Artificial intelligence as a tool to predict cosmetics hydration properties

Sigaki, Higor Yudi Duenha[1]; Fonseca, Carlos Magno Molinaro[1]; Marinho, Beatriz Paudarco[1]; Weihermann, Ana Cristina[1]; Orizzi, Talita Damaris Cucolotto[2]; Waterkemper, Lilian Longuini de Souza[2]; Correa, Raphael Marcos Diniz Souza[1]; Firmino, Ana Raquel Inacio[2]; **Schuck, Desiree Cigaran**[2*];

[1] Diretoria Executiva de Dados, Grupo Boticário, Paraná, Brazil
[2] Diretoria de Qualidade Excelência e Cuidado, Grupo Boticário, Paraná, Brazil

*desirees@grupoboticario.com.br

Abstract

Product claims are the most important tool cosmetic companies use to promote the benefits of their products to the consumers. Choosing the right set of claims is fundamental for the market success of a cosmetic product. However, depending on the type of the claim, substantiation is required usually in the form of long and expensive laboratory instrumental tests. That is the case for claims related to the moisturizing properties of a cosmetic formulation. Here, we develop a machine learning approach to predict the effectiveness of moisturizing formulations. Our approach is conservative, in the sense that it comprises the results of two separate models, one considering the chemical compounds present in the formulation and the other considers the chemical function category of these compounds. By following the proposed framework, we achieve an accuracy of 88% of correct classifications and most importantly we wrongly classify as moisturizing a non-moisturizer product just once. This shows the high reliability and low risk of our model which is intended to support researches when developing new cosmetic formulations.

Keywords: Moisturizers; instrumental testing; machine learning, formulation similarity.

1. Introduction

From a marketing perspective, a product is made more appealing to the consumers given what it claims to do. Cosmetic claims are defined as any public information, usually for marketing purposes, on the content, the nature, the effect, the properties or the efficacy of the product[1]. When it comes to skin or hair, one of the most common and sought-after attributes for cosmetics is the ability to hydrate or to prevent excessive water loss. Moisture is constantly lost from the skin in a process called Transepidermal Water Loss and effective moisturizers should replace this lost moisture and/or help protect skin from further dehydration[2]. As such, claim substantiation is needed in the form of long and expensive laboratory instrumental tests.

In this sense, we propose a machine learning approach to predict the effectiveness of moisturizing formulations. Our method is based on the composition of the formulations and also on the function category of each compound to predict if a formulation will improve skin or hair hydration. We take

advantage of a proprietary dataset of laboratory instrumental tests results on the moisturizing properties of cosmetic formulations. This data makes our model very specialized in distinguishing moisturizer from non-moisturizer products of specific categories, such as products for hair or skin.

We design a conservative approach, which is based on the results of two separate models: one considering the presence/absence of the compounds, and the other takes into account the function categories of the compounds. By doing that, we try to avoid false positives as much as possible, that is, wrongly evaluating a non-moisturizing formulation as moisturizing.

The rest of this work is organized as follows. In the next section, we present the materials and methods underlying our model. We then conclude with a summary of the presented results and a discussion of their implications, such as reducing costs and accelerating the development process of new products.

2. Materials and Methods

2.1. Data

Data was collected from an internal historical database containing instrumental test results. These laboratorial tests employ different methodologies to assess skin and hair hydration according to the literature[3,4,5] always with a statistical comparison between a treatment and a control group. This data was stored as written reports in PDF files, and posteriorly extracted and structured with the help of generative AI tools, namely ChatGPT[6]. After checking and cleaning the data structured by ChatGPT we have the results of 615 formulations tested for skin or hair hydration. Then, we retrieve the compounds for each of these formulations from our internal database. At this point, we have a table in which each row corresponds to a formulation and the columns are the compounds and also the target binary variable extracted with ChatGPT, 0 if the formulation does not promote hydration or 1 if the formulation promotes hydration compared to the control group. We also use metadata containing the chemical function of each compound, for example: adenosine (compound) is a skin conditioning (function).

