quantile-quantile) des résidus [30, 31, 32, 33, 34, 35, 36, 37, 38]. Les ensembles de données utilisés

pour les étapes d'entraînement et de test ont été obtenus en collaboration avec le Dr Lionel Ripoll, professeur au département des sciences fondamentales de l'Université du Québec à Chicoutimi. Tous les éléments de validation des mesures présentés dans cette section seront discutés en détail ultérieurement dans la section 2.1.6 du chapitre 2.

Notre méthode implique l'extraction d'informations de plusieurs feuilles de calcul au format Excel, suivie d'une adaptation de ces données selon nos besoins à l'aide d'un algorithme que nous avons développé spécifiquement à cet effet. Ensuite, nous avons exécuté les algorithmes d'apprentissage automatique et évalué les résultats à l'aide des indicateurs susmentionnées.

Après avoir obtenu des résultats positifs dans cette évaluation, nos algorithmes sont prêts à être utilisés pour suggérer les concentrations des matériaux pour les formulations cosmétiques en étude.

1.6 RÉSULTATS ET CONTRIBUTIONS

Le chapitre 4 (Validation) discute des résultats de notre solution. Comme les lecteurs pourrons le constater, les mesures MSE, RMSE et MAE tendent à converger vers 0, ce qui indique une faible disparité entre les prédictions et les valeurs réelles, tandis que la métrique R^2 tend à converger vers 1, indiquant que le modèle explique la variation globale de la variable dépendante. De plus, dans le graphique de dispersion des résidus par rapport aux prédictions, nous remarquons que les résidus sont répartis de manière aléatoire autour de 0 sur l'axe horizontal. L'histogramme des résidus présente une distribution suivant une courbe normale, tandis que dans le graphique QQ des résidus, il suggère une concordance évidente entre les résidus et la distribution théorique au point central du graphique, et à mesure qu'il se rapproche des extrémités, nous observons un écart de ces résidus par rapport à ladite distribution, suggérant que les données présentent des valeurs plus extrêmes que prévu si elles

provenaient réellement d'une distribution normale. La validation des modèles d'apprentissage automatique a été réalisée en utilisant la méthode de validation croisée *k-fold*, garantissant une analyse robuste et fiable des performances des modèles sur différents ensembles de données.

Nos algorithmes sont disponibles pour le développement et l'utilisation, et peuvent être téléchargés en ligne [1].

1.7 ORGANISATION

Dans le chapitre 2 (Revue de la littérature), nous présentons les bases théoriques essentielles pour une meilleure compréhension de notre proposition. Nous abordons les concepts liés aux modèles d'apprentissage automatique, ainsi que les défis associés à leur mise en œuvre. Nous détaillons les mesures utilisées pour évaluer les prévisions faites par les modèles. Enfin, nous présentons quelques études liées à la prédiction de la concentration de matériaux en utilisant des modèles d'apprentissage supervisé 2.1.3 (Apprentissage supervisé).

Dans le chapitre 3 (Implémentation de l'approche), nous expliquons en détail l'implémentation de la solution proposée, fournissant des informations sur le prétraitement des données, le traitement des données, le développement de l'algorithme d'extraction de données, l'entraînement des modèles d'apprentissage automatique qui nous ont permis d'obtenir des prédictions de concentration des matériaux, et enfin le développement d'une IGU.

Dans le chapitre 4 (Validation), nous décrivons l'étude de cas, en présentant étape par étape la préparation et la collecte des informations, ainsi que en détaillant les résultats de notre question de recherche. Ensuite, nous répondons à la question de recherche formulée au début du mémoire et nous présentons quelques observations sur les difficultés rencontrées pendant le développement de la recherche.

1. https://github.com/Cosmetic-Concentration-Predictor

Enfin, nos contributions et perspectives pour les travaux futurs sont discutées dans le chapitre 5 (Conclusion).

CHAPITRE II
REVUE DE LA LITTÉRATURE

Le thème principal de ce mémoire est la prédiction de la concentration des matériaux dans une formulation cosmétique à l'aide de l'apprentissage automatique. Dans le présent chapitre, nous présentons les fondements essentiels qui sous-tendent la compréhension complète de cette étude.

Dans la section 2.1, nous expliquons ce qu'est une formulation cosmétique. De plus, nous décrirons cinq algorithmes de prédiction qui ont été utilisés afin d'atteindre les objectifs fixés pour cette recherche. Dans ce contexte, nous conduirons une analyse détaillée des indicateurs de performance adoptés pour évaluer l'efficacité et l'efficience des approches de prédiction entreprises.

Enfin, la section 2.2 a pour objectif de fonder et de contextualiser la présente recherche. Nous menons une revue de la littérature. Cette revue, de caractère complet et détaillé, englobe les études les plus pertinentes qui jettent la lumière sur les aspects fondamentaux et les nuances critiques associées aux sujets abordés dans cette enquête, afin de placer notre recherche dans le contexte académique et scientifique actuel.

2.1 DÉFINITIONS

2.1.1 FORMULATION COSMÉTIQUE

Une formulation chimique est un mélange de substances naturelles ou artificielles. Les formulations cosmétiques contiennent typiquement une combinaison d'ingrédients spécifiques, comme des émollients pour l'hydratation, des agents émulsifiants pour la fusion des ingrédients, des conservateurs pour prévenir la prolifération des micro-organismes, des épaississants pour

améliorer la texture et la viscosité, des parfums pour offrir un arôme agréable, et des actifs qui remplissent des fonctions spécifiques en fonction des besoins du produit [5, 40, 41].

La production de produits cosmétiques implique la fusion d'ingrédients actifs avec d'autres composés afin de déterminer leur forme physique et de réguler la libération de ces ingrédients actifs [5]. Cette élaboration aboutit au développement d'un produit cosmétique final, destiné à être appliqué sur la peau, les cheveux, les ongles, les dents, entre autres zones du corps, dans le but de nettoyer, parfumer, modifier l'apparence, protéger ou corriger les odeurs corporelles. L'objectif est d'améliorer l'esthétique corporelle, ces produits étant disponibles sous une variété de formes, notamment des solutions, des crèmes, des lotions, des suspensions, etc. [42].

La sécurité des produits cosmétiques et leur composition sont des aspects importants pour les consommateurs. L'analyse des cosmétiques peut être un défi en raison du grand nombre d'ingrédients présents. Par conséquent, des techniques analytiques sont utilisées pour évaluer la composition et la qualité des produits cosmétiques [42, 43]. Les techniques analytiques couramment utilisées pour évaluer la composition chimique des produits cosmétiques comprennent des méthodes de détection d'ingrédients restreints, des techniques d'évaluation sensorielle et des méthodes de préparation d'échantillons pour les métaux lourds et toxiques [43, 44, 45].

Dans le tableau 2.1, nous pouvons observer un exemple de formulation cosmétique. La première colonne est destinée au code de la matière première, la deuxième colonne se rapporte à la matière première et la troisième colonne indique les concentrations de chaque matériau de la formulation cosmétique.

TABLEAU 2.1 : Composition de la formulation d'exemple
©Christian Gonzalo Frantz Segovia, 2024

Code	Matières Premières	Concentration (%)
EAU010	EAU OSMOSEE TIEDE 30°C	63.1
HUME002	GLYCERINE CODEX VEGETALE 99,5 %	2
ADDI032	SULFATE MAGNESIUM HEPTAHYDRATE	0.6
ADDI014	SEL FIN EPURE SECHE DESULFATE	0.6
EMUL035	ABIL EM 90	1.5
CGRA048	BEURRE DE KARITE fondu	2.5
EMUL044	PLUROL DIISOSTEARIQUE	2.5
CGRA581	CETIOL CC	4
SILI007	HUILE SILICONE VOLATILE DC 345 FLUIDE	6
HUIL040	CETIOL OE	6
CONS037	DEKABEN CP (= ELESTAB CPN)	0.3
CONS023	BIOSOL	0.1
HUME016	HYDROLITE 5	2
ACTI740	NIGHT RESTORE GINKGO + STANDSTILL ROSE DE DAMAS	0.2
ACTI345	NOCTILISS	0.5
ACTI201	MG RELAX	0.4
ADDI049	ORGASOL CARESSE	0.5
EAU010	EAU OSMOSEE TIEDE 30°C	7
ACTI703	BASHYAL POUDRE	0.1
FRAG007	PARFUM BASE EAU D'ECLAT GAMME VRAI	0.1
	TOTAL	100

2.1.2 APPRENTISSAGE AUTOMATIQUE

L'apprentissage automatique est un terme large faisant référence à des algorithmes capables de faire des prédictions intelligentes basées sur de vastes ensembles de données. Il existe quatre stratégies générales dans le domaine de l'IA qui peuvent être mises en œuvre, et celles-ci sont catégorisées comme les approches de : l'apprentissage non supervisé, l'apprentissage supervisé et l'apprentissage par renforcement [46].

L'évolution de l'apprentissage automatique semble atteindre des niveaux de compréhension sémantique comparables à ceux des humains, démontrant, dans certaines situations, une

capacité supérieure à détecter des motifs abstraits. Actuellement, l'application de l'apprentissage automatique s'étend à divers domaines au-delà de la technologie, englobant des secteurs tels que la santé, l'éducation, le marché financier et le divertissement [47].

2.1.3 APPRENTISSAGE SUPERVISÉ

Le diagramme de l'apprentissage automatique supervisé est représenté par la figure 2.1. L'apprentissage automatique supervisé est un paradigme dans lequel des algorithmes sont entraînés sur des données étiquetées pour apprendre la relation entre les entrées et les sorties souhaitées. Il implique la préparation de jeux de données d'entraînement avec des exemples d'entrées et des étiquettes, l'entraînement d'un modèle pour optimiser cette relation, l'évaluation du modèle sur des données de test et l'utilisation ultérieure du modèle pour prédire les étiquettes sur de nouvelles données non étiquetées. Il est appliqué à des tâches telles que la classification et la régression dans divers domaines, tels que le diagnostic médical et la prévision financière [46, 48]. C'est la modalité d'apprentissage automatique que nous abordons dans ce mémoire, car nos données consistent en des échantillons avec des étiquettes connues et notre objectif est de réaliser des prévisions sur la concentration des matériaux dans des échantillons non étiquetés.

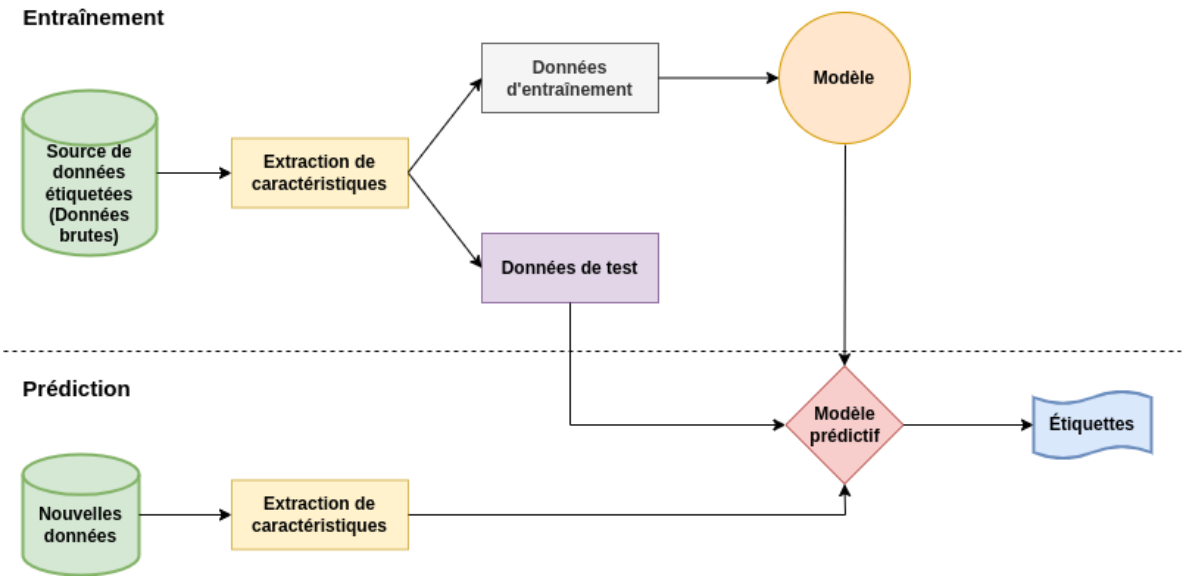


FIGURE 2.1 : Diagramme de flux d'apprentissage supervisé
©Christian Gonzalo Frantz Segovia, 2024

DECISION TREE REGRESSOR

Le *Decision Tree Regressor* (DTR) est un modèle d'apprentissage automatique utilisé pour les problèmes de régression, où l'objectif est de prédire des valeurs continues. Cet algorithme utilise des arbres de décision pour créer une structure hiérarchique qui divise l'ensemble de données en fonction des caractéristiques, étant particulièrement efficace dans les cas où la relation entre les variables d'entrée et la variable de sortie est non linéaire.

Depuis son introduction, l'algorithme DTR a joué un rôle important dans les applications de modélisation prédictive, en particulier dans les problèmes de régression, où l'objectif est de prédire des valeurs continues [49]. Initialement, les arbres de décision étaient principalement explorés par des scientifiques et des experts en apprentissage automatique intéressés par l'innovation. Au fil des ans, le DTR a évolué pour devenir un outil largement reconnu et adopté non seulement par les communautés d'apprentissage automatique, mais aussi dans

divers domaines en dehors de l'informatique [50]. Son application dans des domaines aussi divers que les finances, la santé et les sciences physiques met en évidence sa polyvalence et son adaptabilité.

Un jalon significatif dans la validation du DTR a été observé dans des études de référence, dans lesquelles ses performances ont été comparées à celles d'autres techniques de régression. Ces études, basées sur une variété d'ensembles de données réels, soulignent l'efficacité du DTR pour fournir des prévisions précises dans une large gamme de scénarios [49, 50].

Le DTR est connu pour son interprétabilité, ce qui facilite la compréhension des décisions prises par le modèle. Sa structure arborescente permet d'identifier des modèles et des relations dans des données complexes, ce qui est nécessaire en modélisation prédictive, offrant des performances robustes et une interprétabilité dans une variété de contextes et de domaines d'application [49, 50, 51, 52].

RANDOM FOREST REGRESSOR

Depuis son introduction en 2001, l'algorithme RFR a connu une croissance notable en popularité, tant dans les applications de régression que de classification. Représentant un modèle d'apprentissage automatique appartenant à la famille des méthodes d'ensemble, plus précisément à la catégorie du *bagging*. Cette méthode a évolué pour devenir une approche de classification de référence qui rivalise favorablement avec la *Logistic Regression* (LR) dans divers domaines scientifiques propices à l'innovation.

Dans les premières années suivant son introduction, l'utilisation de l'algorithme RFR était principalement restreinte aux scientifiques et aux experts en apprentissage automatique, intéressés par l'innovation. Cependant, on observe actuellement que le RFR est de plus en

plus reconnu et adopté dans diverses communautés, non seulement informatiques, mais aussi dans des domaines non liés à l'informatique.

En 2018, une étude de référence remarquable a été menée par des chercheurs, visant à évaluer les performances prédictives de la version originale de l'algorithme RFR avec des paramètres par défaut, en comparaison avec la LR, dans le contexte de la classification binaire. Cette étude a porté sur une vaste collection de 243 ensembles de données réels, représentant une large gamme de scénarios d'application. Les résultats ont révélé qu'une proportion impressionnante de 69% de ces ensembles de données a montré que les performances de l'algorithme RFR ont dépassé de manière significative l'alternative représentée par la LR. Ces résultats soulignent la capacité prédictive distincte du RFR et sa pertinence dans les applications de classification binaire [53].

EXTREME GRADIENT BOOSTING

Cet algorithme représente un modèle d'apprentissage automatique appartenant à la famille des méthodes d'ensemble, plus précisément à la catégorie du *boosting*, largement utilisé dans les tâches de classification et de régression, où la justesse est essentielle. Tout comme l'algorithme RFR, le XGBoost est appliqué dans divers domaines qui exigent une justesse élevée, en particulier en raison de sa capacité à gérer des ensembles de données volumineux et complexes.

L'algorithme XGBoost est fondé sur le concept de *boosting*, étant une évolution de cette méthode. Il incorpore deux types de régularisation pour éviter le surapprentissage et construit des arbres de décision qui corrigent les erreurs des arbres précédents. De plus, il utilise une fonction de perte personnalisée qui tient compte de l'erreur résiduelle des modèles précédents, encourageant l'apprentissage des arbres suivants. La note de prévision est calculée

en additionnant les prévisions de tous les arbres, et de nouveaux arbres sont entraînés à chaque itération, s'adaptant aux erreurs commises par les modèles précédents. Il met également en œuvre des mécanismes pour prévenir le surapprentissage [54].

Dans le cadre de notre recherche, nous utilisons l'algorithme XGBoost comme partie intégrante de la stratégie de modélisation prédictive de la concentration de matériaux chimiques dans les formulations cosmétiques. Des études antérieures ont souligné les performances remarquables de l'algorithme XGBoost dans les applications d'apprentissage automatique, consolidant sa position en tant qu'outil précieux pour la présente recherche. Il a été appliqué dans divers domaines, tels que le marché boursier, l'agriculture et la prévision des changements climatiques [55, 56, 57].

K-NEAREST NEIGHBORS

L'un des méthodes les plus courantes de classification dans l'apprentissage automatique est le classificateur k-NN, qui est une procédure d'apprentissage supervisé. Le k-NN est un algorithme non paramétrique, c'est-à-dire qu'il ne fait pas d'hypothèses sur la structure des données sous-jacentes. Ce type de classificateur regroupe des éléments en fonction de leur proximité avec leurs « k »,voisins les plus proches. Il prend en compte uniquement l'environnement immédiat de l'objet, pas la distribution complète des données. Pour que cette approche fonctionne, il est nécessaire de spécifier un nombre entier appelé « k ». Ce paramètre sert à spécifié combien de points de références seront utilisés en combinaison avec une mesure de proximité pour déterminer la classe ou la valeur [58].

Dans une étude réalisée avec le *SEER Breast Cancer Dataset*, qui contenait 4.024 enregistrements, on a cherché à détecter et prédire le cancer du sein. L'objectif était de développer et d'anticiper une méthode utile, fournissant des informations précieuses aux

patients et suggérant une approche fiable pour prédire le cancer du sein. Le modèle k-NN a été utilisé avec d'autres modèles, et dans cette comparaison, il a présenté la plus faible justesse, qui était de 88,08%. Après l'application de *Least Absolute Shrinkage and Selection Operator* (LASSO), les résultats du modèle k-NN se sont légèrement améliorés, mais sont restés inférieurs aux résultats obtenus par les autres modèles [58].

2.1.4 APPRENTISSAGE PROFOND

L'apprentissage profond fait partie d'un ensemble plus large d'approches d'apprentissage automatique qui s'appuient sur les *Réseaux Neuronaux Artificiels* (RNAs) pour apprendre des représentations. Ces réseaux ont la capacité d'apprendre des représentations complexes des données par le biais de multiples couches, permettant l'extraction de caractéristiques à des niveaux de plus en plus abstraits. Cette technique emploie une architecture informatique composée de plusieurs couches de traitement, y compris des couches d'entrée, cachées et de sortie, pour absorber des informations des données. Au lieu de dépendre de caractéristiques présélectionnées, comme dans de nombreuses méthodes traditionnelles d'apprentissage automatique, l'apprentissage profond extrait automatiquement les caractéristiques les plus pertinentes des données pendant l'entraînement. L'apprentissage automatique profond est souvent utilisé pour des tâches telles que la reconnaissance de modèles, la classification, la prévision, le traitement du langage naturel, la vision par ordinateur et de nombreuses autres applications où les données sont complexes ou n'ont pas une représentation facile à exprimer en termes de règles mathématiques ou de caractéristiques spécifiques [24].

