Perfume Chemistry, Innovation & Tradition: How Machine Learning can Help Improve New Fragrance Creation Processes

Sebastián Correa

***Abstract*— As technology development has found its way through many of the before known fields with still traditional procedures, the demand for new intelligent, technology-driven solutions for many of the issues these fields encounter has been surging and expanding constantly. Perfumery, as one of these traditional fields, has recently been trying to evolve its procedures through peak technology, but a great part of its production still remains as manual labor. While it's understandable that this field's nature requires some tests to be manually taken, other processes that don't need such a specific approach can incur slowdowns and resource wastage. A newcomer perfumer's learning rate can be one main example of this. This paper attempts to solve such a problem by developing a Machine Learning model that can predict the outcome of a fragrance based on the raw materials used. The model's outcome was then tested with a veteran's perfumer feedback, whilst analyzing its usage in the perfume field and possible next steps.**

I. INTRODUCTION

The fragrance industry has existed for hundreds of years and offers different industries aromas that are aligned with the most demanded products throughout various national territories. Today, compounds and applications labs operate by implementing automation, robotics, digital tools, and lean management, while monitoring service rate, on-time rate, automation rate, lead time, and cycle time. However, the fragrance creation process as it is finds itself still too tied to manual labor. Although the perfumery industry has benefited from products for managing its shipping and mass production, the manufacturing process of new perfumes and aromas for different ternary products is still carried out without digitization, machinery revision, and resorting directly to human smell.

II. BACKGROUND

A. Perfumery

A perfume, according to Jean Carles [1], is a liquid mixture of top or primary notes (first impact, fresh), middle notes (main perfume character) and base notes (long-lasting) in solvents (ethanol, water, matrix). The role a dominant note plays is the most clearly perceived in the chord, which is the mixture of many raw materials and compounds. It is important to take into account that the primary note is not always the one with the majority percentage on a mixture. As Cristina Wilches [2] mentions, there are many note groups such as floral, citrus, aldehydic, fruity, sweet notes, and many others. Recognizing primary note of a fragrance is a critical task for the perfumer, and it is essential to master how they differ from each other.

B. Perfumery process

Cristina Wilches [2] talks about the process of creating a fragrance and how the perfumer begins to conceive his next creation. When creating a fragrance, the first step is to imaginatively conceptualize the product and its residual odors. Experienced perfumers already have a very clear idea of what raw materials might be used initially, and are generally faster than their amateur counterpart. Once the concept is clear, it is landed into a chemical formula in function of the perfume components, the raw materials to be used, and the different variations that can be taken. For it to be taken as a successful fragrance, Cristina Wilches [2] must go through theoretical and experimental scent tests, preliminary tests, client feedback, and mass production.

The theoretical and experimental scent tests refer to a small sample created from the chemical formula that can be smelled and evaluated to check the proximity of the theoretical and experimental smell, using the purest raw materials and refraining from making any more until the result is satisfactory. To accomplish this, there are a number of companies that are suppliers of raw materials for use in fragrances, which go under strict regulations to differ more than 5,000 existing raw materials using an appointment and technical sheet, safety sheet, and a unique identifier called CAS Number. It is also the perfumer's job to personally test the raw materials and perform a thorough quality control test before using them.

Once the perfumer is satisfied with the test scent, the formula is taken to real-life simulations where the new fragrance would be used [4]. The formula is applied to a base product that the client will ultimately use to verify that the fragrance does not have any unknown side effects and test how the formula performs and is manifested, as there are cases where the fragrance completely loses its smell, or entirely changes [4].

Client tests only occur when all lab tests are positive. It is normal for a fragrance laboratory to have hair dryers, washing machines, mops, hair strands, and laundry soap. If any of the tests fail, the perfumer goes back to the drawing board, and failed trials are collected into a mix of waste that can then be reused to mask other undesirable fragrances and therefore applied to cheap products, generally cleaning-related [4]. If all the tests are successful, the results are presented to

the client that requested the product. If the customer is not satisfied with the fragrance, feedback is received and the perfumer goes back to revising the formula of step 1.

If the client is satisfied with the produced fragrance, it's mass produced per the client's needs, and sent to the factory on which the client's product is being made. The perfumer also partakes in this step as a guide on how to generate large quantities of the formula without compromising results.

