**Titre** : Apprentissage par la machine et big data cosmétique en vue de la modélisation mathématique et analyse statistique prédictive sur l'efficacité de formules cosmétiques

**Résumé** : L’industrie cosmétique est confrontée à une concurrence accrue et aux attentes d’une clientèle exigeantes. Elle doit proposer des solutions innovantes toujours plus efficaces, satisfaisantes d’un point de vue sensoriel tout en respectant les exigences règlementaires. Différents tests sont mis en place pour soutenir les allégations marketing et prouver l’innocuité des produits mis sur le marché. L’emploi de méthodes d’apprentissage automatique peut être utilisée comme solution de prétraitement afin de prédire l’efficacité des produits cosmétiques, tout en permettant une économie de temps et de moyen. Cette approche nécessite la construction d’une base de données afin de lier les compositions de formules cosmétiques aux résultats de tests in vivo sur humain réalisés pour attester de leur efficacité. L’élaboration de cette base de données nécessite une réflexion autour des variables pertinentes pour décrire la composition des formules. Dans le même temps, un protocole permettant d’extraire les résultats d’efficacité jusqu’alors stockés dans des document Word est défini et appliqué pour les tests démontrant le pouvoir hydratant et l’effet lissant des produits. Les forêts aléatoires et les réseaux de neurones sont les deux algorithmes d’apprentissage automatiques choisis pour traiter ce problème de classification supervisée de grandes dimensions. Ces deux modélisations donnent des premiers résultats intéressants pour prédire la classe d’hydratation des cosmétiques à partir de leur composition. Dans le but d’étudier les interactions entre les éléments chimiques d’une formule, les vecteurs de descripteurs sont transformés et utilisés en entrées de réseaux de neurones convolutifs. L’explicabilité des forêts aléatoires et des réseaux de neurones étudié respectivement grâce à permutation de variables et la méthode Layerwize Relevance Propagation, amène trois réflexions. Premièrement, des concepts physico-chimiques sont intégrés par lors de la phase d’apprentissage. Cette idée est supportée par la pertinence des variables sur lesquelles s’appuient les modèles pour justifier leur prédiction d’hydratation. Deuxièmement, l’explicabilité sert à la détection et à l’interprétation de mauvaises classifications, comme c’est le cas pour l’effet lissant où les prédictions se fondent sur des descripteurs du panel volontaire et non sur la composition de la formule. Enfin l’explicabilité offre une opportunité de découvrir de nouveaux composés ou mélanges de composés chimiques prometteurs pour créer des formules hydratantes. Après avoir appris puis étudié les algorithmes d’apprentissage automatique, l’exploitation des réseaux de neurones complètement connectés permet de proposer des formules virtuelles générées avec ou sans contrainte de composition répondant à des critères d’efficacité prédéfinis.

**Mots clefs** : formules cosmétiques, test d’efficacité in vivo, classification supervisée, régression ordinale, réseaux de neurones, forêts aléatoires, explicabilité.

**Title** : Machine learning and big data toward mathematical modelling and predictive analysis on the efficacy of cosmetic formulas

**Abstract** : Cosmetic industry faces several challenges such as fierce competition and demanding customers. It must offer innovative products combining increasing efficacy and sensoriality while respecting regulatory requirements. Several tests are set up to support marketing claims and ensure the safety of products launched onto the market. In addition, machine learning algorithms can be viewed as a preprocessing solution to predict cosmetics efficacy saving resources and time. This approach requires the completion of a database relating cosmetics composition with the results of in vivo tests conducted on human volunteers. A discussion on the relevant features to describe the formulas’ compositions is carried out to create the database. Furthermore, a protocol is set up to extract the efficacy data stored in Word documents, and this procedure is applied to moisturizing and smoothing effect tests. Random forests and neural networks, two machine learning models, are selected to address this supervised classification problem. Those algorithms provide first interesting results to predict cosmetics hydration based on their chemical composition. Then, features vectors are processed and used to train convolutional neural networks in order to study interactions between chemical elements of a formula. Random forest and neural network explainability is explored respectively through features permutation and Layerwize Relevance Propagation technics and leads to different observations. The explanability ranks features according to their relevance in the model’s decision-making. Applied to hydration issue, it highlights the models’ understanding of physico-chemical mechanisms. Secondly, it allows to detect and interpret misclassifications. For instance, it explains the poor results on smoothing effect problem because decision-making is based on volunteers features instead of compositions ones. Finally, explainability provides an opportunity to explore new chemicals or chemicals mixture promising to create moisturizing cosmetics. After the study of machine learning algorithms, fully connected neural networks are used to suggest virtual formulas generated either with or without constraints on their compositions which meet preset efficacy criterions.

**Keywords**: cosmetic formulas, in vivo efficacy, moisturizing, supervised classification, ordinal regression, neural networks, random forest, explainability.