UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Relationship Between Chemical Components and Aroma Profiles: A Machine Learning-Based Approach

Gabriel Vélez Malo

Carolina Lanas Terán

David Sebastián Flores Figueroa

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniería Industrial

Quito, 08 de mayo de 2025

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN

DE TRABAJO DE FIN DE CARRERA

Relationship Between Chemical Components and Aroma Profiles: A Machine Learning-Based Approach

Gabriel Vélez Malo

Carolina Lanas Terán

David Sebastián Flores Figueroa

Nombre del profesor, Título académico **María Gabriela Baldeón Calisto, PhD**

Quito, 08 de mayo de 2025

© DERECHOS DE AUTOR

Nombres y apellidos: Gabriel Vélez Malo

Código: 00212365

Cédula de identidad: 0105973309

Nombres y apellidos: Carolina Lanas Terán

Código: 00320269

Cédula de identidad: 1723422646

Nombres y apellidos: David Sebastián Flores Figueroa

Código: 00322915

Cédula de identidad: 0105070650

Lugar y fecha: Quito, 08 de mayo de 2025

ACLARACIÓN PARA PUBLICACIÓN

UNPUBLISHED DOCUMENT

RESUMEN

Este trabajo investiga la relación predicha entre la estructura química y la percepción olfativa utilizando algoritmos de aprendizaje automático. La colección contiene 44 descriptores de olor para 5,855 moléculas, cada una representada por su SMILES isomérico. Los datos fueron recopilados de bases de datos públicas y privadas y limpiados tras un exhaustivo preprocesamiento. Para traducir los SMILES a un lenguaje comprensible para el aprendizaje automático, se aplicaron las Morgan Fingerprints, creando dos conjuntos de datos con dos valores de radio diferentes (2 y 3). Tres algoritmos de aprendizaje automático —Random Forest, XGBoost y TabNet— fueron entrenados y evaluados utilizando métricas que incluyen Hamming Loss, AUROC, F1-score y Top-k True Skill Statistics ajustadas. Para abordar el desbalance multietiqueta del conjunto de datos, se implementó MLSMOTE, lo que mejoró significativamente la capacidad de generalización de los modelos. Se aplicaron un ANOVA de dos vías y una prueba de Tukey HSD para evaluar la significancia estadística de las interacciones entre el modelo y el conjunto de datos, identificando a XGBoost entrenado con el conjunto de datos MLSMOTE R3B2048 como el modelo de mejor desempeño. Este modelo fue validado adicionalmente utilizando un conjunto de datos externo de compuestos de cacao. Los resultados muestran que el aprendizaje automático, y específicamente XGBoost, puede predecir con precisión los atributos sensoriales. Esto abre la puerta a avances en la creación de fragancias y sabores y proporciona una alternativa basada en datos a las técnicas convencionales de categorización de olores.

Palabras clave: Aprendizaje automático, Clasificación multietiqueta, Predicción de aromas, SMILES isoméricos, Huellas digitales de Morgan (Morgan Fingerprints), XGBoost, Aumento de datos, MLSMOTE, Descriptores olfativos, Predicción de perfiles sensoriales

ABSTRACT

This work investigates the predicted link between chemical structure and olfactory perception using machine learning algorithms. The collection contains 44 odor descriptors for 5,855 molecules, each of which is represented by its isomeric SMILES. It was collected from public and private databases and cleaned after extensive preprocessing. To translate SMILES into a machine learning language, Morgan Fingerprints were applied creating two datasets with two different radius values (2 and 3). Three machine learning algorithms: Random Forest, XGBoost, and TabNet were trained and evaluated using metrics including Hamming Loss, AUROC, F1-score, and adjusted Top-k True Skill Statistics. To address the dataset's multi-label imbalance, MLSMOTE was implemented, significantly improving the models' generalization ability. A two-way ANOVA and Tukey HSD test valued the statistical significance of model and dataset interactions, identifying XGBoost trained on the MLSMOTE R3B2048 dataset as the best-performing model. This model was further validated using an external dataset of cacao compounds. The results show that machine learning, and specifically XGBoost, can accurately predict sensory attributes. This opens the door for advancements in fragrance and taste creation and provides a data-driven alternative for conventional odor categorization techniques.