MULTI-LAYER PERCEPTRON

Le MLP est une architecture fondamentale en informatique neuronale qui joue un rôle essentiel dans l'avancement du domaine. Il s'agit d'un réseau composé de plusieurs couches

interconnectées de neurones artificiels, capables de traiter des informations complexes et de les transmettre entre les couches. Ce type de réseau, connu sous le nom de Perceptron Multicouche, est précieux dans l'apprentissage automatique pour des tâches telles que la classification et la régression, permettant l'apprentissage et la représentation de relations complexes dans les données. Cette architecture est souvent utilisée dans l'apprentissage supervisé, en traitant les données d'entrée, en passant par les couches cachées, qui effectuent des calculs mathématiques complexes, et en générant des résultats dans la couche de sortie. Son mécanisme repose sur la rétropropagation (*backpropagation*), qui ajuste les poids des connexions entre neurones de manière itérative, minimisant la différence entre les prévisions et les valeurs réelles des données d'entraînement [24].

Le MLP est appliqué dans divers domaines, tels que la reconnaissance des formes, le traitement du langage naturel, la vision par ordinateur, l'analyse de séries temporelles, entre autres. Sa capacité d'apprendre et de généraliser à partir des données en fait un outil polyvalent dans les problèmes de prédiction et de classification. De plus, le MLP est partie intégrante de l'apprentissage profond, et malgré l'évolution de techniques et d'architectures plus complexes, telles que les *Convolutional Neural Network*s (CNNs) et les *Recurrent Neural Network*s (RNNs), il reste essentiel dans le domaine de l'apprentissage automatique, représentant un domaine de recherche prometteur en informatique neuromorphique [24]. Bayat et al. [59] ont démontré le fonctionnement efficace d'un classificateur perceptron à couche cachée utilisant un matériel intégré à signaux mixtes, composé de matrices passives de barre transversale *memristive* en oxyde métallique. Cela représente une avancée significative dans l'informatique neuronale, en tirant parti de la technologie *Memristor* améliorée et en atteignant des résultats de classification proches de ceux obtenus en simulations. D'autre part, Chai et al. [60] ont développé un modèle de MLP pour classifier l'hypertension chez les adolescents sur la base de données anthropométriques et sociodémographiques. Le modèle a obtenu des résultats prometteurs, avec une sensibilité, une spécificité et d'autres mesures de performance

considérables. Cette approche peut être utile pour le dépistage des adolescents à haut risque de développer l'hypertension.

2.1.5 ESTIMATION DE PRÉCISION AVEC LA VALIDATION CROISÉE K-FOLD

La méthode de validation croisée *k-fold* est une approche de validation croisée largement utilisée pour évaluer les modèles d'apprentissage automatique. Considérée comme une méthodologie pour estimer la capacité de généralisation d'un modèle sur différents ensembles de données. Dans cette méthode, l'ensemble de données est divisé en k parties, connues sous le nom de *folds*, de taille approximativement égale. Le modèle est ensuite entraîné k fois, en utilisant $k-1$ *folds* comme ensemble d'entraînement à chaque itération, et le *fold* restant comme ensemble de test. Cette approche donne lieu à k ensembles distincts de mesures de performance. À la fin des k itérations, des mesures de performance telles que MSE, RMSE, MAE, R^2 sont calculées pour chaque itération. La performance globale du modèle est ensuite estimée en calculant la moyenne et l'écart type de ces mesures sur les k itérations. Ainsi, la validation croisée *k-fold* estime la capacité de généralisation d'un modèle sur différents ensembles de données, étant un outil utilisé pour l'évaluation et la comparaison des modèles d'apprentissage automatique [39].

2.1.6 INDICATEURS DE PERFORMANCE

Dans notre étude, nous avons utilisé les mesures de performance (mesures d'erreur) MSE(2.1), RMSE(2.2), MAE(2.3) et R^2(2.4), car elles fournissent une vue complète des performances des modèles. Ces mesures peuvent être définies comme des constructions logiques et mathématiques. Elles sont des outils essentiels pour évaluer l'accord entre les prédictions générées par le modèle entraîné et les données réelles (observées) de l'ensemble de test [30, 61, 62]. De plus, nous avons utilisé des graphiques de résidus, des graphiques de disper-

sion des résidus en fonction des prédictions, des histogrammes des résidus et des graphiques QQ (quantile-quantile) des résidus, car ils aident à détecter les motifs ou l'hétéroscédasticité dans les erreurs [33, 34, 37].

MEAN SQUARED ERROR

La MSE(2.1) est une métrique largement utilisée pour évaluer la justesse des modèles par rapport aux valeurs réelles, en particulier dans les problèmes de régression, où l'objectif est de prédire des valeurs continues. Cette métrique calcule la moyenne des carrés des écarts entre les prédictions du modèle et les valeurs réelles observées. Un MSE de magnitude réduite est associé à une performance améliorée du modèle, se rapprochant de la valeur zéro. L'état idéal est représenté par un MSE égal à zéro, indiquant une correspondance exacte entre les prédictions du modèle et les valeurs réelles. En raison de sa sensibilité remarquable, la MSE pénalise fortement les grands écarts, ce qui est approprié lorsque des écarts significatifs doivent être mis en évidence. La fonction de la MSE est dérivable, ce qui est avantageux dans les algorithmes d'optimisation, tels que la descente du gradient. L'unité de la MSE est le carré de l'unité de la variable dépendante, ce qui fournit une interprétation claire et significative de la métrique [30].

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.1}$$

— n est le nombre total d'observations dans les données ;

— y_i représente la valeur réelle à l'observation i ;

— $\hat{y}_i$ représente la valeur prédite par le modèle pour l'observation i ;

Différence entre valeur réelle et prévue : $y_i - \hat{y}_i$ mesure la différence entre la valeur réelle et la valeur prévue pour chaque observation. En élevant cette différence au carré $(y_i - \hat{y}_i)^2$, nous pénalisons davantage les grands écarts, ce qui amplifie l'influence des écarts significatifs. Nous supprimons également le signe négatif comme une fonction absolue, ce qui est pratique dans le calcul.

E.g. $-2+2=0$ cependant $-2^2+2^2=8$

Somme des différences au carré : $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est la somme des différences au carré pour toutes les observations. Cette somme représente la « somme totale des erreurs quadratiques » entre les valeurs réelles et prévues.

Moyenne des erreurs quadratiques : $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est la moyenne des erreurs quadratiques. Cette moyenne fournit une mesure agrégée de la qualité de la prévision, ajustée au nombre d'observations.

ROOT MEAN SQUARED ERROR

Le RMSE(2.2) est une métrique statistique utilisée pour évaluer la justesse des modèles en problèmes de régression, tout comme le MSE. La distinction fondamentale entre elles réside dans l'approche des écarts entre les prévisions du modèle et les valeurs réelles.

Le RMSE, étant la racine carrée du MSE, fournit une mesure de la magnitude moyenne des erreurs entre les prévisions du modèle et les valeurs réelles. La recherche du RMSE par la proximité de la valeur zéro reflète un rendement idéal. Plus le RMSE est proche de zéro, plus la justesse du modèle dans la prédiction des données est élevée. Le RMSE est particulièrement sensible aux grands écarts, en raison de la racine carrée dans sa formule, contrairement à le MSE, qui offre une moyenne plus directe des carrés des écarts. Cette sensibilité accrue du

RMSE aux écarts significatifs contribue à une évaluation plus rigoureuse de la performance du modèle en termes de justesse prédictive [30, 31].

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{2.2}$$

Le RMSE est une version du MSE qui calcule la racine carrée du résultat du MSE. La racine carrée a pour effet de pénaliser plus fortement les grands écarts, ce qui rend le RMSE plus sensible aux écarts significatifs. L'unité du RMSE est la même que l'unité de la variable dépendante, ce qui facilite l'interprétation pratique.

MEAN ABSOLUTE ERROR

Le MAE est une métrique statistique utilisée pour évaluer la justesse des modèles en problèmes de régression. Cette métrique mesure la moyenne des magnitudes absolues des écarts entre les prévisions du modèle et les valeurs réelles observées. Le MAE offre une interprétation directe et intuitive, représentant la moyenne absolue des écarts entre les prévisions du modèle et les valeurs réelles. Comme elle n'implique pas le carré des erreurs, le MAE n'amplifie pas l'influence des grands écarts, ce qui la rend une métrique robuste aux valeurs extrêmes.

Le choix du MAE est approprié dans des situations où toutes les erreurs, quelle que soit leur magnitude, ont la même importance. En comparaison du MSE et du RMSE, le MAE fournit une approche plus linéaire pour évaluer la justesse du modèle en termes d'erreur moyenne absolue [30, 31].

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.3}$$

— n est le nombre total d'observations ou échantillons ;

— y_i représente la valeur réelle à l'observation i ;

— $\hat{y}_i$ représente la valeur prévue pour l'observation i ;

— $|y_i - \hat{y}_i|$ représente la valeur absolue de la soustraction.

COEFFICIENT DE DÉTERMINATION

Le R^2(2.4), souvent noté R^2, est une mesure statistique qui fournit une indication de la qualité de l'ajustement d'un modèle de régression aux données. Le R^2 évalue la proportion de la variabilité de la variable dépendante (celle qui est prédite par le modèle) qui est expliquée par les variables indépendantes (ou prédictives) dans le modèle de régression. Le R^2 est une métrique largement utilisée en régression pour évaluer la pertinence du modèle. Cependant, il doit être utilisé en combinaison avec d'autres mesures et considérations, car il peut avoir des limitations, en particulier dans les modèles plus complexes. En résumé, le R^2 est une mesure précieuse pour évaluer à quel point un modèle de régression s'ajuste aux données et explique la variabilité de la variable dépendante [32].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.4}$$

— n est le nombre total d'observations ou échantillons ;

— y_i représente la valeur réelle à l'observation i ;

— $\hat{y}_i$ représente la valeur prévue pour l'observation i ;

— $\bar{y}$ est la moyenne de la variable dépendante.

Le R^2 varie de 0 à 1. Une valeur de 1 indique que le modèle explique parfaitement la variabilité de la variable dépendante, tandis qu'une valeur de 0 indique que le modèle n'offre aucune explication.

Variance Résiduelle : $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ est la variabilité qui n'est pas expliquée par le modèle. Elle est la somme des carrés des différences entre les prévisions du modèle et les valeurs réelles.

Variance Totale : $\sum_{i=1}^{n}(y_i - \bar{y})^2$ représente la variabilité totale de la variable dépendante. Elle peut être pensée comme la différence entre chaque observation et la moyenne de la variable dépendante élevée au carré.

L'expression $1 - \frac{\textbf{Variance Résiduelle}}{\textbf{Variance Totale}}$ est une manière d'exprimer la proportion de variabilité expliquée par le modèle.

GRAPHIQUES DE RÉSIDUS

Les graphiques de résidus sont utilisés pour évaluer la qualité d'un modèle statistique de régressions linéaires et non linéaires. Les données linéaires font référence à une relation entre deux variables qui peut être représentée par une ligne droite sur un graphique, tandis que les données non linéaires font référence aux relations qui ne peuvent pas être modélisées de manière appropriée par une ligne droite. Dans ce cas, les graphiques de résidus représentent la différence entre les valeurs observées et les valeurs prédites par le modèle, appelées résidus. Les graphiques de résidus aident à diagnostiquer d'éventuels problèmes avec le modèle, tels

que l'hétéroscédasticité, la non-linéarité, la présence des valeurs extrêmes, etc. Ils sont des outils visuels qui aident à évaluer l'adéquation du modèle statistique aux données observées. Les résidus sont les différences entre les valeurs observées et les valeurs prédites par le modèle, représentées par $e_i = y_i - \hat{y}_i$, où ei est le résidu pour l'observation i, y_i est la valeur réelle et $\hat{y}_i$ est la valeur prédite. L'axe horizontal représente les observations ou les valeurs prédites et l'axe vertical représente les résidus. Pour l'interprétation, il est nécessaire d'observer si les résidus sont distribués de manière aléatoire autour de zéro. Cela suggère que le modèle capture efficacement la variabilité des données. Des modèles systématiques tels qu'une tendance ou une structure peuvent indiquer des inadéquations du modèle. Ces graphiques sont souvent utilisés après la construction d'un modèle statistique, pour vérifier si les hypothèses du modèle ont été respectées [33, 34].

GRAPHIQUES DE DISPERSION DES RÉSIDUS EN FONCTION DES PRÉVISIONS

Les graphiques de dispersion des résidus en fonction des prévisions constituent une extension précieuse des graphiques de résidus dans l'analyse des modèles de régression, offrant une vision plus détaillée de la qualité de l'ajustement et aidant à identifier des domaines spécifiques qui peuvent nécessiter des améliorations du modèle statistique. Cette extension ajoute une dimension supplémentaire à l'évaluation de l'ajustement du modèle, présentant une représentation visuelle de la relation entre les résidus et les prévisions du modèle. Ces graphiques offrent une perspective complète de la variabilité des résidus en fonction de différents intervalles de valeurs prévues. L'interprétation de ces graphiques consiste à rechercher tout motif ou tendance discernable dans les points. Une dispersion uniforme des résidus le long des prévisions suggère une distribution aléatoire des erreurs, indiquant un bon ajustement du modèle. Les motifs non aléatoires, tels que les courbes ou d'autres structures, peuvent suggérer des inadéquations du modèle, nécessitant un examen approfondi [33, 34].

HISTOGRAMMES DES RÉSIDUS

L'histogramme des résidus est un outil graphique utilisé dans l'analyse des modèles statistiques, en particulier dans les contextes de régression, pour examiner la distribution des résidus, qui sont les différences entre les valeurs observées et les valeurs prévues par le modèle. Ce type de graphique vise à fournir une visualisation de la dispersion et des motifs dans les résidus, en aidant à identifier de possibles violations des hypothèses du modèle. L'histogramme permet d'évaluer si les résidus suivent une distribution normale. L'observation de motifs aide à identifier des motifs ou des structures dans les résidus, tels que des asymétries, des valeurs extrêmes et d'autres écarts de l'aléa, qui peuvent indiquer des inadéquations du modèle. En résumé, c'est un outil important et complémentaire pour la validation continue des modèles statistiques, fournissant des informations sur la normalité et les motifs dans les résidus, essentiels pour garantir la fiabilité et la justesse du modèle. L'axe horizontal de l'histogramme représente les valeurs des résidus, tandis que l'axe vertical indique la fréquence à laquelle ces valeurs se produisent. Les barres du graphique représentent la distribution des résidus. Dans le contexte des données linéaires, un histogramme symétrique et en forme de cloche suggère que les résidus sont distribués normalement, ce qui est une hypothèse importante pour la validité statistique du modèle linéaire. Dans les données non linéaires, l'histogramme est encore utile pour identifier des motifs systématiques dans les résidus. Il peut indiquer si la non-linéarité a été complètement capturée par le modèle [35, 36].

GRAPHIQUES QQ (QUANTILE-QUANTILE) DES RÉSIDUS

Les graphiques QQ (Quantile-Quantile) des résidus sont des outils graphiques utilisés en statistique pour évaluer si un échantillon de données suit ou non une distribution théorique, telle que la distribution normale. Dans le contexte des modèles de régression, ils aident à

vérifier si les résidus du modèle suivent une distribution théorique et sont utilisés pour évaluer visuellement si les résidus du modèle linéaire ou non linéaire sont distribués de manière normale. Ils permettent de comparer la distribution empirique des résidus avec la distribution théorique assumée (généralement la normale) pour identifier les écarts [37, 38].

L'axe horizontal représente les quantiles théoriques (attendus sous une distribution spécifique, telle que la normale). L'axe vertical représente les quantiles observés des résidus. Si les points suivent approximativement une ligne diagonale, les résidus sont en accord avec la distribution théorique. Pour les données linéaires, un graphique QQ est considéré bon si les points s'approchent d'une ligne diagonale, indiquant que les résidus suivent une distribution normale. Même dans les modèles non linéaires, les graphiques QQ sont utiles pour vérifier si les résidus suivent une distribution théorique, garantissant que les hypothèses du modèle sont respectées [37, 38].

2.2 TRAVAUX DANS LA LITTÉRATURE

Il existe un consensus dans la littérature concernant l'importance de l'utilisation de modèles d'apprentissage automatique pour résoudre ou simplifier les tâches quotidiennes [57, 58, 63, 64, 65].

Shim et al. [66] ont mené une étude utilisant l'apprentissage automatique pour prédire l'Indice de Protection Solaire (IPF) et le niveau de protection contre les rayons ultraviolets (*Ultraviolet A* (UVA)) et le *Protection Against UVA* (PA) en utilisant des informations sur les ingrédients des écrans solaires. Environ 2200 résultats cliniques individuels ont été analysés par DTR, permettant d'obtenir des prédictions précises pour l'IPF et le PA en fonction des caractéristiques des ingrédients, car les simulateurs d'IPF et de PA couramment utilisés utilisent des calculs complexes de la capacité de protection contre les rayons *Ultraviolet V* (UV) basés sur la concentration de chaque ingrédient actif. De plus, afin d'augmenter le

taux de prédiction de l'IPF et le PA, quatre facteurs ont été incorporés, tels que la présence de pigment, la concentration de dioxyde de titane avec un degré de pigment, le type de formulation et le type de produit, dans le modèle de prédiction. Les chercheurs suggèrent que, tout comme dans le contexte de la protection contre les rayons UV, l'incorporation de quatre facteurs des ingrédients a été importante pour améliorer la justesse des prévisions dans ce domaine d'étude, contribuant à l'optimisation de formulations cosmétiques plus efficaces et sûres.

En 2016, Huang et Boutros [67] ont examiné les effets de la sélection des paramètres sur les performances de classification en utilisant l'algorithme d'apprentissage automatique RF. À cette fin, ils ont sélectionné deux ensembles de données génomiques couramment utilisés en biologie computationnelle, en fonction du nombre de variables et du nombre d'échantillons distincts. Ils ont effectué deux étapes d'ajustement de modèle pour les deux ensembles de données. Dans un premier temps, ils ont sélectionné une large gamme de paramètres et ont entraîné un modèle de classification RF pour chaque combinaison, y compris les paramètres par défaut. Les modèles ont été formés sur l'ensemble de données d'entraînement et validés sur un ensemble de données indépendant. Ils ont évalué les performances en utilisant l'*Area Under the Receiver Operating Characteristic Curve* (AUC). Ensuite, ils ont ajusté un modèle de régression RF en utilisant les données de l'étape précédente, où les paramètres ont été définis comme des covariables et l'AUC comme réponse. Cela leur a permis de caractériser l'association entre la justesse de la prédiction et la paramétrisation. Ils ont sélectionné des échantillons aléatoires des deux tiers des ensembles de paramètres pour l'entraînement et le reste a été réservé pour la validation. Leur objectif était de prédire les scores d'AUC retenus et d'évaluer les performances en utilisant les coefficients de corrélation de *Spearman* et de *Lin*. La première étude a analysé les mesures de contrôle de qualité dans le séquençage de nouvelle génération avec quinze caractéristiques (mesures de qualité de séquençage) et a été divisée en 720 échantillons d'entraînement et 576 échantillons de validation. Chaque échantillon a été classé comme « bonne bibliothèque » ou « mauvaise bibliothèque », dans

le but de prédire cette classification. La deuxième étude a porté sur des patients atteints de cancer du poumon non à petites cellules, comprenant trois variables cliniques catégoriques et 12135 abondances continues de *messenger ribonucleic acid* (mRNA). L'objectif était de prédire le résultat du patient, catégorisé comme « sans décès » ou « décès », sur la base des données cliniques et de mRNA. Ils ont observé que les paramètres par défaut ont un potentiel de performance, mais l'ajustement des paramètres a conduit à de meilleures performances du modèle. La plupart des combinaisons de paramètres à haute performance n'ont pas suivi les tendances générales observées dans le processus de sélection de modèles. Les modèles à haute performance peuvent être le résultat du hasard ou du surajustement. Ces résultats mettent en évidence l'importance de l'ajustement des paramètres et suggèrent des améliorations de la justesse de la classification dans les publications existantes. Pour réduire le temps et le travail dans la sélection des paramètres, ils ont appliqué un modèle de régression RF, qui a prédit les performances du modèle plus précisément que la validation croisée *k-fold*, où ils ont utilisé $k = 10$. Le modèle de régression RF a également distingué les ensembles de paramètres à faible performance de ceux à haute performance. En résumé, ajuster les paramètres du RF en fonction des caractéristiques de chaque ensemble de données est une pratique courante et nécessaire pour obtenir les meilleurs résultats en matière de classification.