2.2. Formulation similarity

We use Tanimoto[7] to compute the similarity between each pair of formulations. To do so, we represent each formulation as a binary vector indicating the presence or absence of a compound. This representation is inspired by the *fingerprints* from cheminformatics[8]. Moreover, Tanimoto similarity has been extensively used by many chemoinformatics methods based on the principle that similar molecules have similar properties[9]. Here, inspired by this idea in this strict context of cosmetic products categories, we consider that formulations with similar compounds have similar properties. Tanimoto is also ideally suited for measuring proximity in sparse high-dimensional data[10,11,12]. Given two formulations A and B, such that each formulation is represented by a sparse binary vector in an *n*-dimensional feature space, Tanimoto, or extended Jaccard for binary vectors, is defined as the ratio of the number of compounds common to both formulations (intersection) to the total number of compounds (union). Mathematically, we have

$$T(A, B) = \frac{A \cap B}{A \cup B}$$

2.3. Models

Our approach to the classification task of predicting hydrating properties for a formulation is based on two stages. The first model takes into account the presence/absence of the chemical compounds to predict if a formulation has hydrating properties. The second model considers the chemical function of the compounds to perform the same task. We choose a conservative approach, in which the final prediction considers both models results and they need to agree that a formulation is positive for improving hydration. This is because the cost in time of development of wrongly assuming hydration is far greater than that of preemptively discarding a formulation candidate. All models used here are implemented in Python language via Sci-kit learn[13] and Scipy[14].

Results

Our results are based on data comprising the chemical compounds and functions of 615 cosmetics formulations of specific categories, namely hair and skin. For the task of classifying a formulation regarding the hydration potential we implement two methods. The first model is based on the pairwise Tanimoto similarity regarding the presence/absence of compounds and the second model

takes into account the chemical function of the compounds. The final prediction is made after considering the results of both models, that is, we classify a formulation as positive for improving skin or hair hydration if both models agree so. We consider this approach to be conservative, in the sense that we want to avoid false positives even though by doing that we will also get more false negatives, that is, predicting moisturizing formulas as non-moisturizing.

First, we calculate the Tanimoto similarity regarding the presence/absence of the compounds for all the pairs of formulations. In order to classify if a new formulation improves hair or skin hydration we get the test result of the most similar formula in our database. It is important to remark that we only evaluate formulations that are already expected to improve skin or hair hydration. This makes our models very specialized regarding the categories of formulations that can be effectively evaluated.

For the second model, we train a Random Forest algorithm[15] for the task of classifying if a formulation improves hydration based on the function categories of the compounds. Functional use categories are identifiers describing the function a specific chemical serves[16]. Examples of these functions are: cleaning agent, fragrance, humectant, etc. The input for the algorithm is the total concentration percentage for each function category. If two or more compounds have the same function, we add the concentration percentage values for each compound for that function. We optimize the decision threshold considering the balanced accuracy in a cross-validated strategy and combine it with a grid search algorithm to determine the best combination of parameters of the model.

The final prediction considers the output of the two algorithms. Both predictions need to agree that a formulation improves skin or hair hydration. Figure 1 shows the confusion matrix of the final predictions for a test set comprising approximately 25% of the data which was withheld and never seen by the trained model. The elements c_{ij} of this matrix represent the: c_{00} true negatives, c_{10} false negatives, c_{01} false positives and c_{11} true positives. We note that just one formulation was wrongly classified as positive for improving hydration, which is represented by the false positive in the confusion matrix.