Key words: Machine Learning, Multi-label Classification, Aroma Prediction, Isomeric SMILES, Morgan Fingerprints, XGBoost, Data Augmentation, MLSMOTE, Olfactory Descriptors, Sensory Profile Prediction

TABLE OF CONTENTS

TABLES INDEX

FIGURES INDEX

INTRODUCTION

The prediction of aroma profiles from molecular characteristics is a growing field of study due to its impact on various industries, such as fragrances, food, and pharmaceuticals. These industries often develop aromas and flavors using a traditional trial-and-error method, which, besides being time-consuming and costly, is highly subjective. Saini and Ramanathan (2022) argue that odor is a psychological construct since its classification relies on verbal descriptions provided by individuals, which can be influenced by factors such as age, culture, experience, and the evaluator's vocabulary.

To overcome this limitation, machine learning has proven to be more accurate in classifying aromatic profiles, as they rely on the intrinsic properties of molecules (Sisson et al., 2024). Furthermore, predictive models not only improve efficiency in product development but also facilitate the creation of new odor and flavor combinations that are tailored to consumer preferences (Keller et al., 2017). In this regard, machine learning undoubtedly has the potential to accelerate innovation, reduce costs, and expedite the discovery of new compounds, fragrances, and flavors, establishing itself as a key tool for optimizing and advancing the sensory industry.

In this work, we aim to evaluate the performance of machine learning and deep learning models to predict the olfactory descriptors of molecules based on their SMILES representations. The dataset was compiled from multiple public sources and curated to include 5,855 molecules labeled with 44 standardized aroma descriptors. The models analyzed include Random Forest, XGBoost, and TabNet, each evaluated under different configurations of Morgan Fingerprints (radius 2 and 3) and with and without the application of the MLSMOTE data augmentation technique to mitigate label imbalance. To guarantee reliable model comparison, statistical tests (ANOVA and Tukey HSD) and stratified cross-validation were also used in this study. The findings show that XGBoost had the greatest overall performance across the following metrics: Hamming Loss, AUROC, Top 2 TSS, Top 5 TSS, Precision, Recall, and F1-Score when trained on the SMOTE-augmented dataset using Morgan Fingerprints with radius 3. This suggests that XGBoost has the ability to provide precise and broadly applicable aroma prediction in multi-label classification scenarios.

LITERATURE REVIEW

Predicting olfactory perception from molecular structures has been a longstanding challenge in sensory research. Conventional methods depend on Quantitative Structure-Odor Relationship (QSOR) models, which correlate molecular properties (molecular weight, functional groups, etc.) with statistical techniques such as principal component analysis (PCA) and multiple linear regression (MLR) (Saini & Ramanathan, 2022). But according to Keller et al. (2017) the nonlinearity and complex relationships between molecular structure and odor perception are a frequent problem for QSQR models. Advances in machine learning (ML) and deep learning have made it possible to create more reliable prediction models by automatically identifying patterns in large data bases using techniques like Random Forests, Support Vector Machines (SVMs), and Graph Neural Networks (GNNs) (Sanchez-Lengeling et al., 2019).

A key study by Keller et al. (2017) developed a machine learning model that utilizes chemoinformatic descriptors of odor molecules to predict sensory attributes. The study predicted odor intensity and semantic descriptors such as “sweet”, “fruity”, and “spicy”. They discovered that linear models and Random Forest performed well in capturing important chemical characteristics associated with sensory perception.

Another approach to odor prediction using machine learning was introduced by Saini & Ramathan (2022). They identified the relationship between odor and chemical structure using a multi-label classification approach. They used Random Forest, Binary Relevance, and Classifier Chains models to predict several olfactory descriptors from chemical representations. They emphasized the complexity of QSOR modeling, highlighting that small structural changes can drastically alter perceived odors.

As part of a machine learning framework, XGBoost was used in "Data-Driven Elucidation of Flavor Chemistry" (Kou et al., 2023) to predict olfactory perceptions based on the chemical structures of molecules. When it came to managing high-dimensional chemical datasets and enhancing prediction accuracy, XGBoost was very helpful. The outcomes showed that this model was able to identify intricate patterns in olfactory data, which made them useful for studying taste chemistry and creating new aromatic chemicals.

Sanchez-Lengeling et al. (2019) used GNNs to predict perceptual odor similarities, which significantly outperformed QSOR models. Their research not only showed that deep learning captures better perceptual relationships, resulting in predictions that are more accurate and broadly applicable, but it also has demonstrated great performance in fields such as bioinformatics and cheminformatics, indicating its potential application in odor prediction.