L'informatique joue un rôle essentiel dans le domaine biomédical, en utilisant des études biomédicales pour diagnostiquer diverses maladies, telles que le cancer, le diabète et la démence, qui peuvent conduire à la mort si elles ne sont pas diagnostiquées à un stade précoce. Avec les progrès continus dans le domaine de l'apprentissage automatique, une variété de techniques accessibles sont devenues disponibles pour prédire et prognostiquer ces maladies en se basant sur divers ensembles de données, comprenant des ensembles de données d'images et des ensembles de données au format *Comma-Separated Values* (CSV), provenant de différentes régions du monde. À cet égard, Ghazal et al. [68] ont utilisé un ensemble de données provenant de la plateforme open-source Kaggle. Cet ensemble de données comprend

les antécédents médicaux de 2067 patients diagnostiqués avec le cancer, la démence et le diabète. Il comprend 32 attributs indépendants, tels que l'âge des patients, les gènes du côté maternel et paternel, les informations sur les gènes maternels, le nombre d'avortements antérieurs, et d'autres encore. De plus, il existe un attribut dépendant pour indiquer « oui » ou « non », ainsi qu'un attribut indiquant un genre « ambigu ». Pour ce faire, ils ont adopté une approche de classification multiclasse en utilisant des techniques d'apprentissage automatique telles que le *Support Vector Machine* (SVM) et le k-NN, le jeu de données a été partitionné en deux parties distinctes. Pour l'entraînement, 70% du jeu de données ont été utilisés, ce qui correspond à 1447 instances. Pour le test, 30% du jeu de données ont été employés, totalisant 620 instances. La recherche a révélé que le modèle SVM proposé a atteint une justesse de 92,8% en entraînement et de 92,5% en test pour prédire la démence, le cancer et le diabète liés aux troubles héréditaires multifactoriels. Ces résultats indiquent que le modèle de prédiction basé sur le SVM offre des résultats remarquables par rapport au modèle k-NN qui a obtenu une justesse de 92,8% lors de l'entraînement et de 91,2% lors du test. Ces résultats renforcent l'idée que l'application de ces modèles peut jouer un rôle essentiel dans le pronostic et la prédiction de maladies, fournissant ainsi des perspectives précieuses pour la pratique clinique et la recherche biomédicale. De cette manière, Hassan et al. [58] ont utilisé diverses ressources probabilistes pour établir une approche fiable pour la prédiction du cancer du sein. Ils ont utilisé les modèles d'apprentissage automatique RF, XGBoost, k-NN et MLP. Le jeu de données *SEER Breast Cancer* a été utilisé dans cette étude et a été obtenu du référentiel *IEEE DataPort*. Le jeu de données se compose de 4024 cas avec 14 attributs distincts. Les enregistrements de diagnostics du cancer du sein féminin ont été obtenus entre 2006 et 2010, parmi les 14 attributs pertinents figuraient l'âge, l'origine ethnique, le stade T, le grade, la taille de la tumeur, entre autres. Ils ont opté pour l'application de la méthode d'évaluation des ressources *Least Absolute Shrinkage and Selection Operator* (LASSO), car elle a été utilisée pour atténuer le *overfitting* et réduire les temps d'exécution prolongés, en plus d'identifier les

ressources les plus pertinentes. Le jeu de données a été divisé en 80% pour l'entraînement et 20% pour le test. Les étapes nécessaires pour le développement de leur étude étaient de prédire le cancer du sein avec la régression logistique, les vecteurs de support et les quatre modèles susmentionnés, en plus de comparer entre eux celui qui a obtenu les meilleurs résultats en utilisant tous les ressources par rapport à celui utilisant LASSO. Le modèle RF a atteint une justesse maximale de 90,68% avec l'utilisation de la technique LASSO, comparativement à 90,44% sans celle-ci. En ce qui concerne le modèle probabiliste k-NN, il a obtenu une performance de 88,82% avec l'application de la technique LASSO, par rapport à 88,08% sans celle-ci. Pour le modèle MLP, la justesse était de 89,44% avec l'utilisation de LASSO, tandis qu'elle était de 88,19% sans. Quant au modèle XGBoost, il a atteint un taux de justesse de 90,19% avec l'application de la technique LASSO, contre 89,19% sans celle-ci. Bien que l'application de la méthode LASSO ait eu une influence positive sur les performances du modèle RF, cette influence a été modeste, représentant une différence de 0,24%. Il convient de noter que d'autres modèles, tels que k-NN qui a montré une différence de 0,74%, MLP qui a montré une différence de 1,25% et XGBoost qui a montré une différence de 1%, ont connu de meilleures performances en termes de justesse lors de l'utilisation du LASSO.

Les nanoparticules de médicaments à doses solides ont été examinées à l'aide de modèles théoriques pour étudier la faisabilité du traitement médicamenteux par un traitement vert supercritique. Améliorer la solubilité d'un médicament par nanonisation est d'une grande importance pour l'industrie pharmaceutique, car cela augmente sa biodisponibilité. Par le biais de trois modèles de régression différents, notamment la régression par processus gaussien (*Gaussian Process Regression* - PO-GPR), les k-plus proches voisins (*k-Nearest Neighbors* - PO-KNN) et le perceptron multicouche (*Multi-Layer Perceptron* - PO-MLP), il a été possible de prédire la densité du solvant et la solubilité du médicament Hioscina. Tous les modèles ont été améliorés grâce à l'application de l'algorithme *Political Optimizer* (PO). PO est un algorithme d'optimisation globale conçu pour traiter des problèmes d'optimisation impliquant

la recherche du meilleur résultat possible dans tout l'espace de solution, contrairement aux optimisations locales qui visent uniquement à trouver le meilleur résultat dans une région spécifique. Dans ce sens, le PO est également utilisé pour l'optimisation des hyperparamètres dans les algorithmes d'apprentissage automatique. La présence de l'acronyme PO dans le nom des modèles indique leur utilisation dans ce processus d'optimisation. Les résultats ont révélé que les trois méthodes optimisées ont démontré une grande justesse dans la prédiction de la densité et de la solubilité. En particulier, PO-GPR a atteint le score le plus élevé pour la solubilité (R^2 = 0,9984) et la densité (R^2 = 0,9999), tandis que le modèle PO-MLP a obtenu le score le plus élevé pour la densité (R^2 = 0,9997) et le deuxième score le plus élevé pour la solubilité (R^2 = 0,9945). Le PO-KNN a également présenté des performances solides pour la densité (R^2 = 0,9557) et la solubilité (R^2 = 0,9783). De plus, les modèles PO-GPR et PO-MLP ont montré des taux d'erreur plus faibles en termes de RMSE et de *Average Absolute Relative Deviation* (AARD)% par rapport au PO-KNN. Ces résultats indiquent que le PO-GPR et le PO-MLP ont un potentiel prometteur pour prédire avec justesse la densité et la solubilité, contribuant ainsi de manière significative à l'application réussie de la nanonisation des médicaments et à l'optimisation du processus [69].

Montavon et al. [70] ont utilisé un modèle de réseaux de neurones artificiels multi-tâches d'apprentissage automatique pour explorer les corrélations sous-jacentes entre diverses propriétés électroniques moléculaires telles que l'énergie d'atomisation, la polarisabilité, les valeurs propres des orbitales frontières, le potentiel d'ionisation, l'affinité électronique et les énergies d'excitation. Les structures moléculaires ont été représentées en utilisant le champ de force universel pour générer des géométries moléculaires cartésiennes raisonnables. De plus, les données d'entrée comprenaient les charges nucléaires et les coordonnées cartésiennes de tous les atomes, similaires aux méthodes *ab initio*. La représentation des données utilisée dans l'étude impliquait des matrices de Coulomb qui ont été cartographiées pour toutes les 14 propriétés de la molécule correspondante simultanément. Cette représentation des données

était essentielle pour alimenter le modèle de réseau neuronal profond et pour explorer les corrélations sous-jacentes entre les différentes propriétés moléculaires. Les auteurs de l'étude ont observé une réduction systématique de l'erreur à mesure que l'ensemble d'entraînement augmentait, ainsi qu'une justesse comparable, voire supérieure, aux méthodes quantiques-chimiques modernes, le tout avec un coût informatique négligeable, rendant les prédictions pratiquement instantanées par rapport aux méthodes de référence qui nécessitent plusieurs heures de CPU pour l'entraînement. Bien que l'étude se soit concentrée sur de plus petites molécules organiques, les auteurs suggèrent une applicabilité à une variété de tailles et de compositions. En conclusion, la combinaison de bases de données fiables avec l'apprentissage automatique représente une approche prometteuse pour explorer l'espace des composés chimiques et concevoir de nouveaux composés meilleurs.

Başkor et al. [71] ont étudié l'applicabilité potentielle du dexkétoprofène, un analgésique puissant avec une bonne solubilité dans l'eau et peu d'effets indésirables, dans le cadre de la formulation de produits pharmaceutiques. L'efficacité du dexkétoprofène a été attestée à la fois par des essais cliniques chez l'homme et des expériences sur des animaux. Cependant, le défi le plus important réside dans l'optimisation de sa formulation, notamment en raison du goût amer du principe actif. Dans ce contexte, une analyse statistique et l'application d'algorithmes d'apprentissage automatique ont été utilisées pour prédire les caractéristiques idéales de la formulation du dexkétoprofène. L'ensemble de données utilisé contenait des informations sur 54 formulations différentes, couvrant 10 variables, telles que la friabilité, la dureté, la variation de poids et la vitesse de dissolution. Chaque formulation a été préparée sous forme de comprimés, en variant les niveaux de revêtement avec l'eudragit (15,16% et 17,34%) et d'autres variables, telles que la quantité de Prosolv ODT, Emdex et MagnaSweet, ainsi que la pression de compression des comprimés. La méthodologie adoptée comprenait l'application de sept modèles d'apprentissage automatique, couvrant k-NN, *Support Vector Regression* (SVR), *Classification and Regression Tree* (CART), *Bootstrap Aggregating* (bagging), RF, *Gradient*

Boosting Machine (GBM) et XGBoost. Les données ont été partitionnées en ensembles d'entraînement (85%) et de test (15%), et différentes configurations de paramètres hyperboliques ont été implémentées pour optimiser les performances de chaque modèle. Les résultats ont révélé que les modèles basés sur des arbres, en particulier GBM et XGBoost, ont montré une plus grande efficacité dans la prédiction des caractéristiques souhaitées du dexkétoprofène. La stratégie proposée a permis de réduire la nécessité d'essais physiques extensifs, en anticipant des valeurs intermédiaires et, de cette manière, en optimisant la formulation. Cependant, il est important de noter que la rareté des données disponibles et la complexité de l'identification d'optimums globaux représentent des défis à relever. En résumé, l'application de l'apprentissage automatique s'est avérée un outil précieux pour l'optimisation de la formulation du dexkétoprofène, avec un potentiel pour avoir un impact positif sur l'industrie pharmaceutique en accélérant le développement de nouveaux médicaments de manière plus efficace et économique. Cette étude contribue au progrès de cette ligne de recherche, soulignant la pertinence de l'utilisation des techniques d'apprentissage automatique dans le contexte pharmaceutique pour l'optimisation et l'efficacité des formulations.

2.2.1 DISCUSSION

L'utilisation de modèles d'apprentissage automatique pour la prédiction de la concentration de matériaux dans les formulations cosmétiques émerge comme une stratégie nécessaire, soutenue par les résultats significatifs observés dans diverses disciplines de recherche. L'employabilité de ces modèles, observée dans les études de Shim et al. [66], est mise en évidence par la réalisation d'une analyse de régression utilisant la technique de l'arbre de décision, prenant en considération la concentration des filtres UV, la présence de pigment, la quantité et le degré de pigmentation du dioxyde de titane, le type de formulation et la nature du produit, a permis le développement de modèles prédictifs avancés pour la prédiction de l'(IPF) et de l'Activité Anti-Pigmentation (PA). De plus, dans l'étude de Başkor et al. [71], l'application de

l'apprentissage automatique pour l'optimisation de la formulation du *Dexketoprofène* a été observée. Les modèles d'apprentissage automatique ont été entraînés avec diverses données, réduisant l'utilisation d'ingrédients actifs et ayant un impact positif sur les coûts et le temps de développement. Malgré des défis tels que la difficulté à trouver des valeurs optimales mondiales avec des données limitées et la taille significative du nouveau jeu de données, la recherche souligne que les modèles basés sur les arbres, tels que GBM et XGBoost, ont montré des résultats prédictifs supérieurs. L'approche novatrice de cette étude offre un potentiel pour accélérer le développement de nouvelles formulations médicamenteuses, contribuant à l'efficacité et à l'économie de ressources dans l'industrie pharmaceutique et pouvant également être extrapolée au domaine des cosmétiques. La spécificité des réglages des paramètres, soulignée par Huang et Boutros [67], devient une pratique courante et nécessaire, indiquant la capacité de ces modèles de s'adapter à la singularité de chaque ensemble de données, un aspect important dans la prédiction des concentrations de matériaux dans les formulations chimiques. La recherche met l'accent sur l'analyse des effets de la paramétrisation dans les algorithmes d'apprentissage automatique non paramétriques, en soulignant l'application réussie d'ajustements pour améliorer la justesse des prévisions. L'approche englobe des méthodes de sélection exhaustives et révèle que la sensibilité des paramètres est intrinsèquement spécifique aux ensembles de données, nécessitant des ajustements personnalisés pour chaque ensemble. En explorant deux ensembles de données génomiques distincts, les chercheurs ont identifié des corrélations divergentes entre les paramètres et les scores de performance, mettant en évidence l'importance de l'ajustement approprié pour les données de différents niveaux de variabilité. L'étude souligne que le processus d'ajustement du modèle est important en apprentissage automatique et met en garde contre les risques d'une sélection de paramètres négligente, pouvant entraîner des modèles sous-optimaux et la perte potentielle de découvertes significatives.

Les études biomédicales étendent davantage le champ d'application, révélant la contribution des modèles d'apprentissage automatique à la prédiction de maladies multifactorielles. Dans l'étude de Ghazal et al. [68], l'application de l'apprentissage automatique dans la classification des maladies médicales et biomédicales est mise en avant. Le modèle proposé utilise les techniques de SVM et k-NN pour prédire la démence, le cancer et le diabète sur la base de troubles d'hérédité génétique multifactorielle. L'analyse des résultats de prédiction a pris en compte divers paramètres de performance statistique. En utilisant l'historique des troubles génétiques multifactoriels des patients, le modèle a atteint une justesse de classification de test de 92,5%, mettant en évidence l'impact significatif de l'historique médical sur la prédiction. En éclairant l'application d'algorithmes d'apprentissage automatique et de techniques de corrélation pour prédire des maladies complexes telles que le cancer, la démence et le diabète, ces résultats peuvent également être extrapolés à différents domaines des sciences médicales. Dans l'étude menée par Hassan et al. [58], des méthodes d'apprentissage automatique ont été utilisées pour faire des prédictions liées au cancer du sein. Les chercheurs ont innové en introduisant la technique de sélection de caractéristiques LASSO. L'intégration de cette approche avec l'algorithme RF a résulté en une justesse remarquable de 90,68%, mettant en évidence une amélioration significative par rapport aux efforts antérieurs.

La convergence des résultats de différentes recherches met en évidence non seulement une tendance, mais aussi la nécessité d'intégrer des modèles d'apprentissage automatique dans les recherches axées sur le développement de formulations cosmétiques, compte tenu de leur pertinence face aux défis contemporains et à la complexité inhérente au sujet spécifique de cette recherche, qui est l'utilisation de modèles d'apprentissage automatique pour suggérer la concentration de matériaux dans les formulations cosmétiques.

CHAPITRE III
IMPLÉMENTATION DE L'APPROCHE

Dans ce chapitre, nous décrivons la mise en œuvre de l'approche de cette recherche. Dans la section 3.1, nous détaillons les étapes entreprises pour la préparation des fichiers de données. Dans la section 3.2, nous présentons la description détaillée du traitement des données appliqué avant leur extraction. Dans la section 3.3, nous abordons le processus de développement de l'algorithme utilisé pour l'extraction des données. De plus, cette section décrit la modélisation des données, aboutissant à deux ensembles de données, le premier se référant aux données avant la modélisation et le second après la modélisation. Dans la section 3.4, nous expliquons notre première approche, tout comme dans la section 3.5, nous expliquons notre deuxième approche, et dans la section 3.6, nous exposons en détail la troisième approche et l'utilisation des algorithmes des modèles, ainsi que les défis que nous avons rencontrés.

Ensuite, la section fournit des explications sur la façon dont nous avons utilisé chaque modèle, y compris les étapes d'entraînement et de test. Dans la section 3.7, nous montrons comment nous avons organisé l'interface graphique utilisateur (IGU) de prédiction et son fonctionnement intégré à l'interface de programmation d'application (*Interface de Programmation Applicative* (API)) que nous avons développée. Enfin, dans la section 3.8, nous expliquons le test de comparaison que nous avons réalisé entre l'application du *OneHotEncoder* de la section 3.5 et l'application du *LabelEncoder*.

3.1 PRÉPARATION DES FICHIERS DE DONNÉES

Avant de commencer le développement de l'algorithme d'extraction des données, nous avons procédé à leur validation. Cette validation a été essentielle pour l'extraction des formulations complètes, incluant tous les matériaux et les informations essentielles.

Initialement, nous avons procédé à une validation manuelle des tableaux dans chaque feuille de calcul. Cette étape était nécessaire pour identifier des problèmes tels que l'absence de codes pour certains produits, des codes de produits incomplets ou des codes de produits incorrects. Compte tenu du grand nombre de fichiers contenant plusieurs feuilles de calcul, voir annexes A et B, nous avons choisi d'automatiser cette tâche. Nous avons développé un algorithme en Python pour parcourir tous les fichiers d'un dossier en localisant les tableaux, en identifiant l'en-tête avec le mot "Code" dans la première colonne et la concentration dans la troisième colonne, voir tableau 2.1. Nous avons utilisé l'algorithme pour calculer la somme des concentrations, qui devrait totaliser 100% dans le meilleur des cas, étant donné que certaines formulations ont une somme totale très proche de cette valeur. Ensuite, nous avons accédé manuellement à chaque feuille de calcul pour identifier celles qui avaient des valeurs inférieures ou supérieures à 100%, voir exemple dans l'annexe C. En plus des problèmes mentionnés précédemment, nous avons identifié des tableaux avec des en-têtes différents, plusieurs tableaux dans une seule feuille de calcul, un ordre différent des colonnes et des tableaux dans des positions différentes dans les feuilles de calcul.

3.2 TRAITEMENT DES DONNÉES

Toutes les incohérences liées aux données et à la structure des tableaux ont été traitées manuellement. Ce n'est qu'après ce traitement que nous avons pu exécuter l'algorithme pour extraire intégralement les données des tableaux. Les incohérences rencontrées étaient variées, comme illustré dans les tableaux 3.1 et 3.2, ainsi que les annexes D et E, respectivement.

TABLEAU 3.1 : Incohérence des données et solution
©Christian Gonzalo Frantz Segovia, 2024

Incohérence	Solution
Tableau de formulation sans le code du produit	Recherche du code du produit dans d'autres tableaux
Tableau de formulation avec code produit avec informations complémentaires	Recherche du code du produit dans d'autres tableaux
Tableau de formulation avec un code de produit incorrect	Recherche du code du produit dans d'autres tableaux

TABLEAU 3.2 : Incohérence des tableaux et solution
©Christian Gonzalo Frantz Segovia, 2024

Incohérence	Solution
Tableau sans l'en-tête correct	Ajustement de l'en-tête
Multiples tableaux dans une feuille de calcul	Correction pour que chaque tableau soit dans sa propre feuille de calcul
Tableau avec un ordre de colonnes différent	Ajustement pour que chaque tableau ait le même ordre par rapport aux colonnes *Code* et *Concentration (%)*
Tableau à n'importe quelle position dans la feuille de calcul	Ajustement pour que le tableau soit situé dans la première colonne
Tableau sans l'en-tête	Ajout d'en-tête

Dans le tableau 3.3, vous pouvez visualiser la formulation chimique d'exemple, où chaque ligne comprend le code, le nom du matériau et sa concentration en pourcentage. Il est important de noter que le code de la matière première est composé de deux parties : la première est une séquence de lettres, pouvant également être alphanumérique, que nous appelons « acronyme », suivie d'un ensemble de chiffres que nous appelons « code ».