III. PROPOSED MODELS

A. To predict a fragrance dominant note

The invention of new fragrances is, essentially, a combination of a fixed quantity of a substance that may or may had not gone through previous preparation. The amount and selection of such substances are given by the perfumer through what is normally trial and error to get close to the envisioned result. This process is not done without previous investigation, such as beforehand knowledge of results when mixing specific substances. Moreover, one can predict the outcome of a mixture made from *n* number of mixtures of substances that have been tested before, allowing one to have *i* number of substances in *j* number of mixtures in order to create an *c* fragrance. Such configuration can be represented through

$$\sum_{i=1}^{n} s_i q_i = c_j$$

When analyzing the way a perfume process happens, it resembles many applications within various machine learning methods such as linear regression or classification, which means that they can be implemented in order to compare the best results if a valid data-set is given. The following is a breakdown of some machine learning methods considered to be applied to the data-set.

Accuracy Comparison	
Method	Advantage
Logistic Regression	It finds the probability that an instance belongs to a class.
K-Nearest Neighbors	Each data point is plotted as a point in a defined space.
Gaussian Naive Bayes	Easy to implement, does not require a lot of training data. Fast. It is used to make predictions in real-time.

Table 1. Algorithm advantages description.

IV. DATA SET

Cristina Wilches [2] provided a small data-set which can be found by clicking here.The data-set showcases five different types of primary notes that can be found when making harmonic mixtures of fragrances: floral, woody, aldehydic, fruity and sweet. All of the records take into account their base ingredient which is phenyl ethyl alcohol (rose essential oil).

A. Trends in research

The research trends of artificial intelligence have been very high ever since 1998 and 2007, and it only grew further afterward. The trend still showed growth remarks in 2017 [3]. While there are many methods such as neural networks that can be used without knowing precisely the exact calculations on a given training, but still knowing the underlying logic, there are way simpler ones that will later be discussed.

At its most basic sense, machine learning uses algorithms that learn and optimise their operations by analysing input data to make predictions within an acceptable range [4]. These algorithms tend to make more accurate predictions the more data is fed to them. Three broad categories can be made according to their purposes and the way the underlying algorithm works. These three categories are: supervised, unsupervised and semi-supervised [4]. In supervised machine learning algorithms, a labelled training dataset is used first to train the underlying algorithm. This trained algorithm is then fed on the unlabelled test dataset to categorise them into similar groups [4]. Supervised learning algorithms suit well with two types of problems: classification problems; and regression problems. In classification problems, the underlying output variable is discrete. When optimizing these algorithms, there is a risk in reaching an overfitting or underfitting problem. Overfitting occurs when the model model tries to cover all the data points or more than the required data points present, so the algorithm starts caching noise and inaccurate values [9]. Underfitting occurs when the model model is not able to capture the underlying trend of the data. It is because of this that regularization is often used to impose constraints on a neural network so that it limits a weight's value in hopes to reduce these two issues.

In the following subsections, the commonly used supervised machine learning algorithms for disease prediction [4] will now be briefly described

B. Logistic regression

Logistic regression comes from plotting all data in a plane and identify a line that is the closest to all of the dots in the plot[4]. It helps in finding the probability that a new instance belongs to a certain class. Since it is a probability, the outcome lies between 0 and 1. Therefore, to use the LR as a binary classifier, a threshold needs to be assigned to differentiate two classes. The idea is that, given that very straight line, one can predict where a new value could land on the graph given the already present data.

C. K-nearest neighbors

The K-nearest neighbor (KNN) algorithm is one of the simplest and earliest classification algorithms [5]. This algorithm is discrete, and does not require to consider probability values. The 'K' in the KNN algorithm is the number of nearest neighbors considered to take from. The selection of different values for 'K' can generate different classification results for the same sample object [4]. In short, the algorithm The KNN algorithm is a discrete classifier that assumes that

similar things are close to each other. In other words, similar things are close to their alike. One of the main challenges of K-nearest algorithms is to select the appropriate K for the data fed into the model[4]. To find a correct value, the algorithm is run multiple times with different values in order to pick that which reduces the number of errors encountered while maintaining the algorithm's ability to make accurate predictions when the answer is not in the given data [5].

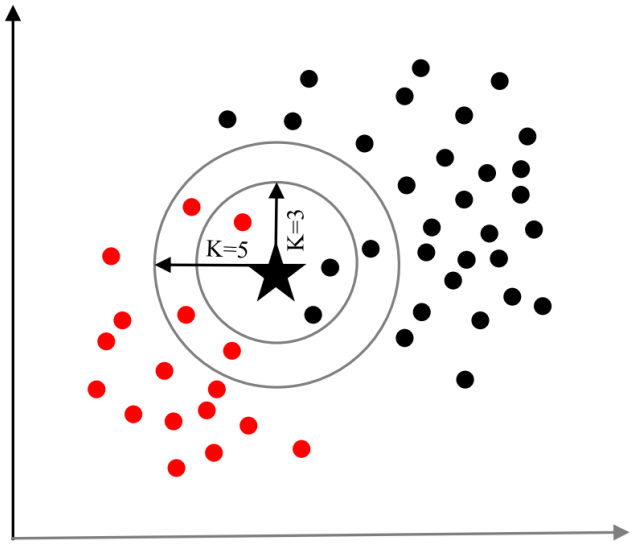

Fig. 1. A simplified illustration of the K-nearest neighbour algorithm. When K = 3, the sample object ('star') is classified as 'black' since it gets more 'vote' from the 'black' class. However, for K = 5 the same sample object is classified as 'red' since it now gets more 'vote' from the 'red' class [4].