Additionally, TabNet, a deep learning model developed by Arık & Pfister (2021), was proposed as a new method for tabular data learning. TabNet improves feature interpretability and learning efficiency by using sequential attention mechanisms, in contrast to conventional decision tree-based models like Random Forest and XGBoost.

To assess the performance of odor prediction models, researchers employed different metrics whether the task is classification or regression. The micro-averaged F1-score is commonly used for multi-label classification because it balances precision and recall, making sure there is a fair evaluation across all odor labels (Saini & Ramanathan, 2022). For Graph Neural Networks and Random Forest models, authors used the AUROC (Area Under the Receiver Operating Characteristic Curve) to evaluate how effectively they differentiate odor descriptors (Sanchez-Lengeling et al., 2019).

METHODOLOGY

Datasets and Preprocessing Operations.

The database used in this study was built by combining both public and private datasets. ChemTasteDB, FlavorDB, Goodscents, IFRA_2019, Keller, Leffingwell, and Sharma_2021b were among the public datasets that were sourced from GitHub and compiled in the Pyrfume repository (2024), for more information about the databases see annex 2. Two more datasets related to distilled rum and gin were included in order to expand the variety of aroma profiles (Camilo et al., 2023). These were provided by researchers from *Universidad de los Andes (UNIANDES)* based on previous studies. The databases included key information for each molecule, such as CAS number, compound name, IUPAC name, PubChem CID, SMILES notation, and odor descriptors. Initially, the dataset contained 361,640 molecules. However, after preprocessing the data, such as filtering duplicates, eliminating blank data, the database comprised 5,855 observations.

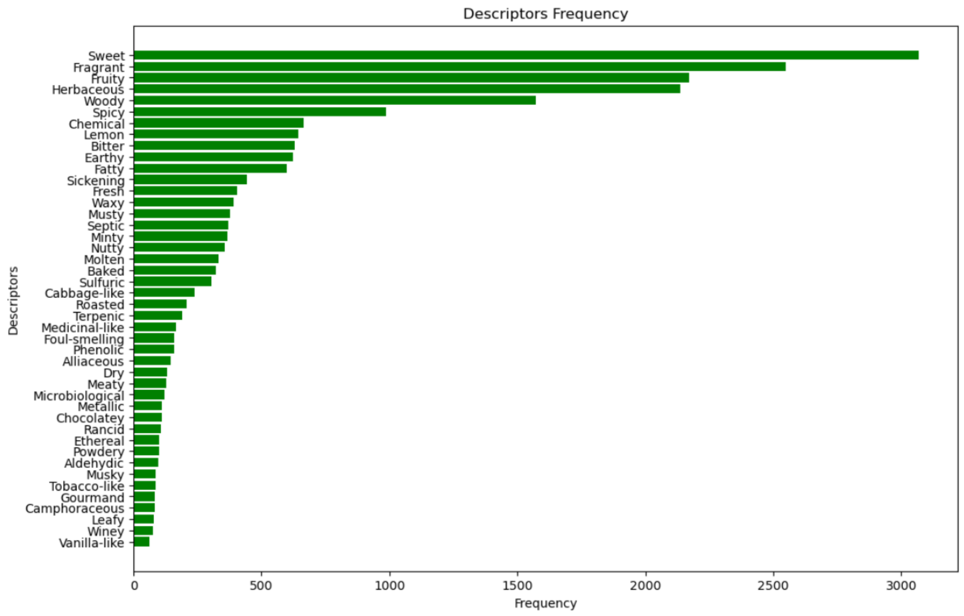
SMILES (Simplified Molecular Input Line Entry System) is a linear notation that uses ASCII character strings to uniquely and unambiguously describe molecular structures. There are two main variants of SMILES: Canonical SMILES that provide a standardized representation of the molecule's two-dimensional structure, without including information about chirality or isotope specification, and Isomeric SMILES, which in addition to basic connectivity, incorporate details about stereoisomeric configuration and isotope specification, allowing a more comprehensive description of the molecule's three-dimensional structure (Weininger, 2022). For this study, Isomeric SMILES were chosen as they provide more detailed structural information, which may enhance model performance by improving the accuracy of odor descriptors predictions.

PubChemPy Python library was used to integrate chemical data of Isomeric SMILES and to verify those already recorded. The PubChem database, which contains comprehensive information on chemical substances, is accessible through this library. PubChemPy is a Python package that serves as a wrapper for the PubChem REST API, enabling users to connect with the PubChem database and get chemical compound data with ease (Gqamana, 2024). This tool simplifies the search for chemical compounds by name, substructure, chemical standardization, graphical representation, and obtaining detailed chemical properties.