TABLEAU 3.3 : Exemples des données des matières premières

acronym	code	concentration
EAU,	10,	0.615
HUME,	2,	0.01
ADDI,	32,	0.011
ADDI,	14,	0.016
EMUL,	35,	0.004
CGRA,	48,	0.013
EMUL,	44,	0.037
CGRA,	581,	0.02
SILI,	7,	0.04
HUIL,	40,	0.04
CONS,	37,	0.017
CONS,	23,	0.024
HUME,	16,	0.04
ACTI,	740,	0.012
ACTI,	345,	0.011
ACTI,	201,	0.026
ADDI,	49,	0.001
EAU,	10,	0.05
ACTI,	703,	0.008
FRAG,	7,	0.005

3.3 DÉVELOPPEMENT D'ALGORITHMES POUR L'EXTRACTION DE DONNÉES

Nous avons développé l'algorithme en Python dans le but d'extraire les codes des matériaux ainsi que leurs concentrations chimiques respectives, qui ont ensuite été enregistrées dans un fichier au format CSV, tableau 3.3.

3.3.1 EXTRACTION DE DONNÉES

L'algorithme parcourt chaque fichier Excel dans le dossier spécifié dans le code, extrayant les données nécessaires des tableaux contenus dans les feuilles de calcul. Cette extraction de données est effectuée en recherchant le terme *Code* dans l'en-tête du tableau, situé dans la

première colonne de chaque feuille. Dès que le terme *Code* est identifié, l'algorithme parcourt chaque ligne du tableau pour extraire le code du matériau et sa concentration respective. Les concentrations des matériaux de la formulation se trouvent dans la troisième colonne de la feuille, tandis que d'autres informations sont ignorées. Lors de la recherche dans le tableau, nous sautons les lignes qui ne contiennent pas de matériaux, pouvant être des lignes vides, pouvant indiquer le numéro de la phase dans laquelle le matériau a été utilisé ou pouvant contenir une légende. Le code du matériau est divisé en deux parties, la première étant ce que nous appelons un acronyme et la deuxième partie étant ce que nous appelons un code. Pour ce faire, nous avons développé deux expressions régulières. Les figures 3.2 et 3.4 illustrent les acronymes, et les figures 3.1 et 3.3 illustrent les codes identifiés par l'expression régulière, respectivement.

```
([A-Za-z]+\s?(?=\d)(?:\d+[A-Za-z]+)?)
```

FIGURE 3.1 : Regex utilisée pour capturer l'acronyme du code du matériau

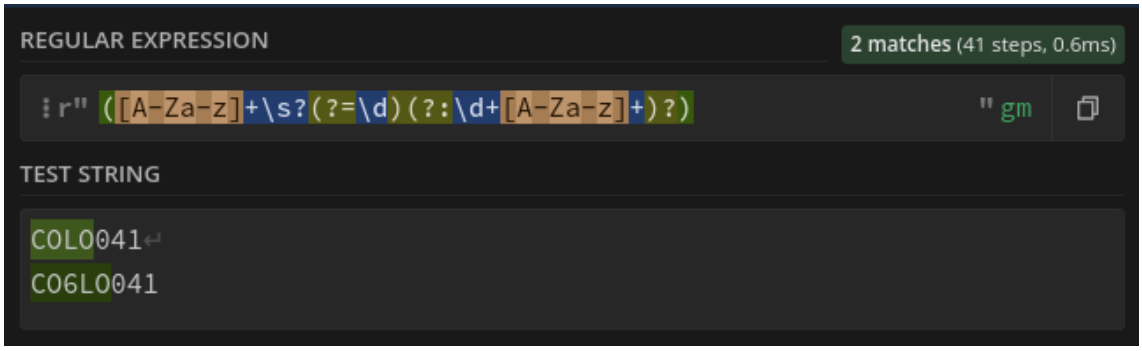


FIGURE 3.2 : Exemples d'extraits de acronymes identifiés par l'regex 3.1

Ci-dessous, nous détaillons l'expression régulière en ses parties constitutives :

— ([A-Za-z]+) : Il s'agit d'un groupe de capture qui correspond à une ou plusieurs lettres majuscules ou minuscules. La classe de caractères [A-Za-z] représente n'importe quelle lettre de l'alphabet, et le quantificateur + indique une ou plusieurs occurrences de ces lettres.

— \s ? : Il s'agit d'un caractère facultatif d'espace blanc. Le \s correspond à n'importe quel caractère d'espace blanc, et le ? indique que la présence de ce caractère est facultative.

— (?=\d) : Il s'agit d'un lookahead positif, qui est une construction qui vérifie si une expression donnée se produit plus loin dans la séquence, sans consommer de caractères. Dans ce cas, nous cherchons à nous assurer qu'il y a un chiffre (\d) devant. Le résultat pratique de cela est que la séquence de lettres et d'espaces ne sera capturée que si elle est suivie d'un chiffre.

— (? :\d+[A-Za-z]+) ? : Il s'agit d'un groupe de non-capture (groupé par ? :) qui correspond à une séquence optionnelle d'un ou plusieurs chiffres suivis d'une ou plusieurs lettres. Le ? à la fin rend tout ce groupe facultatif.

En utilisant le code du matériau CO6LO041 comme base, l'expression régulière ci-dessus extrait CO6LO.

```
(\s?\d*(?!.*[A-Za-z]))
```

FIGURE 3.3 : Regex utilisée pour capturer le code du code du matériau

FIGURE 3.4 : Exemples d'extraits de code identifiés par l'regex 3.3
©Christian Gonzalo Frantz Segovia, 2024

Ci-dessous, nous détaillons l'expression régulière en ses parties constitutives :

— (\s ?\d*) : Ce est un groupe de capture qui correspond à zéro ou un caractère d'espace blanc suivi de zéro ou plusieurs chiffres. Le \s correspond à n'importe quel caractère d'espace blanc, le ? indique que l'espace blanc est facultatif, et le * correspond à zéro ou plusieurs chiffres.

— (?!.[A-Za-z]) : Ce est un lookahead négatif qui vérifie s'il n'y a aucune lettre (majuscule ou minuscule) nulle part devant la séquence. Le .[A-Za-z] correspond à n'importe quelle séquence suivie d'une lettre, et le (?! ...) nie cette condition.

En utilisant le code de matériau CO6LO041 comme base, l'expression régulière ci-dessus extrait 041.

Le tableau 3.3 présente des exemples de données sur les matières premières obtenues à l'aide de l'algorithme d'extraction de données et enregistrées dans un fichier CSV. Ces informations ont été acquises dans le but de prédire la concentration des matières premières dans les produits cosmétiques en utilisant les modèles d'apprentissage automatique développés au cours de notre étude. L'application de l'algorithme mentionné a permis l'extraction de

77038 données, et cette extraction a été effectuée sans manipulation préalable des données, c'est-à-dire que ces données sont brutes.

3.3.2 MODÉLISATION DES DONNÉES

Après avoir extrait les données des feuilles de calcul, nous avons procédé à leur modélisation. Une fois que nous avons obtenu les données pour une formulation spécifique, comme illustré dans le tableau 3.3, nous avons conservé cet ensemble de données et appliqué une opération de rotation, déplaçant le premier matériau en dernière position dans l'ensemble. Cette opération a été répétée séquentiellement jusqu'à ce que chaque matériau occupe la première position, suivant cette séquence jusqu'à ce que le matériau qui était initialement en dernière position occupe la première position, comme le template de données illustré dans la figure 3.5 et l'exemple illustratif avec les vrais données en annexe F. Ainsi, pour une formulation comprenant un total de dix matériaux, nous avons obtenu un total de 10 variations distinctes de la formulation. Ce processus a été réalisé pour toutes les formulations disponibles.

Grâce à cette stratégie, nous avons cherché à obtenir une représentation plus complète et diversifiée du jeu de données. La rotation des lignes des formulations nous a permise d'explorer différentes perspectives d'entraînement. La prise en compte de l'ordre des matériaux dans les formulations est importante, car la disposition séquentielle peut influencer l'apprentissage des modèles d'apprentissage automatique, en particulier les algorithmes sensibles à la séquence d'entraînement. La rotation des lignes, en créant de nouvelles dispositions des matériaux, visait à fournir une exploration plus approfondie de l'espace des caractéristiques, améliorant ainsi l'entraînement des modèles.

- item 1, %, item 2, %, item 3, %, . . ., item N-1, %, item N, %
- item 2, %, item 3, %, item 4, %, . . ., item N, %, item 1, %
- . . .
- item N, %, item 1, %, item 2, %, . . ., item N-2, %, item N-1, %

FIGURE 3.5 : Exemple de rotation des matériaux
©Christian Gonzalo Frantz Segovia, 2024

En plus de cette modélisation, nous avons identifié une moyenne totale équivalente à seize matériaux dans les formulations et complété avec des zéros lorsque la formulation comptait moins de matériaux, comme illustré dans la figure 3.6.

Grâce à cette stratégie, notre objectif était de mitiger les disparités dans le nombre de matériaux présents dans les formulations. En incorporant des lignes supplémentaires pour les formulations qui contenaient à l'origine moins de 16 matériaux, nous avons introduit des valeurs nulles pour les caractéristiques spécifiques (acronyme, code et concentration) afin de uniformiser structurellement l'ensemble de données. Cette uniformisation a aidé à réduire les distorsions et les déséquilibres pendant l'entraînement des modèles d'apprentissage automatique.

- item 1, %, item 2, . . ., %, item 14, %, 0, 0(%), 0, 0(%)
- item 2, %, item 3, %, item 4, %, . . ., 0, 0(%), item 1, %
- . . .
- 0, 0(%), item 1, %, item 2, %, . . ., item 14, %, 0, 0(%)

FIGURE 3.6 : Exemple d'ajustement avec la moyenne de dix matériaux par formulation
©Christian Gonzalo Frantz Segovia, 2024

Après l'application de la modélisation des données mentionnée dans cette sous-section, l'algorithme mentionné a permis l'extraction de 1679522 enregistrements. Avec ce modèle, nous recevons un ensemble d'éléments et de leurs concentrations, à partir d'un élément cible, et nous prédisons ainsi sa concentration. De cette manière, nous espérons que le modèle

fournira une estimation plus précise que lors de l'extraction précédente, où la modélisation des données n'a pas été réalisée.

3.4 PREMIÈRE APPROCHE : DÉVELOPPEMENT D'ALGORITHMES D'APPRENTISSAGE AUTOMATIQUE

Nous avons considéré un échantillon de données contenant des informations sur les matériaux, leurs acronymes (*acronym*), leurs codes (*code*) et leurs concentrations correspondantes (*concentration*).

Notre première approche s'est basée sur des expériences en prévision statistique simple utilisant les données extraites 3.3.1 sans l'application de la modélisation des données 3.3.2. La prévision était basée sur un seul matériau, indépendamment du nombre de matériaux dont les concentrations devaient être prévues. En d'autres termes, la prévision de la concentration d'un matériau n'affectait pas la prévision des autres. Pour cette approche, nous avons implémenté un algorithme utilisant le modèle de DTR en Python avec la bibliothèque d'apprentissage automatique Scikit-Learn [72, 73].

Pour que notre algorithme d'apprentissage automatique puisse traiter nos données d'entrée, celles-ci doivent être au format numérique. Cependant, dans nos données, nous avons une colonne appelée *acronym* qui est au format de variable catégorielle. Pour permettre l'utilisation de ces données dans notre algorithme, nous avons appliqué des techniques de prétraitement, telles que le *OneHotEncoder*. Dans ce sens, Scikit-Learn a été utilisé pour diviser les ensembles de données en sous-ensembles d'entraînement et de test et ensuite pour fournir l'encodeur. La technique du OneHotEncoder a généré des colonnes binaires pour chaque catégorie unique, indiquant la présence (1) ou l'absence (0) de cette catégorie pour chaque observation. Par exemple, « EAU », « ACTI » et « ADDI » seraient représentés comme [1, 0, 0], [0, 1, 0] et [0, 0, 1], respectivement, comme présenté dans le tableau 3.4. Cette

technique a traité et interprété correctement les variables catégorielles. Cela a permis une modélisation plus précise et efficace.

TABLEAU 3.4 : Transformation OneHotEncoder
©Christian Gonzalo Frantz Segovia, 2024

Données non traitées	Données traitées
ACTI, 201, 0.004	[1, 0, 0, 201, 0.004]
ADDI, 49, 0.005	[0, 1, 0, 49, 0.005]
EAU, 10, 0.07	[0, 0, 1, 10, 0.07]
ACTI, 703, 0.001	[1, 0, 0, 703, 0.001]
FRAG, 7, 0.001	[0, 0, 0, 7, 0.001]
EAU, 10, 0.631	[0, 0, 1, 10, 0.631]

3.4.1 DECISION TREE REGRESSOR

En créant une instance de la classe DecisionTreeRegressor, comme illustré dans la figure 3.7, nous avons choisi de conserver les valeurs par défaut des hyperparamètres. Tout d'abord, le critère d'évaluation de la qualité des divisions dans les nœuds de l'arbre a été déterminé par le paramètre *criterion*. Avec la valeur *squared_error*, ce critère est particulièrement adapté aux problèmes de régression, visant à minimiser la moyenne des carrés des différences entre les prédictions du modèle et les valeurs observées. Le deuxième hyperparamètre est le *splitter*, qui a spécifié la stratégie utilisée pour diviser les nœuds internes de l'arbre. Avec la valeur par défaut *best*, le modèle a sélectionné la division optimale en fonction du critère *criterion*, contribuant ainsi à la construction d'un arbre maximisant la qualité des prédictions. De plus, les hyperparamètres *min_samples_split* et *min_samples_leaf* ont régulé la complexité de l'arbre. Ces hyperparamètres ont défini le nombre minimum d'échantillons nécessaires pour effectuer une division dans un nœud interne et le nombre minimum d'échantillons requis pour une feuille, respectivement. Avec des valeurs par défaut de 2 et 1, ils ont influencé la quantité minimale de données nécessaires pour prendre des décisions structurelles lors de la construction de l'arbre.

```
self.model = DecisionTreeRegressor(
    criterion="squared_error",
    splitter="best",
    min_samples_split=2,
    min_samples_leaf=1
)
```

FIGURE 3.7 : Constructeur DecisionTreeRegressor.

3.5 DEUXIÈME APPROCHE : DÉVELOPPEMENT D'ALGORITHMES D'APPRENTISSAGE AUTOMATIQUE

Tout comme dans la première approche 3.4, nous considérons l'échantillon de données contenant des informations sur les matériaux, leurs acronymes (*acronym*), leurs codes (*code*) et leurs concentrations (*concentration*) correspondantes.

Dans notre deuxième approche, pour la prédiction des concentrations en fonction des codes et des acronymes fournis, nous avons choisi quatre modèles : le RFR, le XGBoost, le k-NN et le MLP qui ont été développés en utilisant *Python* avec la bibliothèque d'apprentissage automatique *Scikit-Learn* [72, 73]. Le choix des modèles RFR, XGBoost, k-NN et MLP s'est basé sur leurs caractéristiques distinctes et complémentaires, dans le but de surmonter les limitations individuelles de chaque algorithme pour la tâche proposée. Le RFR a été sélectionné pour sa capacité à gérer le surajustement en utilisant plusieurs arbres de décision. Cependant, son interprétabilité et sa performance sur de petits ensembles de données, considérés comme des points faibles, ne s'appliquent pas à notre étude, étant donné le grand volume de données obtenu pour cette recherche. En complément, le XGBoost apporte une efficacité dans l'ajustement des hyperparamètres et offre de bonnes performances, bien qu'il soit sensible au choix de ces paramètres. Le k-NN, bien qu'il soit sensible à l'échelle des données et exigeant en termes de stockage, fournit des informations précieuses avec son

approche basée sur la similarité des données. Enfin, le MLP offre la flexibilité d'un modèle de réseau neuronal profond, bien qu'il nécessite un grand volume de données et un ajustement minutieux des hyperparamètres. En combinant ces modèles, nous cherchons à aborder les limitations individuelles de chaque algorithme, en exploitant leurs avantages pour prédire les concentrations de matériaux dans les formulations cosmétiques.

Comme mentionné dans la section 3.4, la technique de prétraitement *OneHotEncoder* a été appliquée, et en plus de cette technique, nous avons appliqué la validation croisée *k-fold* où nous avons calculé la moyenne et l'écart type des métriques calculées pour chaque *fold*.

3.5.1 RANDOM FOREST REGRESSOR

Lors de la création de l'instance de la classe RandomForestRegressor, comme illustré dans la figure 3.8, nous avons fourni en argument la valeur 100 pour l'hyperparamètre *n_estimators*, qui a régulé le nombre d'arbres de décision dans le modèle de forêt aléatoire. De plus, le *criterion* a spécifié la fonction utilisée pour mesurer la qualité des divisions dans chaque arbre de décision de la forêt. Avec la valeur par défaut *squared_error*, la métrique de l'erreur quadratique moyenne a été employée. En outre, les hyperparamètres *min_samples_split* et *min_samples_leaf* ont déterminé le nombre minimum d'échantillons nécessaires pour effectuer une division dans un nœud interne et le nombre minimum d'échantillons requis pour une feuille, respectivement. Avec des valeurs par défaut de 2 et 1, ces paramètres ont exercé un contrôle sur la complexité des arbres individuels dans la forêt, ce qui a contribué à régulariser le modèle et à atténuer les problèmes de surajustement. L'hyperparamètre *max_features* a contrôlé le nombre maximal de caractéristiques à considérer lors de la recherche de la meilleure division dans chaque arbre de la forêt. Avec la valeur par défaut de 1.0, le modèle a considéré toutes les caractéristiques pour chaque division. L'hyperparamètre *bootstrap* a déterminé si l'échantillonnage avec remplacement est utilisé lors de la construction des

arbres dans la forêt. Avec la valeur par défaut *True*, chaque arbre a été construit à partir d'un échantillon aléatoire des données d'entraînement, permettant l'inclusion d'instances répétées et la formation de jeux de données uniques pour chaque arbre. De plus, nous avons spécifié l'hyperparamètre *random_state* comme étant 42, qui a été utilisé pour initialiser le générateur de nombres pseudo-aléatoires interne, responsable de la partition aléatoire des données lors de la construction des arbres, permettant ainsi que le processus de construction du modèle soit reproductible.

```
self.model = RandomForestRegressor(
    n_estimators=100,
    criterion="squared_error",
    min_samples_split=2,
    min_samples_leaf=1,
    max_features=1.0,
    bootstrap=True,
    random_state=42
)
```

FIGURE 3.8 : Constructeur RandomForestRegressor.