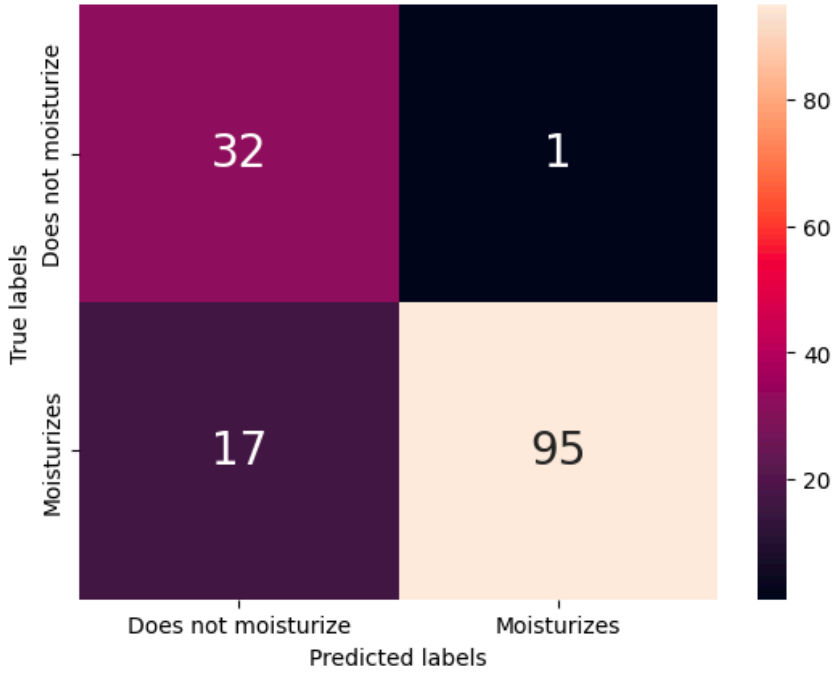

Figure 1. Predicting moisturizing properties of a cosmetic formulation. This confusion matrix shows the good performance achieved by the method we implemented. In particular, we note that just one formulation was wrongly classified as moisturizing which indicates the high reliability of our model.

Discussion

We have proposed a fast and reliable approach to predict the moisturizing properties of a cosmetic formulation. Our models are very specialized regarding the categories of products evaluated. The application of our methodology in the development of moisturizing cosmetic products is intended to support researchers when choosing the right compounds and concentrations for a new product before sending it to the laboratory instrumental tests. In this sense, it has the potential of reducing costs and accelerating the research and development of new moisturizing products.

Conclusion

While machine learning methods may not yet entirely replace laboratory procedures, these methods are already supporting researchers in several basic and applied research scenarios. Our work has demonstrated the usefulness of machine learning models for predicting the moisturizing properties of a cosmetic product in a strict context of types of formulations. Our results thus help reduce the shortage of machine learning research applied in the cosmetic industry.

Conflict of Interest Statement

The authors are employees of Grupo Boticário. However, the authors declare that the research was conducted in the absence of a potential conflict of interest.

References

[1] P. Romanowski And R. Schueller. Beginning Cosmetic Chemistry. 3rd edition. Allured Publishing Corporation. 2009.
[2] Purnamawati, Schandra et al. The Role of Moisturizers in Addressing Various Kinds of Dermatitis: A Review. Clinical medicine & research vol. 15,3-4 (2017): 75-87.
[3] Drozdenko, R.; Weinstein, C.; Weintein, S. Application of electrical hygrometric measures to the evaluation of hair moisturizing products. J. Soc. Cosmet. Chem. 43. 1992.
[4] Kurebavashi, A. Análise Sensorial na Determinação de Atributos do Consumidor como Valor Agregado ao Produto;
[5] Frosch, P. J., Kligman, A. M. The soap chamber test: a new method for assessing the irritancy of soaps. J. Am. Acad. Dermatology, v.1, p.35-41, 1979.
[6] OpenAI. (2023). ChatGPT (Feb 13 version) [Large language model]. https://chat.openai.com.
[7] Tanimoto, T. T. An elementary mathematical theory of classification and prediction. International Business Machines Corporation New York 1958.
[8] Capecchi, A., Probst, D. & Reymond, JL. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform 12, 43 2020.
[9] Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, B.K., Shoichet, John J.: Relating protein pharmacology by ligand chemistry. Nat. Biotechnol. 25(2), 197–206. 2007.
[10] Anastasiu, D.C., Karypis, G. Efficient identification of Tanimoto nearest neighbors. Int J Data Sci Anal 4, 153–172 2017.
[11] Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. J Cheminform 7, 20 (2015).
[12] Strehl, A., Ghosh, J.: Relationship-based clustering and visualization for high-dimensional data mining. Informs J. Comput. 15(2), 208–230. 2003.
[13] Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), pp.2825–2830.
[14] Jones, E. et al. SciPy: Open source scientific tools for Python (2001).
[15] Hastie T, Tibshirani R, & Friedman J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York, Springer.
[16] United States Environmental Protection Agency. Function Categories. https://comptox.epa.gov/chemexpo/functional_use_categories/. Accessed in May 2024.