Since various sources utilized different terminology to refer to the same odor quality, descriptor names were first standardized to eliminate repetition across datasets. Similar descriptors were grouped under a unified label (e.g., Fruity, Non-citrusy fruity, and Tropical-fruity were all classified as Fruity). The frequency of occurrences for each unified description was then quantified by creating a frequency table. Ninety percent of the cumulative frequency's descriptors were chosen. This process led to a final set of 44 different descriptors, being Sweet the descriptor with the highest frequency in the database, 3068 times, and Vanilla-like the last descriptor appearing 63 times. This is graphically represented in Graph 1.

Figure 1

Descriptors frequency in the Data Base

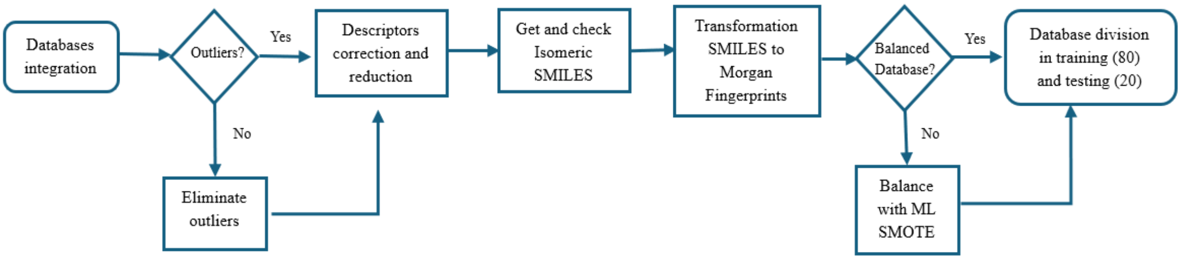


The next step in the study involved SMILES representations to a proper interpretation by machine learning models. Morgan Fingerprints was chosen for molecular feature extraction based on its effectiveness in prior studies. It has demonstrated success in a variety of machine learning techniques, such as random forests, graph neural networks, and classifier chains (Keller et al., 2017; Sanchez-Lengeling et al., 2019; Saini & Ramanathan, 2022).

The Morgan Fingerprints methodology is a graph-based molecular representation system that uses the molecule's SMILES notation to record its structural properties and connections. This approach, which is incorporated into the RDKit library, uses the Extended-Connectivity Fingerprints (ECFP) method, which views the bonds between molecules as connections and atoms as nodes in a graph. This enables the use of a hashing algorithm to generate unique IDs (Rogers & Hahn, 2010). Morgan fingerprints are more informative than other fingerprinting methods such as PubChem fingerprints or FP2 fingerprints because they can flexibly capture chemical structures by considering substructures of different sizes through a radius parameter (Zhou & Skolnick, 2024). According to Zhou and Skolnik (2024), a bit of 2048 and a radius of 2 offer an optimal balance between computational efficiency and information richness. This study, however, also wanted to analyze whether adjusting the radius parameter could improve odor prediction in machine learning models. To compare their effects, two separate databases were generated: one utilizing a radius of 2 and another one with a radius of 3, while maintaining a bit size of 2048 in both cases. It is important to mention that both databases have the same molecules, the only difference is the parameters on the Morgan Fingerprints. The databases will be referred to as R2B2048 for the first one and R3B2048 for the second database. The process is visually represented in Graph 2 and the final database can be found in Annex 5.

Figure 2

Preprocessing flowchart


Databases integration
Outliers?
Yes
No
Eliminate outliers
Descriptors correction and reduction
Get and check Isomeric SMILES
Transformation SMILES to Morgan Fingerprints
Balanced Database?
Yes
No
Balance with ML SMOTE
Database division in training (80) and testing (20)

Performance with Cacao Molecules.

To validate the performance of the optimal model-database combination, the selected model was evaluated using a new, independent dataset provided by *Universidad de los Andes* (UNIANDES) from FlavorDB2 (FlavorDB2, 2025), consisting of 276 molecules found in cacao. This dataset went through the same preprocessing steps as both datasets R2B2048 & R3B2048, but to ensure the proper functionality of the model when applied to new data, the cacao dataset was aligned with the 47 descriptors previously selected during preprocessing of the original dataset. However, only 35 of those descriptors were present in the cacao data. This reduction was due to the absence of certain descriptors in the cacao samples that were available in the original dataset; for instance, descriptors like *Alliaceous* were present in the original dataset but not in the cacao set. Consequently, just the overlapping descriptors were kept, giving the two datasets a straightforward uniform representation.