3.5.2 EXTREME GRADIENT BOOSTING

Lors de la création d'une instance de la classe XGBRegressor de la bibliothèque XGBoost illustrée dans la figure 3.9, nous avons spécifié *reg :squarederror* comme la valeur attribuée à l'hyperparamètre *objective*. Il s'agit d'un argument du constructeur qui a pour objectif de déterminer la fonction de perte (*loss function*) qui sera optimisée pendant la phase d'entraînement du modèle. De manière simplifiée, la fonction objective est une métrique que l'algorithme tente de minimiser ou de maximiser, en fonction du problème existant.

```
self.model = xgb.XGBRegressor(objective="reg:squarederror")
```

FIGURE 3.9 : Constructeur xgb.XGBRegressor.
©Christian Gonzalo Frantz Segovia, 2024

3.5.3 K-NEAREST NEIGHBORS

Lors de la création d'une instance de la classe KNeighborsRegressor, comme illustré dans la figure 3.10, nous avons choisi de conserver les valeurs par défaut des hyperparamètres. Tout d'abord, l'hyperparamètre *n_neighbors*, défini avec la valeur par défaut de 5, a influencé la lissage ou la granularité des prédictions. Un autre hyperparamètre est le *weights*, qui a déterminé la stratégie de pondération attribuée aux voisins lors de la prédiction. Avec la valeur par défaut *uniform*, tous les voisins ont des poids égaux. De plus, l'hyperparamètre *algorithm* a spécifié l'algorithme utilisé pour calculer les voisins les plus proches. Avec la valeur par défaut "auto", l'algorithme le plus approprié a été sélectionné automatiquement en fonction des données d'entrée et des paramètres fournis. Quant à l'hyperparamètre *leaf_size*, avec la valeur par défaut de 30, il a déterminé la taille de la feuille passée à l'algorithme de construction de l'arbre. L'hyperparamètre *p* est utilisé pour la métrique de distance *Minkowski*, qui est une généralisation de la métrique euclidienne. La valeur par défaut pour *p* est 2, ce qui indique que nous utilisons la distance euclidienne. Enfin, l'hyperparamètre *metric* a spécifié la métrique de distance utilisée pour calculer la similarité entre les points dans l'espace des caractéristiques. Avec la valeur par défaut *minkowski*, le modèle utilise la métrique *Minkowski*, qui est une généralisation de la distance euclidienne et de la distance de *Manhattan*.

```
self.model = KNeighborsRegressor(
    n_neighbors=5,
    weights="uniform",
    algorithm="auto",
    leaf_size=30,
    p=2,
    metric="minkowski"
)
```

FIGURE 3.10 : Constructeur KNeighborsRegressor.
©Christian Gonzalo Frantz Segovia, 2024

3.5.4 MULTI-LAYER PERCEPTRON

Lors de l'instanciation de la classe MLPRegressor, comme illustré dans la figure 3.11, nous avons attribué des valeurs spécifiques à plusieurs hyperparamètres pour configurer le modèle de MLP. L'argument *activation* a été défini comme *logistic*, ce qui indique l'utilisation de la fonction d'activation logistique dans les couches cachées. De plus, l'hyperparamètre *solver* a déterminé l'algorithme d'optimisation utilisé pour entraîner le réseau neuronal et ajuster ses poids. Avec la valeur par défaut *adam*, l'algorithme d'optimisation basé sur le gradient stochastique est adaptatif, efficace pour les problèmes de régression avec des ensembles de données de taille modérée à grande. La régularisation du modèle a été contrôlée par l'hyperparamètre *alpha*, établi à 0.0001, pour prévenir le surajustement. La taille des lots de données utilisés pendant l'entraînement du réseau neuronal a été déterminée par le paramètre *batch_size*. La valeur par défaut de ce paramètre, *auto*, signifiait que la taille du lot était ajustée automatiquement en fonction de la taille de l'ensemble d'entraînement. En ce qui concerne le taux d'apprentissage, nous avons opté pour une approche *learning_rate="constant"*, maintenant le taux inchangé pendant l'entraînement, et *learning_rate_init* a été défini comme 0.01, représentant le taux d'apprentissage initial. Le pas de dégradation de l'apprentissage a été contrôlé par l'hyperparamètre *power_t* pendant l'entraînement. La configuration par défaut

est de 0.5, indiquant une décroissance du pas d'apprentissage selon un taux de racine carrée inverse du nombre total de mises à jour de gradient. Le nombre maximal d'itérations a été limité à 500 grâce à l'hyperparamètre *max_iter*. La randomisation des données d'entraînement avant chaque époque a été assurée par le paramètre *shuffle*, défini par défaut comme *True*. L'hyperparamètre *tol* a défini la tolérance pour la convergence du processus d'optimisation pendant l'entraînement du modèle. Avec la configuration par défaut de 0.0001, l'entraînement a été interrompu lorsque l'amélioration des performances du modèle est devenue insignifiante. L'utilisation de l'hyperparamètre *momentum* a eu pour effet de lisser la trajectoire de mise à jour des poids pendant l'entraînement, aidant le modèle à éviter les minima locaux ou les régions plates de la fonction de coût. L'utilisation de *nesterovs_momentum* a accéléré le processus de convergence du modèle. Il ajuste la manière dont les mises à jour des poids du modèle sont effectuées. La valeur par défaut de l'hyperparamètre *validation_fraction* est définie à 0.1, indiquant qu'une fraction de 10% des données d'entraînement a été réservée pour la validation pendant l'entraînement du modèle. Les hyperparamètres *beta_1* et *beta_2* ont été définis respectivement à 0.9 et 0.999, représentant les valeurs de décroissance pour les estimations des moments du premier ordre (moyenne mobile du gradient) et du deuxième ordre (moyenne mobile du gradient au carré) dans l'algorithme *Adam*. L'epsilon a été défini à 1e-08, évitant la division par zéro et traitant les cas d'instabilité numérique lors du calcul des mises à jour des poids dans l'algorithme d'optimisation. *n_iter_no_change* a été défini à 10, déterminant le nombre maximal d'époques consécutives sans amélioration du score de validation avant d'arrêter l'entraînement. Ce paramètre a été utilisé comme critère d'arrêt anticipé pour éviter le surajustement et améliorer l'efficacité de l'entraînement. *max_fun* a contrôlé le nombre maximal de fois que la fonction de perte ou de coût (également appelée fonction objectif) peut être évaluée pendant le processus d'optimisation. L'architecture du réseau a été déterminée par l'argument *hidden_layer_sizes=(100, 50)*, indiquant deux couches cachées avec 100 et 50 neurones respectivement. Enfin, pour assurer la reproductibilité, *random_state* a été fixé

à 42, déterminant la graine pour la randomisation interne du processus d'entraînement. Ces choix spécifiques d'hyperparamètres ont été guidés par des considérations expérimentales et par l'ajustement fin du modèle pour répondre aux caractéristiques et aux exigences spécifiques du problème en question.

```
self.model = MLPRegressor(
    activation="logistic",
    solver="adam",
    alpha=0.0001,
    batch_size="auto",
    learning_rate="constant",
    learning_rate_init=0.01,
    power_t=0.5,
    max_iter=500,
    shuffle=True,
    tol=0.0001,
    momentum=0.9,
    nesterovs_momentum=True,
    validation_fraction=0.1,
    beta_1=0.9,
    beta_2=0.999,
    epsilon=1e-08,
    n_iter_no_change=10,
    max_fun=15000,
    hidden_layer_sizes=(100, 50),
    random_state=42
)
```

FIGURE 3.11 : Constructeur MLPRegressor.

3.6 TROISIÈME APPROCHE : DÉVELOPPEMENT D'ALGORITHMES D'APPRENTISSAGE AUTOMATIQUE

Dans la troisième approche, nous considérons l'échantillon de données contenant des informations sur les matériaux, leurs acronymes (*acronym*), leurs codes (*code*) et leurs concentrations (*concentration*) correspondantes sur lesquels nous avons appliqué la modélisation des données, comme indiqué à la section 3.3.2. Nous avons également utilisé les indicateurs de performance détaillés à la section 2.1.6.

Dans notre troisième et dernière approche, nous conservons les modèles sélectionnés dans la section 3.5 pour les mêmes raisons mentionnées précédemment.

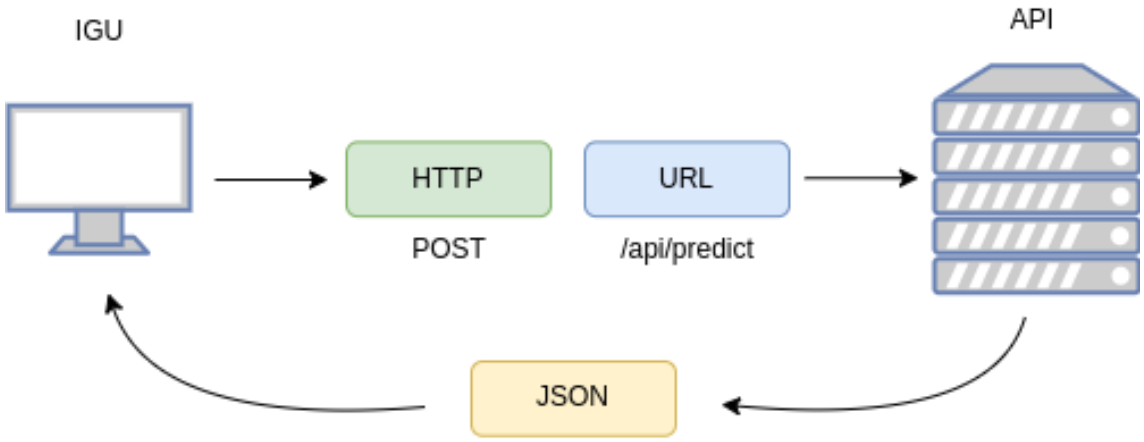
Comme mentionné dans la section 3.5, la technique de prétraitement *OneHotEncoder* a été appliquée, et en plus de cette technique, nous avons également appliqué la technique de validation croisée *k-fold* où nous avons calculé la moyenne et l'écart type des métriques calculées pour chaque *fold*.

En raison de la robustesse et de la capacité de généralisation de ces modèles, associées à la grande quantité d'enregistrements dans l'ensemble de données, l'application du RFR, du XGBoost, du k-NN et du MLP a permis de prédire les concentrations en fonction des acronymes et des codes.

3.7 LOGICIEL DE PRÉVISION

Nous avons développé une API et une IGU où un utilisateur pourra prédire les formulations cosmétiques, illustré dans la figure 3.12. Au fur et à mesure que l'acronyme et le code du matériau sont ajoutés dans le formulaire, le système effectuera une requête vers l'API qui renverra la concentration pour ce matériau en temps réel. Nous avons construit l'API en

utilisant le langage de programmation *Python* avec le *framework Flask* et développé l'IGU avec le *framework Vue.js*.




```
[
  {
    "acronym": "EAU",
    "code": "003",
    "concentration": 0.1732582013177667
  },
  {
    "acronym": "GELI",
    "code": "005",
    "concentration": 0.024000000000000018
  },
  {
    "acronym": "HUIL",
    "code": "044",
    "concentration": 0.07821290152207389
  }
]
```


FIGURE 3.12 : Diagramme de l'API
©Christian Gonzalo Frantz Segovia, 2024

3.8 TESTANT LES EFFETS DU CHANGEMENT D'ENCODEUR

Nous avons comparé les résultats de l'application de la validation croisée *k-fold* obtenus dans la troisième approche où le *OneHotEncoder* a été utilisé avec les résultats obtenus en utilisant le *LabelEncoder*. La technique *LabelEncoder* a transformé les variables catégorielles

en nombres entiers, attribuant une valeur unique à chaque variable. Par exemple, « ACTI » peut être mappé comme 0, « ADDI » comme 1, et ainsi de suite, comme illustré dans le tableau 3.5. Les deux techniques ont traité et interprété correctement les variables catégorielles. Cela a permis une modélisation plus précise et efficace.

TABLEAU 3.5 : Transformation LabelEncoder
©Christian Gonzalo Frantz Segovia, 2024

Données non traitées	Données traitées
ACTI, 201, 0.004	0, 201, 0.004
ADDI, 49, 0.005	1, 49, 0.005
EAU, 10, 0.07	2, 10, 0.07
ACTI, 703, 0.001	0, 703, 0.001
FRAG, 7, 0.001	3, 7, 0.001
EAU, 10, 0.631	2, 10, 0.631

CHAPITRE IV

VALIDATION

Dans ce chapitre, nous décrivons les expérimentations réalisées pour valider les approches proposées au Chapitre 3. Nous avons mené à bien l'implémentation et le développement de l'algorithme pour le traitement et l'extraction des données, ainsi que l'utilisation d'un algorithme d'apprentissage automatique dans la première approche et de quatre algorithmes d'apprentissage automatique développés pour la deuxième et la troisième approche. Pour les deux approches, nous avons utilisé le langage de programmation *Python* avec la bibliothèque d'apprentissage automatique *Scikit-Learn* [72, 73].

Nous avons évalué la justesse de chacun des modèles proposés. Dans la première approche, le DTR et dans les deuxième et troisième approches, le RFR, le XGBoost, le k-NN et le MLP.

Dans la section 4.1 (Étude de cas), nous détaillons cette étude de cas en présentant les différentes étapes de sa conception pour chaque approche, depuis l'obtention des résultats jusqu'à l'analyse. Dans la section 4.2 (Testant les effets du changement d'encodeur), nous testons les effets du changement de l'encodeur utilisé dans la troisième approche. Quant à la section 4.3 (Discussion), nous nous consacrons à l'analyse des données collectées et à une réflexion approfondie à la lumière de la littérature actuelle, où nous répondons à la question de recherche.

4.1 ÉTUDE DE CAS

Dans la sous-section 4.1.1 (Conception), nous présentons les conditions et les étapes impliquées dans la conception de notre approche pour le développement de l'algorithme de

collecte et de modélisation des données, ainsi que l'utilisation des algorithmes d'apprentissage automatique. Puis, dans la sous-section 4.1.2 (Préparation et collecte), nous décrivons les critères relatifs à l'extraction de données et à l'entraînement des modèles. Enfin, dans la sous-section 4.1.3 (Résultats), nous exposons les résultats obtenus à partir des mesures de performance et des graphiques, permettant l'analyse des modèles entraînés.

4.1.1 CONCEPTION

Nos algorithmes représentent des solutions pour atteindre l'objectif proposé. Le premier algorithme exécute l'extraction et le traitement des données, qui seront utilisés par les modèles de prédiction, comme détaillé dans la section 4.1.2 (Préparation et collecte). Dans la première approche, le modèle d'apprentissage automatique proposé a subi un entraînement et une prédiction en utilisant ses ensembles de données correspondants. La prédiction a été effectuée de manière individuelle pour chaque matériau dans l'ensemble, sans tenir compte de la quantité et du type spécifique des matériaux demandés. Dans la deuxième approche, les quatre modèles d'apprentissage automatique proposés effectuent l'entraînement et la prédiction de leurs ensembles de données respectifs, en tenant compte de la quantité et du type spécifique des matériaux demandés. Dans la section 4.1.3 (Résultats), nous évaluons les résultats. Dans les première et deuxième approches, l'évaluation s'est faite au moyen de métriques et d'analyses. Dans la troisième approche, l'évaluation s'est faite au moyen de métriques et de graphiques. Nous avons réalisé une analyse pour comparer les résultats entre les modèles.

4.1.2 PRÉPARATION ET COLLECTE

À cette étape, nous avons développé l'algorithme pour l'extraction et le traitement des données. Les données, comme mentionné précédemment, ont été obtenues en collaboration avec le Dr Lionel Ripoll, professeur au département des sciences fondamentales de l'Université

du Québec à Chicoutimi. Au total, 258 fichiers Excel ont été analysés, totalisant 4806 tableaux. Avant la mise en œuvre des méthodes qui ont traité les données, nous avons acquis 77038 enregistrements. Après l'application des étapes de traitement, nous avons obtenu un total de 1679522 enregistrements. Ainsi, nous avons travaillé avec les deux groupes de données, appelés 1er groupe et 2e groupe. Les fichiers CSV contenant ces ensembles de données ont été importés par l'algorithme et divisés en ensembles d'entraînement (80%) et de test (20%).

4.1.3 RÉSULTATS

PREMIÈRE APPROCHE

Dans la première approche, nous avons entraîné et testé le modèle DTR où nous avons appliqué la technique de prétraitement *OneHotEncoder* avec le 1er groupe de données. Ensuite, nous avons choisi un ensemble initial de matériaux de manière aléatoire pour prédire leurs concentrations, comme indiqué dans le tableau 4.1. Dans cette approche, les matériaux ont été traités de manière indépendante, ce qui indique qu'ils n'étaient pas intégrés dans une formulation cosmétique spécifique, sans tenir compte des autres matériaux de la formulation.

Dans le tableau 4.1, nous présentons les résultats de la prédiction des matériaux choisis de manière aléatoire. Ils ont été soumis à l'évaluation du Dr. Lionel Ripoll, spécialiste du domaine, qui a validé les concentrations prédites, affirmant qu'elles se rapprochent des quantités typiques utilisées dans les formulations cosmétiques. Après ce retour positif, nous avons décidé de passer à la deuxième approche, où nous travaillons sur la prédiction de tous les matériaux d'une formulation cosmétique, comme indiqué dans la sous-section 4.1.3 (Deuxième approche).

Acronym	Code	Concentration estimée
EAU	003	0,19636482
CONS	014	0,00132341
ACTI	627	0,01
GELI	005	0,024
HUIL	044	0,07848684
HUIL	051	0,03963529
EXTR	118	0,02138462
ACTI	183	0,01625
ACTI	019	0,03953488
HUME	016	0,01639566
CONS	023	0,00099747
FRAG	003	0,01

TABLEAU 4.1 : Prédictions de la première approche.

DEUXIÈME APPROCHE

Dans la deuxième approche, nous avons appliqué la méthode de validation croisée *k-fold*, avec *k* égale à 5, et l'avons appliquée à l'ensemble des données. Nous avons entraîné et testé les modèles RFR, XGBoost, k-NN et MLP où nous avons appliqué la technique de prétraitement *OneHotEncoder* avec le 1er groupe de données pour chaque *fold*, dans tous les modèles. Ainsi, nous avons obtenu le tableau 4.2 avec les moyennes et les écarts-types des métriques.

	MSE		RMSE		MAE		R^2	
	Moyenne	**Écart-type**	**Moyenne**	**Écart-type**	**Moyenne**	**Écart-type**	**Moyenne**	**Écart-type**
RFR	**0,00769**	0,00031	**0,08768**	0,00175	**0,03305**	**0,00047**	**0,66089**	0,01503
XGBoost	0,00801	0,00027	0,08974	0,00151	0,0364	0,00049	0,64482	0,01416
k-NN	0,01283	0,00048	0,11325	0,00211	0,0384	0,00068	0,43429	0,02330
MLP	0,00921	**0,00026**	0,09596	**0,00136**	0,04341	0,00253	0,59398	**0,01202**

TABLEAU 4.2 : Métriques des modèles après avoir appliqué la validation croisée *k-fold*.

Les résultats du tableau 4.2 révèlent que le modèle RFR a présenté la plus petite moyenne de MSE (0,00769) et MAE (0,03305), suggérant une tendance à avoir une erreur quadratique moyenne et une erreur absolue moyenne plus faibles dans ses prédictions. De plus, le RFR a obtenu la plus grande valeur moyenne de R^2 (0,66089), indiquant sa capacité à expliquer environ 66,09% de la variation des données. En revanche, le modèle MLP a présenté les plus petits écarts-types pour MSE (0,00026) et RMSE (0,00136), indiquant une moindre variabilité dans les métriques d'erreur par rapport aux autres modèles. Bien que le MLP ait une moyenne légèrement plus élevée pour MSE (0,00921) par rapport au RFR, ses plus petits écarts-types suggèrent une plus grande cohérence dans ses prédictions.

TROISIÈME APPROCHE

Dans la troisième approche, nous avons utilisé les données après leur modélisation, comme indiqué à la section 3.3.2. Nous avons appliqué la méthode de validation croisée *k-fold*, avec *k* égale à 5, sur l'ensemble des données. Nous avons entraîné et testé les modèles RFR, XGBoost, k-NN et MLP où nous avons appliqué la technique de prétraitement *OneHotEncoder* sur les données pour chaque *fold*, dans tous les modèles. Ainsi, nous avons obtenu le tableau 4.3 avec les moyennes et les écarts-types des métriques.

	MSE		**RMSE**		**MAE**		**R^2**	
	Moyenne	**Écart-type**	**Moyenne**	**Écart-type**	**Moyenne**	**Écart-type**	**Moyenne**	**Écart-type**
RFR	**0,00621**	**0,00004**	**0,07882**	**0,00028**	**0,02713**	**0,0001**	**0,6725**	0,00186
XGBoost	0,00664	0,00005	0,08147	0,00031	0,03037	0,00011	0,65011	**0,00157**
k-NN	0,01225	0,00011	0,11067	0,00049	0,03446	0,00024	0,35436	0,0041
MLP	0,00739	0,00013	0,08597	0,00077	0,03681	0,00208	0,61039	0,00622

TABLEAU 4.3 : Métriques des modèles après avoir appliqué la validation croisée *k-fold*.

Dans le tableau 4.3, nous observons une meilleure performance globale des métriques des modèles par rapport à la deuxième approche. Tous les modèles ont présenté des valeurs

plus faibles de MSE, RMSE et MAE, indiquant une tendance à avoir une erreur moyenne plus faible et une plus grande justesse dans les prévisions. Dans cet ensemble de données, le modèle RFR s'est une fois de plus démarqué, montrant la plus petite moyenne de MSE (0,00621), RMSE (0,07882) et MAE (0,02713), ainsi que la plus grande valeur moyenne de R^2 (0,6725).