D. Gaussian naive Bayes

Naïve Bayes is a classification technique based on the Bayes' theorem [6]. This theorem can describe the probability of an event based on the prior knowledge of conditions related to that event, while assuming that a particular feature in a class is not directly related to any other feature [4]. That said, features for that class could have interdependence among themselves [7].

E. Decision tree classifier

A decision tree models the decision logics as tests and corresponds outcomes for classifying data items into a tree-like structure [4]. The nodes of a decision tree normally have multiple levels where the first or top-most node is called the root node. Depending on the test outcome, the classification algorithm branches towards the appropriate child node where the process of test and branching repeats until it reaches the leaf node [8]. The leaf or terminal nodes correspond to the decision outcomes. Decision trees have been found easy to interpret and quick to learn, and are a common component to many medical diagnostic protocols [9]. When traversing the tree for the classification of a sample, the outcomes of all tests at each node along the path will provide sufficient information to conjecture about its class, as seen Fig. 3.

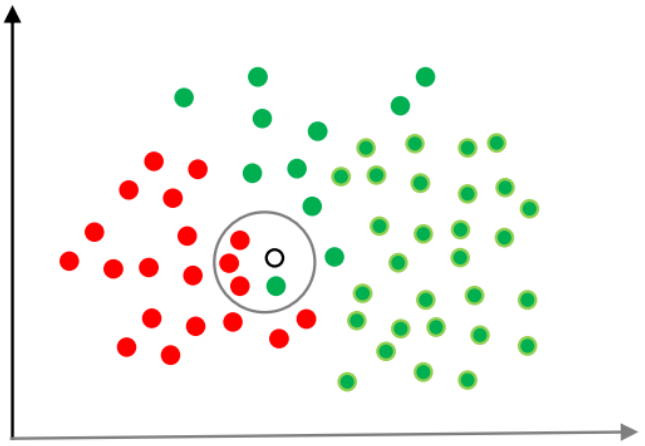

Fig. 2. An illustration of the Naïve Bayes algorithm. The 'white' circle is the new sample instance which needs to be classified either to 'red' class or 'green' class [4].

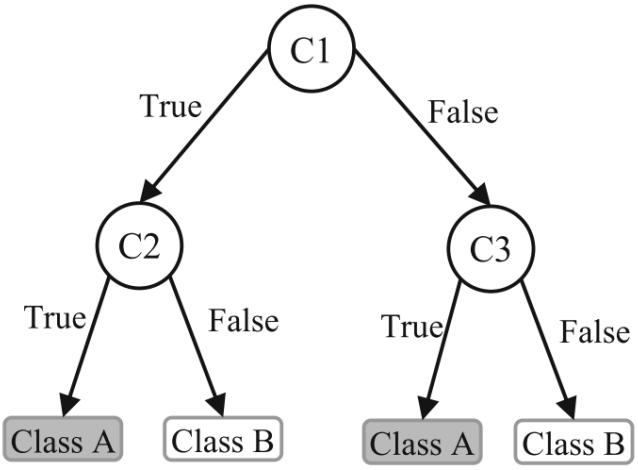

Fig. 3. An illustration of a Decision tree. Each variable (C1, C2, and C3) is represented by a circle and the decision outcomes (Class A and Class B) are shown by rectangles. In order to successfully classify a sample to a class, each branch is labelled with either 'True' or 'False' based on the outcome value from the test of its ancestor node [4].

F. Support Vector Machines (SVM)

Support vector machine algorithms can classify both linear and non-linear data [4]. It works by first mapping each data item into an n-dimensional feature space where n is the number of features, so that it then can identify the hyperplane that separates the items into two classes while maximising the marginal distance for both classes and minimising the classification errors [10]. Each data point is plotted first as a point in an n-dimension space (where n is the number of features) with the value of each feature being the value of a specific coordinate. To perform the classification, we then need to find the hyperplane that differentiates the two classes by the maximum margin. Figure 2 provides a simplified illustration of this classifier.