Après avoir analysé les données du tableau 4.3, nous avons choisi le modèle RFR pour présenter les meilleures métriques parmi les modèles testés. Par conséquent, nous avons réajusté le modèle sur l'ensemble complet des données d'entraînement avant de réaliser toute prédiction, les résultats sont présentés dans le tableau 4.4. Comme vous pouvez constater, les résultats ne sont pas significativement différents, mais légèrement inférieurs.

MSE	**RMSE**	**MAE**	**R^2**
0,00625	0,07907	0,02707	0,66892

TABLEAU 4.4 : Métriques du modèle RFR utilisant le OneHotEncoder, entraînement avec le 2e groupe de données.

Dans le tableau 4.5, nous présentons les résultats correspondant à une formulation d'exemple, mettant en évidence les concentrations réelles par rapport aux concentrations estimées par le modèle RFR. Nous avons arrondi les concentrations estimées à trois décimales, facilitant la comparaison avec les concentrations réelles.

Acronym	Code	Concentration	
		Réelle	Estimée
SOL	040	**0,5**	0,297
HUME	002	**0,033**	0,043
ADDI	115	**0,09**	0,08
GELI	018	**0,006**	0,005
CGRA	017	**0,06**	0,028
CGRA	063	**0,017**	0,022
HUIL	156	**0,02**	0,025
EMUL	024	**0,061**	0,041
CGRA	045	**0,03**	0,015
HUIL	095	**0,041**	0,039
HUIL	004	**0,03**	0,056
HUIL	143	**0,042**	0,063
CONS	063	**0,014**	0,009
CONS	092	**0,002**	0,003
EXTR	763	**0,009**	0,019
EXTR	265	**0,027**	0,023
ACTI	223	**0,017**	0,015
PARF	345	**0,002**	0,002

TABLEAU 4.5 : Concentration réelle versus concentration estimée pour une formulation exemple.

Nous avons effectué une analyse visuelle complémentaire pour mieux comprendre la performance et l'adéquation du modèle, comme illustré dans les graphiques ci-dessous et leurs analyses respectives.

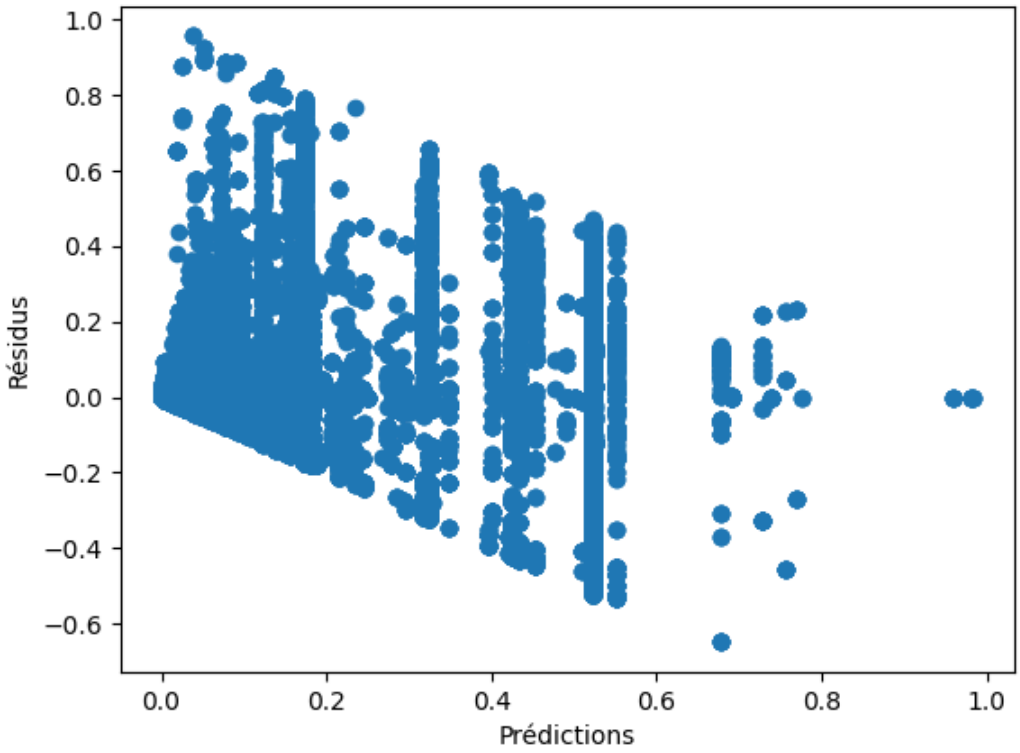


FIGURE 4.1 : Graphique de dispersion des résidus par rapport aux prédictions
©Christian Gonzalo Frantz Segovia, 2024

Dans la figure 4.1, nous observons qu'il n'y a pas de points en dessous d'une ligne diagonale imaginaire partant de (0,0 ; 0,0) jusqu'à environ (0,68 ; -0,7). Nous observons également une ligne diagonale supérieure imaginaire partant de (0,0 ; 1,0) jusqu'à environ (0,9 ; 0,0). Cette distribution suggère que les résidus sont principalement concentrés dans la plage de prédictions proche de 0 sur l'axe horizontal, ce qui indique une bonne justesse dans cet intervalle. De plus, nous remarquons une densité élevée de points près du point (0,0 ; 0,0). Ce regroupement de points suggère une concentration significative de résidus dans cette région spécifique des prédictions, indiquant une relation ou une caractéristique spéciale entre les prédictions dans cette plage et les résidus associés. De plus, le graphique présente quelques lignes verticales presque continues. Ces lignes indiquent des variations distinctes dans les résidus à des points spécifiques des prédictions, probablement liées à des motifs spécifiques dans les données ou dans le modèle.

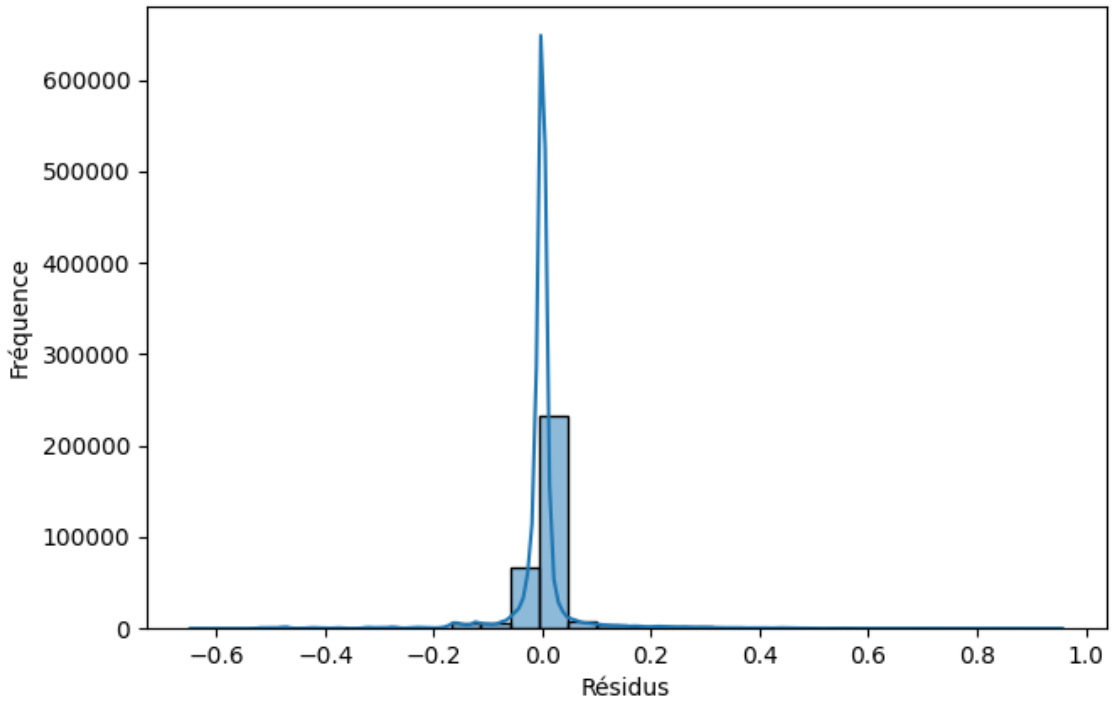


FIGURE 4.2 : Histogrammes des résidus
©Christian Gonzalo Frantz Segovia, 2024

Dans la figure 4.2, il y a une asymétrie dans la distribution des résidus par rapport à la référence zéro. Dans cet histogramme, il y a deux barres en évidence, une à gauche et une à droite de zéro, indiquant une asymétrie dans cette région. La hauteur des barres varie, mais dans tous les cas, elle est supérieure à celles aux extrémités. Cette asymétrie suggère que les résidus ne suivent pas une distribution symétrique, ce qui peut être un indice de la présence de valeurs aberrantes ou extrêmes. De plus, la présence de queues de distribution des deux côtés, avec une hauteur proche de zéro, indique une dispersion considérable dans les données.

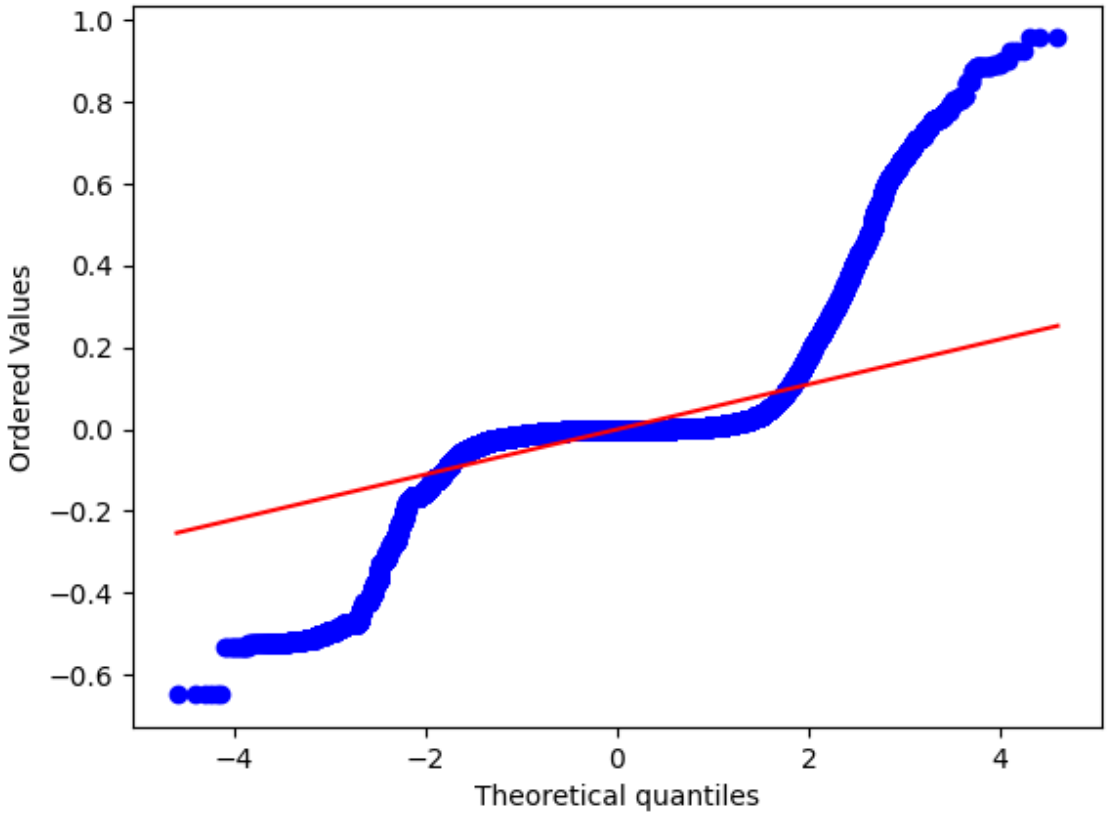


FIGURE 4.3 : Graphiques QQ (quantile-quantile) des résidus
©Christian Gonzalo Frantz Segovia, 2024

Dans la figure 4.3, nous observons un motif dans les points qui suggère un accord évident entre les résidus et la distribution théorique au point central du graphique, ce qui indique une adhérence satisfaisante à la distribution théorique dans cette région. Cependant, à mesure que l'on se rapproche des extrémités, nous observons un écart de ces résidus par rapport à ladite distribution. Cela suggère que les données ont des valeurs plus extrêmes que ce qui serait attendu si elles provenaient vraiment d'une distribution normale.

LOGICIEL DE PRÉVISION

L'IGU développée par nous consiste en un formulaire où nous pouvons ajouter des matériaux un par un. Après que les champs acronyme et code soient remplis, figure 4.4, le

système effectue une requête à notre API qui répond avec la concentration du matériau, figure 4.5.

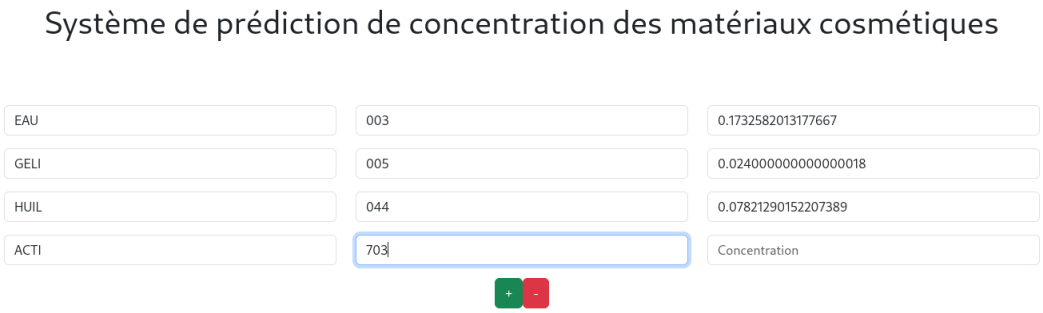


FIGURE 4.4 : Interface Graphique Utilisateur - Remplissage du formulaire
©Christian Gonzalo Frantz Segovia, 2024

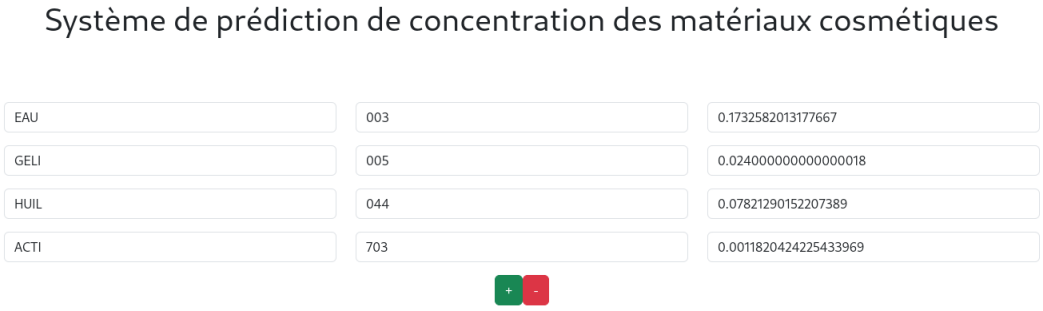


FIGURE 4.5 : Interface Graphique Utilisateur - Formulaire avec la concentration
©Christian Gonzalo Frantz Segovia, 2024

4.2 TESTANT LES EFFETS DU CHANGEMENT D'ENCODEUR

Dans cette étude, nous avons choisi de tester les effets du changement d'encodeur dans la codification de la variable catégorique acronyme. Ainsi, nous avons pu identifier quel encodeur a produit des résultats plus précis et fiables, nous aidant dans la sélection de l'approche la plus appropriée pour répondre à notre question de recherche.

	MSE		RMSE		MAE		R^2	
	Moyenne	**Écart-type**	**Moyenne**	**Écart-type**	**Moyenne**	**Écart-type**	**Moyenne**	**Écart-type**
RFR	**0,00621**	**0,00004**	**0,07882**	**0,00028**	**0,02713**	**0,0001**	**0,6725**	0,00186
XGBoost	0,00660	0,00005	0,08124	0,00033	0,02989	0,00013	0,65206	**0,00178**
k-NN	0,01213	0,00014	0,11012	0,00062	0,03418	0,00032	0,36080	0,00551
MLP	0,00764	0,00010	0,08738	0,00059	0,03764	0,00368	0,59747	0,00494

TABLEAU 4.6 : Métriques des modèles après avoir appliqué la validation croisée *k-fold* - LabelEncoder.

Lors de l'analyse des performances du modèle RFR dans l'application des deux encodeurs, nous avons remarqué qu'il a produit les mêmes résultats dans les deux cas, et ces résultats étaient supérieurs par rapport aux autres modèles évalués. Avec l'application du OneHotEncoder, tableau 4.3, le RFR a montré une MSE plus faible, une MAE plus faible et un R^2 plus élevé. Ces mesures indiquent que le modèle était capable de faire des prévisions plus précises et d'expliquer un pourcentage plus élevé de la variabilité des données de réponse par rapport aux autres modèles. De même, avec l'application du LabelEncoder, tableau 4.6, le RFR a de nouveau présenté des performances supérieures par rapport aux autres modèles, avec des valeurs plus faibles de MSE, RMSE et MAE, ainsi qu'un R^2 plus élevé. Ces résultats suggèrent que le RFR est un choix fiable pour notre modélisation prédictive, quel que soit le type d'encodeur utilisé.

4.3 DISCUSSION

Dans cette section, nous discutons de l'évaluation obtenue à partir de la mise en œuvre de l'étude de cas de la section 4.1. De plus, afin d'atteindre l'objectif défini pour ce mémoire, nous répondons à la question de recherche.

4.3.1 RÉPONSE À LA QUESTION DE RECHERCHE

QR : Est-il possible de prédire avec précision la concentration des matériaux dans la formulation des produits cosmétiques à l'aide de modèles d'apprentissage automatique ?

Sur la base des résultats présentés, des modèles d'apprentissage automatique, y compris RFR, XGBoost, k-NN et MLP, ont été développés et évalués pour prédire la concentration de matériaux dans les formulations de produits cosmétiques. Le modèle RFR a obtenu les meilleures performances globales, avec les mesures les plus basses de MSE, RMSE, MAE proches de 0 et le R^2 le plus élevé proche de 1, ce qui suggère qu'il est un choix prometteur pour cette tâche. En effet, son MSE de 0,00625 nous indique une faible magnitude moyenne d'erreurs au carré, ce qui indique une justesse correcte dans les prévisions, son RMSE de 0,07907 nous suggère que les prévisions sont proches des valeurs réelles et son MAE de 0,02707 nous indique une performance acceptable dans la minimisation des erreurs absolues moyennes. De son côté, le R^2 dont la valeur était de 0,66892 nous indique que le modèle est capable d'expliquer environ 66,89% de la variabilité des données, ce qui est une performance raisonnable. Par conséquent, sur la base des résultats de cette étude, il est possible de prédire la concentration des matériaux pour aider à la formulation de produits cosmétiques de manière adéquate en utilisant des modèles d'apprentissage automatique, à condition que le bon modèle soit choisi et que les caractéristiques des résidus soient soigneusement prises en compte lors de l'analyse.

Sur la base de l'analyse des métriques, en particulier du R^2, nous constatons que les résultats obtenus sont relativement modestes. Par conséquent, nous en déduisons que la prédiction précise des concentrations des matériaux n'est pas atteinte. Cependant, ces résultats peuvent être utilisés comme des outils auxiliaires dans la détermination des concentrations des matériaux pour les formulations cosmétiques.

Ghazal et al. [68] ont appliqué des techniques d'apprentissage automatique pour prédire des affections médicales critiques telles que le cancer, la démence et le diabète. Ils ont utilisé un ensemble de données multifactoriel lié à des troubles génétiques héréditaires, et en utilisant des techniques de SVM et de k-NN, ils ont obtenu des taux de justesse élevés lors de l'entraînement et du test. Dans notre étude, nous nous sommes concentrés sur la prédiction de la concentration des matériaux dans les formulations de produits cosmétiques en utilisant des modèles tels que RFR, XGBoost, k-NN et MLP. Bien qu'il existe des différences de contexte et de mesures d'évaluation, cette étude corrobore nos conclusions. Nous avons observé qu'à l'instar de notre étude, le modèle probabiliste k-NN a eu les moins bonnes performances par rapport aux autres modèles. Cependant, il a quand même présenté des prédictions cohérentes, ce qui confirme la réponse à notre question de recherche.

CHAPITRE V
CONCLUSION

Après l'analyse des données et la sélection du modèle RFR basée sur les résultats obtenus à travers la validation croisée *k-fold*, nous avons pu suggérer la prédiction de la concentration des matériaux dans les formulations cosmétiques. Les résultats obtenus indiquent que le modèle RFR a présenté les meilleures performances parmi les modèles testés, avec une valeur de R^2 de 0,66892, démontrant que le modèle est capable d'expliquer environ **66,89%** de la variabilité des données.

Bien que ce soit une performance raisonnable, il existe encore des limitations. Un plus grand nombre d'attributs améliorerait les performances des prédictions. De plus, cette recherche suggère qu'il existe encore des possibilités d'amélioration et de raffinement des modèles développés, comme différents ajustements de paramètres et l'extrapolation à partir d'autres modèles d'apprentissage automatique. Cependant, les résultats obtenus fournissent une bonne base pour de futures recherches et des applications pratiques dans l'industrie cosmétique. L'application et la comparaison de différents modèles d'apprentissage automatique, ainsi que l'exploration des effets de l'utilisation de différents encodeurs, ont fourni des insights pertinents pour le développement d'approches de prédiction de concentration de matériaux.

Pour les futures recherches, il est recommandé d'explorer d'autres techniques d'ingénierie des caractéristiques et de sélection de variables, ainsi que l'utilisation d'algorithmes d'apprentissage automatique plus avancés. De plus, l'incorporation de données supplémentaires et l'élargissement de la portée de l'étude à d'autres formulations cosmétiques pourraient enrichir davantage la capacité prédictive des modèles développés.

Dans un contexte pratique, les modèles développés dans cette étude ont le potentiel d'être utilisés comme des outils dans l'industrie chimique, aidant à optimiser les processus, à assister le développement de la formulation de produits et à garantir la qualité. En fournissant des prédictions de concentration de matériaux, ces modèles ont le potentiel d'avoir un impact positif sur l'efficacité opérationnelle et la prise de décisions dans les entreprises du secteur.

En résumé, cette étude représente une avancée importante vers le développement de modèles prédictifs pour l'industrie chimique, soulignant l'importance de l'application de techniques avancées d'apprentissage automatique et de validation croisée pour résoudre des problèmes complexes et difficiles dans ce domaine.

BIBLIOGRAPHIE

[1] D. Petruzzi, “Global skin care market size 2012-2025,” 2022. [En ligne]. Repéré à : https://www.statista.com/statistics/254612/global-skin-care-market-size/

[2] H. Chamberland, *Cosmétiques et soins personnels - Profil industriel / ce rapport a été réalisé par le Ministère de l’économie [...]*. Québec : Ministère de l’économie, de l’innovation et des exportations, 2014. [En ligne]. Repéré à : https://numerique.banq.qc.ca/patrimoine/details/52327/2406177

[3] Y. Pena. (2022) Canada cosmetics and beauty products market. En ligne ; consulté le 15 décembre 2023. [En ligne]. Repéré à : https://www.trade.gov/market-intelligence/canada-cosmetics-and-beauty-products-market

[4] A. Ainurofiq, A. Maharani, F. Fatonah, H. N. Halida, et T. Nurrodlotiningtyas, “Pre-formulation study on the preparation of skin cosmetics,” *Science and Technology Indonesia*, vol. 6, n° 4, p. 273 – 284, 2021.

[5] S. Magdassi et E. Touitou, *Cosmeceutics and Delivery Systems*. CRC Press, 2023.

[6] V. B. Martins, J. Bordim, G. A. P. Bom, J. G. D. S. Carvalho, C. R. B. Parabocz, et M. L. Mitterer Daltoé, “Consumer profiling techniques for cosmetic formulation definition,” *Journal of Sensory Studies*, vol. 35, n° 2, 2020.

[7] F. Calvo, J. M. Gómez, L. Ricardez-Sandoval, et O. Alvarez, “Integrated design of emulsified cosmetic products : A review,” *Chemical Engineering Research and Design*, vol. 161, p. 279 – 303, 2020.

[8] A. S. Abouzied, S. M. Alshahrani, U. Hani, A. J. Obaidullah, A. A. Al Awadh, A. A. Lahiq, et H. J. Al-fanhrawi, “Assessment of solid-dosage drug nanonization by theoretical advanced models : Modeling of solubility variations using hybrid machine learning models,” *Case Studies in Thermal Engineering*, vol. 47, 2023.

[9] Y. Nakai, A. Noda, et E. Yamamoto, “Algorithm for the early prediction of drug stability using bayesian inference and multiple measurements : Application for predicting the stability of silodosin tablets,” *Journal of pharmaceutical and biomedical analysis*, vol. 233, p. 115442, 2023. [En ligne]. Repéré à : https://doi.org/10.1016/j.jpba.2023.115442

[10] D. M. West et J. R. Allen. (2018) How artificial intelligence is transforming the world. En ligne; consulté le 28 novembre 2023. [En ligne]. Repéré à : https://www.brookings.edu/articles/how-artificial-intelligence-is-transforming-the-world/

[11] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu, J. Qiu, K. Hua, W. Su, J. Wu, H. Xu, Y. Han, C. Fu, Z. Yin, M. Liu, R. Roepman, S. Dietmann, M. Virta, F. Kengara, Z. Zhang, L. Zhang, T. Zhao, J. Dai, J. Yang, L. Lan, M. Luo, Z. Liu, T. An, B. Zhang, X. He, S. Cong, X. Liu, W. Zhang, J. P. Lewis, J. M. Tiedje, Q. Wang, Z. An, F. Wang, L. Zhang, T. Huang, C. Lu, Z. Cai, F. Wang, et J. Zhang, "Artificial intelligence : A powerful paradigm for scientific research," *The Innovation*, vol. 2, nº 4, p. 100179, 2021. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/S2666675821001041

[12] J. Zhao, "Estimation of inorganic crystal densities using gradient boosted trees," *Frontiers in Materials*, vol. 9, 2022.

[13] D. D. Solomon, Sonia, K. Kumar, K. Kanwar, S. Iyer, et M. Kumar, "Extensive review on the role of machine learning for multifactorial genetic disorders prediction," *Archives of Computational Methods in Engineering*, 2023.

[14] T. P. Nagarhalli, S. Mhatre, S. Patil, et P. Patil, "The review of natural language processing applications with emphasis on machine learning implementations," 2022, Conference paper, p. 1353 – 1358.

[15] P. Pant et P. Srivastava, "Cost-sensitive model evaluation approach for financial fraud detection system," 2021, Conference paper, p. 1606 – 1611.

[16] S. Soriano-Meseguer, E. Fuguet, A. Port, et M. Rosés, "Suitability of skin-pampa and chromatographic systems to emulate skin permeation. influence of ph on skin-pampa permeability," *Microchemical Journal*, vol. 190, 2023.

[17] P. Das et M. K. Das, *Physical, chemical, and microbiological stability of nanocosmetics*. Academic Press, 2022.

[18] N. Kamaruzaman et S. M. Yusop, "Determination of stability of cosmetic formulations incorporated with water-soluble elastin isolated from poultry," *Journal of King Saud University - Science*, vol. 33, nº 6, p. 101519, 2021. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/S1018364721001804

[19] L. H. Do, R. M. Law, et H. I. Maibach, “Dose response effect of chemical surface concentration on percutaneous penetration in human : In vivo + in vitro,” *Regulatory Toxicology and Pharmacology*, vol. 132, 2022.

[20] J. Daneluz, J. d. S. Favero, V. d. Santos, V. Weiss-Angeli, L. B. Gomes, A. S. Mexias, et C. P. Bergmann, “The influence of different concentrations of a natural clay material as active principle in cosmetic formulations,” *Materials Research*, vol. 23, n° 2, p. e20190572, 2020. [En ligne]. Repéré à : https://doi.org/10.1590/1980-5373-MR-2019-0572

[21] S. Russell et P. Norvig, *Artificial Intelligence : A Modern Approach*, 3e éd. USA : Prentice Hall Press, 2009.

[22] T. W. Edgar et D. O. Manz, “Chapter 6 - machine learning,” dans *Research Methods for Cyber Security*, T. W. Edgar et D. O. Manz, éds. Syngress, 2017, pp. 153–173. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/B9780128053492000066

[23] R. S. Sutton et A. G. Barto, *Reinforcement Learning : An Introduction*. Cambridge, MA, USA : A Bradford Book, 2018.

[24] I. H. Sarker, “Machine learning : Algorithms, real-world applications and research directions,” *SN Computer Science*, vol. 2, n° 3, p. 160, March 22 2021. [En ligne]. Repéré à : https://doi.org/10.1007/s42979-021-00592-x

[25] M. Golmohammadi et M. Aryanpour, “Analysis and evaluation of machine learning applications in materials design and discovery,” *Materials Today Communications*, vol. 35, 2023.

[26] C. J. Kuster, J. Baumann, S. M. Braun, P. Fisher, N. J. Hewitt, M. Beck, F. Weysser, L. Goerlitz, P. Salminen, C. R. Dietrich, M. Wang, et M. Ernst, “In silico prediction of dermal absorption from non-dietary exposure to plant protection products,” *Computational Toxicology*, vol. 24, 2022.

[27] H. Ye, Z. Du, H. Lu, J. Tian, L. Chen, et W. Lin, “Using machine learning methods to predict voc emissions in chemical production with hourly process parameters,” *Journal of Cleaner Production*, vol. 369, 2022, cited by : 1.

[28] B. Huwaimel, T. Nafea Alharby, J. Alanazi, et M. Alanazi, “Computational estimation of drug’s concentration distribution through a microporous membrane using artificial intelligence approach,” *Journal of Molecular Liquids*, vol. 380, 2023.

[29] J. Fine, P. R. Wijewardhane, S. D. B. Mohideen, K. Smith, J. R. Bothe, Y. Krishnamachari, A. Andrews, Y. Liu, et G. Chopra, “Learning relationships between chemical and physical stability for peptide drug development,” *Pharmaceutical Research*, vol. 40, n° 3, p. 701 – 710, 2023.

[30] D. Chicco, M. J. Warrens, et G. Jurman, “The coefficient of determination r-squared is more informative than smape, mae, mape, mse, and rmse in regression analysis evaluation,” *PeerJ Comput Sci*, vol. 7, p. e623, 2021, pMID : 34307865. [En ligne]. Repéré à : https://doi.org/10.7717/peerj-cs.623

[31] C. J. Willmott et K. Matsuura, “Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance,” *Climate Research*, vol. 30, n° 1, p. 79 – 82, 2005, cited by : 3008 ; All Open Access, Bronze Open Access. [En ligne]. Repéré à : https://www.scopus.com/inward/record.uri?eid=2-s2.0-30444437204&doi=10.3354%2fcr030079&partnerID=40&md5=cec06ec7e7fca18c27a3f0aad3b94b61

[32] E. Kasuya, “On the use of r and r squared in correlation and regression,” *Ecological Research*, vol. 34, n° 1, pp. 235–236, 2019. [En ligne]. Repéré à : https://esj-journals.onlinelibrary.wiley.com/doi/abs/10.1111/1440-1703.1011

[33] M. Imran et A. Akbar, “Diagnostics via partial residual plots in inverse gaussian regression,” *Journal of Chemometrics*, vol. 34, n° 1, 2020.

[34] S. Azman et D. Pathmanathan, “The glm framework of the lee–carter model : a multi-country study,” *Journal of Applied Statistics*, vol. 49, n° 3, p. 752 – 763, 2022. [En ligne]. Repéré à : https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092524047&doi=10.1080%2f02664763.2020.1833183&partnerID=40&md5=fbc6511c505924f9715da38e76030cff

[35] G. Sannino, P. Melillo, S. Stranges, G. De Pietro, et L. Pecchia, “Short term heart rate variability to predict blood pressure drops due to standing : A pilot study,” *BMC medical informatics and decision making*, vol. 15, p. S2, 09 2015.

[36] X. Zhao, Y. Zhang, S. Xie, Q. Qin, S. Wu, et B. Luo, “Outlier detection based on residual histogram preference for geometric multi-model fitting,” *Sensors (Basel)*, vol. 20, n° 11, p. 3037, May 2020.

[37] M. García Ben et V. J. Yohai, “Quantile-quantile plot for deviance residuals in the generalized linear model,” *Journal of Computational and Graphical Statistics*, vol. 13, n° 1, p. 36 – 47, 2004, cited by : 31. [En ligne]. Repéré à : https://www.scopus.com/inward/record.uri?eid=2-s2.0-85011468615&doi=10.1198%2f1061860042949_a&partnerID=40&md5=8743bc07b10df77e33eb1075dc00670a

[38] E. Weine, M. S. McPeek, et M. Abney, “Application of equal local levels to improve q-q plot testing bands with r package qqconf,” *Journal of Statistical Software*, vol. 106, 2023.

[39] T.-T. Wong et P.-Y. Yeh, “Reliable accuracy estimates from k-fold cross validation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, n° 8, p. 1586 – 1594, 2020, cited by : 324. [En ligne]. Repéré à : https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088149663&doi=10.1109%2fTKDE.2019.2912815&partnerID=40&md5=ca364e716d7467dc172b6432e280ae0e

[40] M. Lukic, I. Pantelic, et S. Savic, “A comparison of myribase and doublebase gel : Does qualitative similarity of emollient products imply their direct interchangeability in everyday practice ?” *Dermatologic Therapy*, vol. 33, n° 6, 2020, cited by : 3 ; All Open Access, Gold Open Access, Green Open Access. [En ligne]. Repéré à : https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088805155&doi=10.1111%2fdth.14020&partnerID=40&md5=6d53f6bc8eb74c1120063e8a42bc4522

[41] Y. Vinetsky et S. Magdassi, *Microcapsules in Cosmetics*, 2023.

[42] A. Feliczak-Guzik et I. Nowak, *Analysis of cosmetic products in biological matrices*, 2022, cited by : 0. [En ligne]. Repéré à : https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159598194&doi=10.1007%2f978-3-030-95660-8_46&partnerID=40&md5=b998ed09bf748f7ae44b48fc8425c86b

[43] E. Mérat, A. Roso, M. Dumaine, et S. Sigurani, *Sensory evaluation of cosmetic functional ingredients*, 2021.

[44] R. Pratiwi, N. N. Auliya As, R. F. Yusar, et A. A. A. Shofwan, “Analysis of prohibited and restricted ingredients in cosmetics,” *Cosmetics*, vol. 9, n° 4, 2022.

[45] A. Papadopoulos, N. Assimomytis, et A. Varvaresou, “Sample preparation of cosmetic products for the determination of heavy metals,” *Cosmetics*, vol. 9, n° 1, 2022.

[46] K. G. Robinson et R. E. Akins, “Chapter 24 - machine learning in epigenetic diseases,” dans *Medical Epigenetics (Second Edition)*, ser. Translational Epigenetics, T. O. Tollefsbol, éd. Academic Press, 2021, vol. 29, pp. 513–525. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/B9780128239285000384

[47] J. A. Nichols, H. W. Herbert Chan, et M. A. B. Baker, “Machine learning : Applications of artificial intelligence to imaging and diagnosis,” *Biophys Rev*, vol. 11, n° 1, pp. 111–118, Feb 2019, this article does not contain any studies with human participants or animals performed by any of the authors.

[48] O. Gorodetskaya, Y. Gobareva, et M. Koroteev, “A machine learning pipeline for forecasting time series in the banking sector,” *Economies*, vol. 9, n° 4, 2021.

[49] L. Vanneschi et S. Silva, *Decision Tree Learning*. Cham : Springer International Publishing, 2023, pp. 149–159. [En ligne]. Repéré à : https://doi.org/10.1007/978-3-031-17922-8_6

[50] M. Jena et S. Dehuri, “Decision tree for classification and regression : A state-of-the-art review,” *Informatica SI*, vol. 44, n° 4, pp. 405—-420, 2020.

[51] I. E. Genrikhov, E. V. Djukova, et V. I. Zhuravlev, “On full regression decision trees,” *Pattern Recognition and Image Analysis*, vol. 27, pp. 1–7, 2017. [En ligne]. Repéré à : https://doi.org/10.1134/S1054661817010047

[52] M. Czajkowski et M. Kretowski, “The role of decision tree representation in regression problems – an evolutionary perspective,” *Applied Soft Computing*, vol. 48, pp. 458–475, 2016. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/S1568494616303325

[53] R. Couronné, P. Probst, et A.-L. Boulesteix, “Random forest versus logistic regression : a large-scale benchmark experiment,” *BMC Bioinformatics*, vol. 19, n° 1, p. 270, 2018. [En ligne]. Repéré à : https://doi.org/10.1186/s12859-018-2264-5

[54] A. Kadiyala et A. Kumar, “Applications of python to evaluate the performance of decision

tree-based boosting algorithms," *Environmental Progress and Sustainable Energy*, vol. 37, n° 2, p. 618 – 623, 2018.

[55] A. Almaafi, S. Bajaba, et F. Alnori, "Stock price prediction using arima versus xgboost models : the case of the largest telecommunication company in the middle east," *International Journal of Information Technology (Singapore)*, vol. 15, n° 4, p. 1813 – 1818, 2023.

[56] Q. Gong, H. Huang, et B. Zhang, "Grain yield prediction model based on the analysis of climate and irrigated area conditions in the wheat grain-filling period," *Irrigation and Drainage*, vol. 72, n° 2, p. 422 – 438, 2023.

[57] A. Galich, S. Nieland, B. Lenz, et J. Blechschmidt, "How would we cycle today if we had the weather of tomorrow ? an analysis of the impact of climate change on bicycle traffic," *Sustainability (Switzerland)*, vol. 13, n° 18, 2021.

[58] M. M. Hassan, M. M. Hassan, F. Yasmin, M. A. R. Khan, S. Zaman, Galibuzzaman, K. K. Islam, et A. K. Bairagi, "A comparative assessment of machine learning algorithms with the least absolute shrinkage and selection operator for breast cancer detection and prediction," *Decision Analytics Journal*, vol. 7, p. 100245, 2023. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/S2772662223000851

[59] F. M. Bayat, M. Prezioso, B. Chakrabarti, H. Nili, I. Kataeva, et D. Strukov, "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits," *Nature Communications*, vol. 9, n° 1, p. 2331, 2018. [En ligne]. Repéré à : https://doi.org/10.1038/s41467-018-04482-4

[60] S. S. Chai, W. L. Cheah, K. L. Goh, Y. H. R. Chang, K. Y. Sim, et K. O. Chin, "A multilayer perceptron neural network model to classify hypertension in adolescents using anthropometric measurements : A cross-sectional study in sarawak, malaysia," *Computational and mathematical methods in medicine*, vol. 2021, p. 2794888, 2021. [En ligne]. Repéré à : https://doi.org/10.1155/2021/2794888

[61] A. Botchkarev, "A new typology design of performance metrics to measure errors in machine learning regression algorithms," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, p. 45 – 76, 2019.

[62] S. Wright, "Correlation and causation," *Journal of Agricultural Research*, vol. XX, n° 7,

pp. 557–585, 1921.

[63] R. Rani, S. Kumar, R. Patil, et S. Pippal, “A machine learning model for predicting innovation effort of firms,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, p. 4633, 08 2023.

[64] L. L. Thornton, D. E. Carlson, et M. R. Wiesner, “Predicting emerging chemical content in consumer products using machine learning,” *Science of The Total Environment*, vol. 834, p. 154849, 2022. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/S0048969722019428

[65] M. Akrom, S. Rustad, et H. Kresno Dipojono, “Machine learning investigation to predict corrosion inhibition capacity of new amino acid compounds as corrosion inhibitors,” *Results in Chemistry*, vol. 6, p. 101126, 2023. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/S221171562300365X

[66] J. Shim, J. M. Lim, et S. G. Park, “Machine learning for the prediction of sunscreen sun protection factor and protection grade of uva,” *Experimental Dermatology*, vol. 28, n° 7, pp. 872–874, 2019. [En ligne]. Repéré à : https://onlinelibrary.wiley.com/doi/abs/10.1111/exd.13958

[67] B. F. Huang et P. C. Boutros, “The parameter sensitivity of random forests,” *BMC Bioinformatics*, vol. 17, n° 1, p. 331, 09 2016. [En ligne]. Repéré à : https://doi.org/10.1186/s12859-016-1228-x

[68] T. M. Ghazal, H. Al Hamadi, M. Umar Nasir, Atta-Ur-Rahman, M. Gollapalli, M. Zubair, M. Adnan Khan, et C. Yeob Yeun, “Supervised machine learning empowered multifactorial genetic inheritance disorder prediction,” *Computational intelligence and neuroscience*, vol. 2022, p. 1051388, 2022.