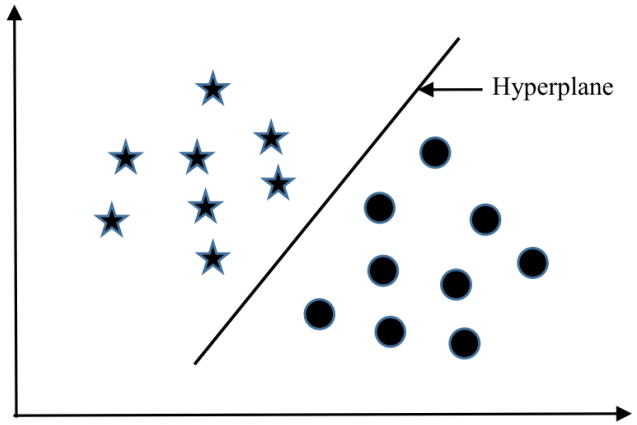

Fig. 4. A simplified illustration of how the support vector machine works, where a Support Vector Machine has identified a hyperplane which maximises the separation between the 'star' and 'circle' classes [4].

G. Resume

These previously described machine learning methods will be the ones used when training the data-set given by Cristina Wilches [2], and the implementation and results will now be discussed.

V. EXPERIMENTAL METHODS

Each of the previously discussed and chosen machine learning algorithms were started up in a Google Colab file, on which first a test data-set was used to configure the different algorithms and show the distribution of the data, to then shift to the data-set given by Cristina Wilches[2] in order to tinker them in order to try and obtain the best results of each. All of the methods' results are then compared from one another to pick the one with the best results. Such configuration can be found by clicking here.

VI. PERFORMANCE EVALUATION

Table 1 gives a breakdown on the results on the prediction results that each of the machine learning methods obtained after training them. Of them all five, all were above the seventieth percentile, but only two of them, being the Gaussian Naive Bayes and Decision Tree Classifier algorithms, reached the eightieth percentile, being the latter the one most accurate with a precision of 86%.

Accuracy Comparison	
Machine Learning Method	Accuracy result
Logistic Regression	79%
K-Nearest Neighbors	70%
Gaussian Naive Bayes	86%
Decision Tree Classifier	88%
Support Vector Machine	77%

Table 2. Accuracy comparison between machine learning approaches.

VII. RESULTS ANALYSIS

In order to check the validity of the predictions in a work-like environment, Cristina Wilches [2] that provided the data set gave the algorithm ten different combinations of different raw materials in order to test the prediction accuracy of the algorithm. Of the ten fragrances, 8 of them were accurate, while 2 of them missed the correct value.

The feedback Cristina Wilches [2] gave to the algorithm was that it has value to create perfumes, and also to have an internal evaluation panel, where she devised that all people, even if they are accountants, should have basic principles of smell and know how to communicate through the vocabulary of a perfumer when working in said industry. For new perfumers, Cristina Wilches [2] had the idea of a future software implementation to assist newcomers, since a beginner has everything to smell. Beginner perfumers make a good assistant, but the only way to gain experience is to learn by doing [2].

VIII. CONCLUSIONS

While the perfume industry shows a lot of interest in what technology can bring to its manufacturing process, the complexity of the processes within it and the manual labor that goes into creating a new fragrance is a major hurdle that can hamper this step. Because its quality evaluation is very restricted, it is difficult to implement technological solutions that can mimic the manual evaluation of a specialized chemist. Furthermore, while the chemical industry has turned to component description and production machines, the predictive field, digitized by Artificial Intelligence, is one that has yet to be fully explored and appears to harbor intriguing opportunities, such as assistance to beginners and task automation for experienced perfumers.

REFERENCES

[1] Carles, J. A method of creation in perfumery. Soap Perfum. Cosmet. 1962, 35, 328–335.
[2] M. Wilches Cárdenas, "Sobre la Industria de la Perfumería", Bogotá, D.C., 2021.
[3] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123135.
[4] S. Uddin, A. Khan, M. Hossain and M. Moni, "Comparing different supervised machine learning algorithms for disease prediction", BMC Medical Informatics and Decision Making, vol. 19, no. 1, 2019. Available: 10.1186/s12911-019-1004-8 [Accessed 27 May 2022].
[5] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7.
[6] Lindley DV. Fiducial distributions and Bayes' theorem. J Royal Stat Soc. Series B (Methodological). 1958;1:102–7.
[7] I. Rish, "An empirical study of the naive Bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, vol. 3, 22, pp. 41–46: IBM New York
[8] Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106.
[9] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Informat. 2006;2:59–77.
[10] Joachims T. Making large-scale SVM learning practical. SFB 475: Komplexitätsreduktion Multivariaten Datenstrukturen, Univ. Dortmund, Dortmund, Tech. Rep. 1998. p. 28.