[69] A. S. Abouzied, S. M. Alshahrani, U. Hani, A. J. Obaidullah, A. A. Al Awadh, A. A. Lahiq, et H. J. Al-fanhrawi, “Assessment of solid-dosage drug nanonization by theoretical advanced models : Modeling of solubility variations using hybrid machine learning models,” vol. 47, p. 103101, 2023. [En ligne]. Repéré à : https://www.sciencedirect.com/science/article/pii/S2214157X23004070

[70] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, et O. A. von Lilienfeld, “Machine learning of molecular electronic

properties in chemical compound space," *New Journal of Physics*, vol. 15, nº 9, p. 095003, sep 2013. [En ligne]. Repéré à : https://dx.doi.org/10.1088/1367-2630/15/9/095003

[71] B. Atakan, T. Y. Pirinçci, M. Burcu, Ö. Yıldız, et U. Tamer, "Estimating the optimal dexketoprofen pharmaceutical formulation with machine learning methods and statistical approaches," *Healthc Inform Res*, vol. 27, nº 4, pp. 279–286, 2021. [En ligne]. Repéré à : http://e-hir.org/journal/view.php?number=1088

[72] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, et É. Duchesnay. (2011) scikit-learn : Machine learning in Python. En ligne ; consulté le 28 novembre 2023. [En ligne]. Repéré à : https://scikit-learn.org/stable/

[73] V. K. Mishra, R. Binyala, P. Sharma, et S. Singh, *Predictive Analysis of Stock Prices Through Scikit-Learn*. John Wiley and Sons, Ltd, 2023, ch. 20, pp. 397–404. [En ligne]. Repéré à : https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119879831.ch20

APPENDICE A

STRUCTURE DE FICHIERS

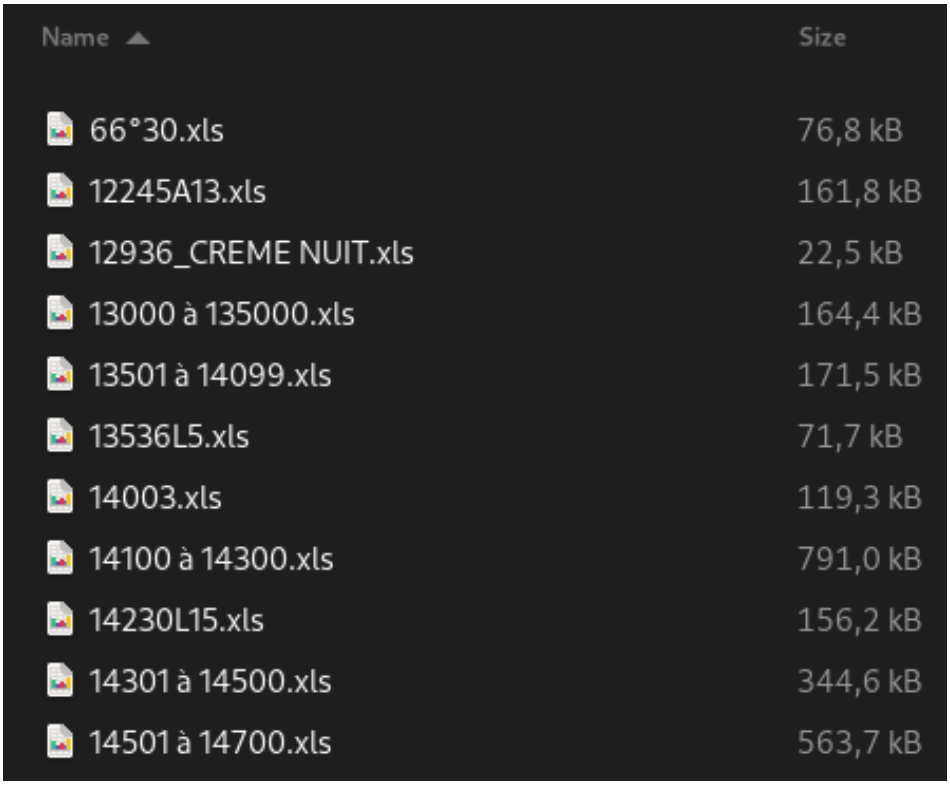


FIGURE A.1 : Structure de fichiers

APPENDICE B

NOMBRE TOTAL ET TAILLE DES FICHIERS

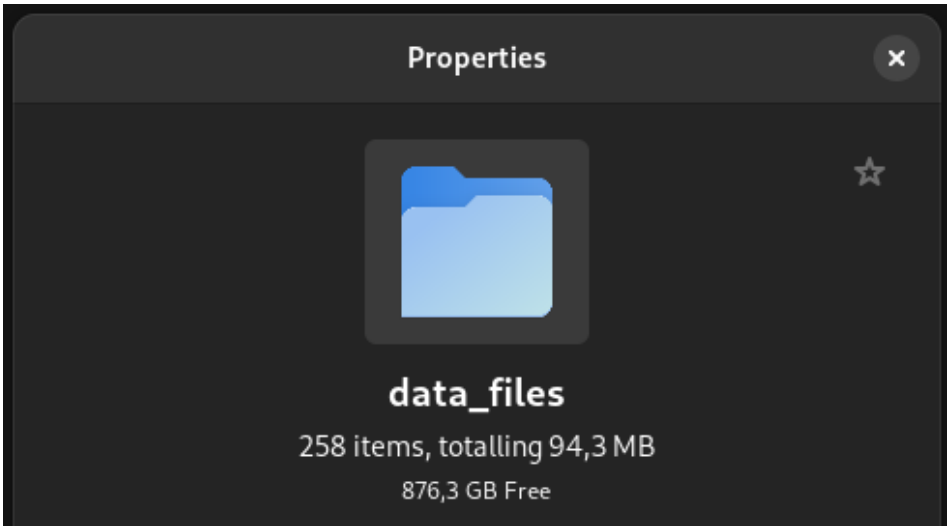


FIGURE B.1 : Nombre total et taille des fichiers

APPENDICE C

SOMME DES CONCENTRATIONS

Code	Nom	Pour 1 99.47%
PHASE I	**PHASE I**	
MTEAU007	EAU OSMOSEE CHAUDE 60°C	76.50%
MTACTI505	ACIDE KOJIQUE	4.60%
MTACTI278	ARBUTINE	1.00%
MTGELI015	SEPIGEL 305	1.30%
PHASE II	**PHASE II**	
MTSILI009	SILICONE DC 2501 WAX	3.00%
MTCONS084	ECO DERMOSOFT GMCY	2.00%
LaboMTEXTR795	GLYCOLYSAT REGLISSE UP	10.00%
MTHUME018	SENSIVA SC 50	0.20%
PHASE III	**PHASE III**	
MTABC106	LICORICE PTH / MARUZEN	0.31%
MTHUME017	BUTYLENE GLYCOL-1,3	0.56%

FIGURE C.1 : Somme des concentrations inférieures à 100%

Code	Nom	Pour 1 100.11%
PHASE I		
MTEAU008	EAU OSMOSEE TIEDE 50°C	82.21%
MTACTI278	ARBUTINE	3.33%
MTACTI543	ACTIWHITE PW LS 9860	2.00%
PHASE IA		
MTGELI018	ECO KELTROL CG-BT	2.80%
MTABC001	HISPAGEL 100	1.00%
MTHUME002	ECO GLYCERINE CODEX VEGETALE 99,5 %	4.00%
PHASE II		
MTABC015	BUTYLENE GLYCOL 1-3 0000254412 / UNIVAR (puis passer sur MTHUME(	1.69%
MTEXTR324	VIAPURE LICORICE WHITE 97	1.00%
PHASE III		
MTHUME016	HYDROLITE 5	1.60%
MTCONS037	DEKABEN CP (= ELESTAB CPN)	0.15%
MTCONS023	BIOSOL	0.20%
MTCONS064	DERMOSOFT OCTIOL	0.13%

FIGURE C.2 : Somme des concentrations supérieures à 100%

APPENDICE D

INCOHÉRENCE DES DONNÉES

Code	Nom	Pour 1 100.00%
PHASE I		
MTEAU004	EAU OSMOSEE CHAUDE 80°C	36.35%
MTACTI031	ECO HYALURONATE SODIUM (CRISTALHYAL)	0.50%
MTABC112	NATROSOL 250 HHX PHARM / AQUALON	1.00%
PHASE II		
MTMOUS009	MASSOCARE T 20 / TWEEN 20	0.50%
MTABC179	PARFUM ROSE SOPHIA (ADVANCED FRAGRANCE CREATION)	0.05%
PHASE III		
MTCONS037	DEKABEN CP (= ELESTAB CPN)	0.30%
	BIOSOL	0.10%
MTCONS064	DERMOSOFT OCTIOL	0.20%
MTHUME016	HYDROLITE 5	1.00%
PHASE IV		
MTACTI834	ICHTYOCOLLAGENE PH	60.00%

FIGURE D.1 : Tableau de formulation sans le code du produit

Code	Nom	Pour 1 100.00%
PHASE I		
MTEAU008	EAU OSMOSEE TIEDE 50°C	82.81%
MTACTI278	ARBUTINE	2.00%
MTACTI626	LIPIDURE PMB PH 10	4.00%
MTGELI015	SEPIGEL 305	1.00%
PHASE II		
MTABC015	BUTYLENE GLYCOL 1-3 0000254412 / UNIVAR (puis passer sur MTHUME0	0.49%
MTEXTR324	VIAPURE LICORICE WHITE 97	0.50%
PHASE III		
MTACTI469	LUMISKIN (TM)	3.00%
MTSILI009	SILICONE DC 2501 WAX	3.00%
PHASE IV		
MTMOUS009	MASSOCARE T 20 / TWEEN 20	0.30%
LaboMTPARF377	PARFUM TIFFANY Y / PARFEX	0.30%
PHASE V		
MTHUME016	HYDROLITE 5	1.00%
MTCONS037	DEKABEN CP (= ELESTAB CPN)	0.55%
LaboMTCONS023	BIOSOL	0.60%
MTCONS064	DERMOSOFT OCTIOL	0.45%

FIGURE D.2 : Tableau de formulation avec code produit avec informations complémentaires

Code	Nom	Pour 1 100.00%
PHASE I		
MTEAU005	EAU OSMOSEE CHAUDE 75°C	74.15%
MTADDI003	DISSOLVINE NA2 P (=EDTA BD)	1.55%
MTHUME002	ECO GLYCERINE CODEX VEGETALE 99,5 %	3.50%
MTGELI080	ARISTOFLEX AVC	1.60%
PHASE II		
MTABC012	ESTASAN GT 860 3575 / QUIMASSO (puis passer sur MTHUIL018)	6.00%
MTACTI469	LUMISKIN (TM)	5.00%
AMTHUIL190	CRODAMOL AB-LQ-(RB) HB03913 / CRODA	1.00%
MTSILI048	GRANSIL GCM-5 / IMCD	0.50%
PHASE III		
MTACTI543	ACTIWHITE PW LS 9860	2.50%
PHASE IV		
MTHUME017	BUTYLENE GLYCOL-1,3	1.30%
MTEXTR324	VIAPURE LICORICE WHITE 97	0.50%
PHASE V		
MTPARF382	PARFUM TFG 1410-Rosée-Irisée-Epicée	1.10%
PHASE VI		
MTHUME016	HYDROLITE 5	0.50%
MTCONS037	DEKABEN CP (= ELESTAB CPN)	0.35%
MTCONS023	BIOSOL	0.20%
MTCONS064	DERMOSOFT OCTIOL	0.25%

FIGURE D.3 : Tableau de formulation avec un code de produit incorrect
©Christian Gonzalo Frantz Segovia, 2023

APPENDICE E

INCOHÉRENCE DES TABLEAUX

LaboF		
PHASE I	PHASE I	
MTEAU004	EAU OSMOSEE CHAUDE 80°C	53.75%
MTHUME002	GLYCERINE CODEX VEGETALE 99,5 %	5.00%
MTGELI018	KELTROL CG-BT	1.40%
LaboMTEXTR091	distillat bio rose de damascena 1:20	7.00%
MTEMUL076	BIOPHILIC H PCR NEGATIF	3.00%
PHASE II	PHASE II	
MTCGRA581	CETIOL CC	6.00%
MTCGRA014	LANETTE O OR	1.50%
MTCGRA006	CUTINA GMS-V	2.50%
MTCGRA004	CERABEIL BLANCHE SELECTION PASTILLES / CIRE ABEILLE	3.00%
LaboMTCGRA010	BEURRE DE CACAO DESODORISE BIO	3.00%
LaboMTCGRA011	BEURRE DE KARITE BIO	1.50%
MTHUIL051	CETIOL LC	1.00%
MTHUIL017	CRODAMOL IPM (myristate isopropyle)	3.00%
PHASE III	PHASE III	
LaboMTCONS075	GEOGARD 221	3.00%
LaboMTACTI782	FIBER BOOSTER GINGKO	2.50%
LaboMTACTI047	LIFTISS EXTRAIT SEC SCX	1.30%
MTACTI438	CIRE BLE NOIR	1.55%
	TOTAL	100.00%

FIGURE E.1 : Tableau sans l'en-tête correct

Code	Nom	Pour 1 100.00%
PHASE I		
MTEAU009	EAU OSMOSEE TIEDE 40°C	53.49%
MTACTI270	ECO SODIUM BENZOATE	4.45%
MTADDI001	ECO ACIDE CITRIQUE MONOHYDRATE GRANULE IP	5.40%
MTGELI018	ECO KELTROL CG-BT	1.80%
MTEXTR002	ECO ALOE VERA GEL POWDER QM 200X ORGANIC CERTIFIED PLUS	4.11%
PHASE II		
MTCGRA014	LANETTE O OR	1.00%
PHASE III		
MTMOUS080	ECO SULFETAL LAB-E	12.50%
LaboMTMOUS035	ECO AMPHOTENSID GB 2009	9.00%
MTMOUS035	ECO PLANTACARE 818 UP	6.00%
MTMOUS039	ECO LAMESOFT PO 65	0.25%
MTMOUS034	ECO ORAMIX CG 110	0.50%
PHASE IV		
MTCONS069	ECO DERMOSOFT 1388 / ECO	1.50%

Code	Nom	Pour 1 100.00%
	PHASE I	
MTEAU005	EAU OSMOSEE CHAUDE 75°C	39.80%
LaboMTEMUL145	SALACOS GE-318	12.00%
LaboMTEMUL146	SALACOS GE-118	12.00%
	DIPROPYLENE GLYCOL (INTERCHIMIE)	31.00%
MTADDI033	TRI SODIUM CITRATE CRISTALLISE DIHYDRATE CRYST. / EMPROVE	1.07%
MTADDI001	ECO ACIDE CITRIQUE MONOHYDRATE GRANULE IP	2.53%
MTPARF104	PARFUM ORANGE VERTE K66-1165	1.60%

FIGURE E.2 : Multiples tableaux dans une feuille de calcul
©Christian Gonzalo Frantz Segovia, 2023

Code	Nom	Qt Th	Pour 1
		700	100.00%
PHASE I			
MTEAU004	EAU OSMOSEE CHAUDE 80°C	188.3	26.90%
MTACTI218	ECO LIPACIDE C8G	14	2.00%
PHASE II			
MTADDI098	ECO SODIUM HYDROXYDE / SOUDE PHARMA	8.4	1.20%
PHASE III			
MTADDI011	PURAC HiPure 90 / ACIDE LACTIQUE	13.09	1.87%
PHASE IV			
MTEAU009	EAU OSMOSEE TIEDE 40°C	257.46	36.78%
MTACTI219	ACNET	8.4	1.20%
PHASE V			
MTMOUS029	ORONAL LCG	70	10.00%
MTMOUS033	MIRANOL C2M CONC NP	98	14.00%
MTPARF399	PARFUM LUNA	15.4	2.20%
PHASE VI			
MTCONS010	PHENONIP	18.2	2.60%
MTCONS037	DEKABEN CP (= ELESTAB CPN)	8.75	1.25%

FIGURE E.3 : Tableau avec un ordre de colonnes différent

	Code	Nom	Pour 1 100.00%
	PHASE I		
1	MTEAU003	EAU OSMOSEE FROIDE	35.42%
2	MTADDI014	SEL FIN EPURE SECHE DESULFATE	0.50%
3	MTHUME005	MONOPROPYLENE GLYCOL USP/EP	12.00%
4	MTCONS012	POB METHYLE / NIPAGIN M	1.10%
5	MTCONS007	GERMALL 115	3.15%
6	MTCONS009	KATHON CG	0.52%
7	MTEXTR004	AROMATIQUE BLEUET EDC 398	2.20%
8	MESOL009	SOL. 0,5% COLORANT BLUE N°1 W 092	3.51%
	PHASE II		
9	MTSILI007	HUILE SILICONE VOLATILE DC 345 FLUIDE	17.00%
10	MTHUIL019	ECO ISOPROPYLPALMITAT / PALMITATE D'ISOPROPYLE	9.00%
11	MTHUIL018	ECO MASSOCARE MCT (ex ESTASAN GT 8-60 3575) / TRIGLYCERID	9.00%
12	MTHUIL013	HUILE PARAFFINE WHITOL CODEX FLUIDE 15	6.10%
13	MTCOLO019	COLORANT VERT AU GRAS W 7200	0.5005%

FIGURE E.4 : Tableau à n'importe quelle position dans la feuille de calcul
©Christian Gonzalo Frantz Segovia, 2023

		100.00%
	PHASE I	
MTEAU004	EAU OSMOSEE CHAUDE 80°C	70.50%
MTGELI017	CARBOPOL ULTREZ 10	1.80%
MTGELI032	JAGUAR HP 105	1.00%
	PHASE IA	
MTADDI017	TRIETHANOLAMINE CODEX 99 % - INEOS	2.55%
	PHASE II	
MTHUME002	ECO GLYCERINE CODEX VEGETALE 99,5 %	1.00%
MTADDI003	DISSOLVINE NA2 P (=EDTA BD)	1.20%
MTACTI003	ALLANTOINE	1.20%
MTHUME005	MONOPROPYLENE GLYCOL USP/EP	1.50%
	PHASE III	
MTCONS009	KATHON CG	1.05%
MTMOUS500	COMPERLAN KD	2.00%
MTCONS010	PHENONIP	1.50%
MTMOUS074	POLYSORBATE 20 A épuisement revenir sur MTMOUS009	2.00%
MTMOUS508	GEROPON TC42	7.00%
MTMOUS513	MIRATAINE CBS	4.00%
MTPARF543	PARFUM MANDARINE 10688/B	1.70%

FIGURE E.5 : Tableau sans l'en-tête
©Christian Gonzalo Frantz Segovia, 2023

APPENDICE F

ROTATION DES MATÉRIAUX AVEC LES VRAIS DONNÉES

Acronym	Code	Concentration estimée
EAU	005	0,498
EMUL	145	0,1
EMUL	146	0,1
HUME	005	0,15
HUME	024	0,15
ADDI	033	0,0007
ADDI	001	0,0003
PARF	104	0,001

TABLEAU F.1 : Exemple de rotation des matériaux - Première itération

Acronym	Code	Concentration estimée
EMUL	145	0,1
EMUL	146	0,1
HUME	005	0,15
HUME	024	0,15
ADDI	033	0,0007
ADDI	001	0,0003
PARF	104	0,001
EAU	005	0,498

TABLEAU F.2 : Exemple de rotation des matériaux - Deuxième itération

Acronym	Code	Concentration estimée
EMUL	146	0,1
HUME	005	0,15
HUME	024	0,15
ADDI	033	0,0007
ADDI	001	0,0003
PARF	104	0,001
EAU	005	0,498
EMUL	145	0,1

TABLEAU F.3 : Exemple de rotation des matériaux - Troisième itération

Acronym	Code	Concentration estimée
PARF	104	0,001
EAU	005	0,498
EMUL	145	0,1
EMUL	146	0,1
HUME	005	0,15
HUME	024	0,15
ADDI	033	0,0007
ADDI	001	0,0003

TABLEAU F.4 : Exemple de rotation des matériaux - Dernière itération