After aligning the descriptors, the best-performing machine learning model identified through the two-way ANOVA and Tukey analysis, was evaluated by training it exclusively with the original dataset and then testing it with the cacao dataset. This method was created to mimic a real-world situation in which the model comes upon completely fresh, unseen data. The objective was to evaluate the model's performance and capacity for generalization when used on a customized dataset with varying aromatic profiles and chemical compositions.

It’s important to mention that the fingerprint parameters used for this new dataset were selected based on the optimal ratio determined by the prior ANOVA analysis.

Data Augmentation.

To start working on the machine learning algorithms, it was necessary to assess the dataset’s balance. For this purpose, metrics like MeanIR and IRLbl were employed; MeanIR measures the overall imbalance across labels, whereas IRLbl assesses the imbalance at the label level. Additionally, visual tools and representations such as histograms of label distribution and a co-occurrence matrix were implemented to analyze correlations between the labels in the dataset (Keller et al., 2017). This imbalance ratio according to Noorhalim et al. (2019), can be classified into three categories: mild imbalance, when this ratio is less than three; medium imbalance, when the ratio is between three and six; and finally, when the ratio is more than six, is considered as extreme imbalance ratio.

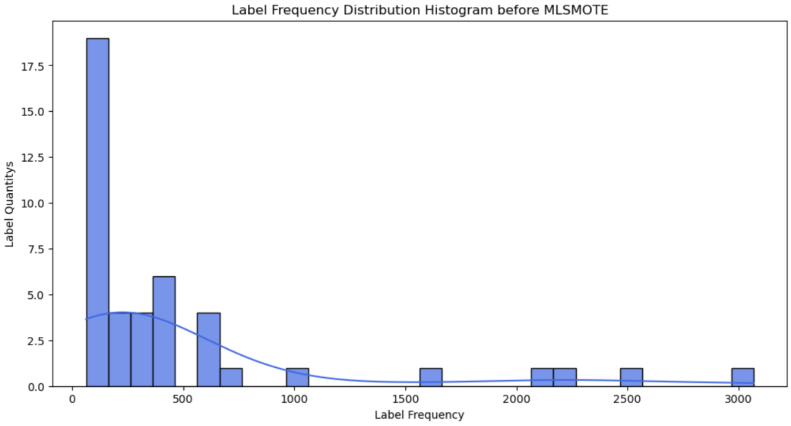
Machine learning models generally typically assume a balanced distribution of classes. A data augmentation method known as the Multi-Label Synthetic Minority Over-Sampling Technique (MLSMOTE) was studied to solve this problem. Sukhwani (2021) claims that MLSMOTE is a development of the Synthetic Minority Over-Sampling Technique (SMOTE), which was created especially for multi-label classification issues. MLSMOTE is designed to handle data where each instance may be linked with multiple labels at the same time, whereas SMOTE creates fresh synthetic samples for minority classes in single-label classification tasks. This is especially useful in cases where certain combinations of labels are infrequent, which could lead to biased models with reduced generalization ability.

MLSMOTE starts with identifying minority class instances, unlike conventional oversampling methods, MLSMOTE creates new observations based on real data to avoid overfitting, instead of simply replicating these instances. Once minority instances are identified, MLSMOTE applies a K-Nearest Neighbors (KNN) algorithm to find similar samples within the feature space. From these neighbors, the algorithm interrogates new synthetic data points by combining feature values in a weighted manner (Sukhwani, 2021). By guaranteeing that synthetic instances are uniformly dispersed throughout the data space instead of being grouped in one area as happens with duplication techniques, this procedure aids in producing more realistic samples. One of the main characteristics of MLSMOTE is that it maintains the correlation between labels, which means that new samples are produced based on the relational structure of labels in the dataset as well as numerical properties.

A notable discrepancy in the dataset was found by the label imbalance analysis. With a Mean Imbalance Ratio (MeanIR) of 6.1, it can be concluded that the dataset exhibited a severe class imbalance. Some labels are much less common than the dominating sweet label (3,068 samples), based on the Imbalance Ratio per Label (IRLbl). Most labels contain fewer than 1,000 samples, according to a label distribution histogram Graph 2, with the most common labels being Sweet, Fragrant, and Fruity.

Figure 3

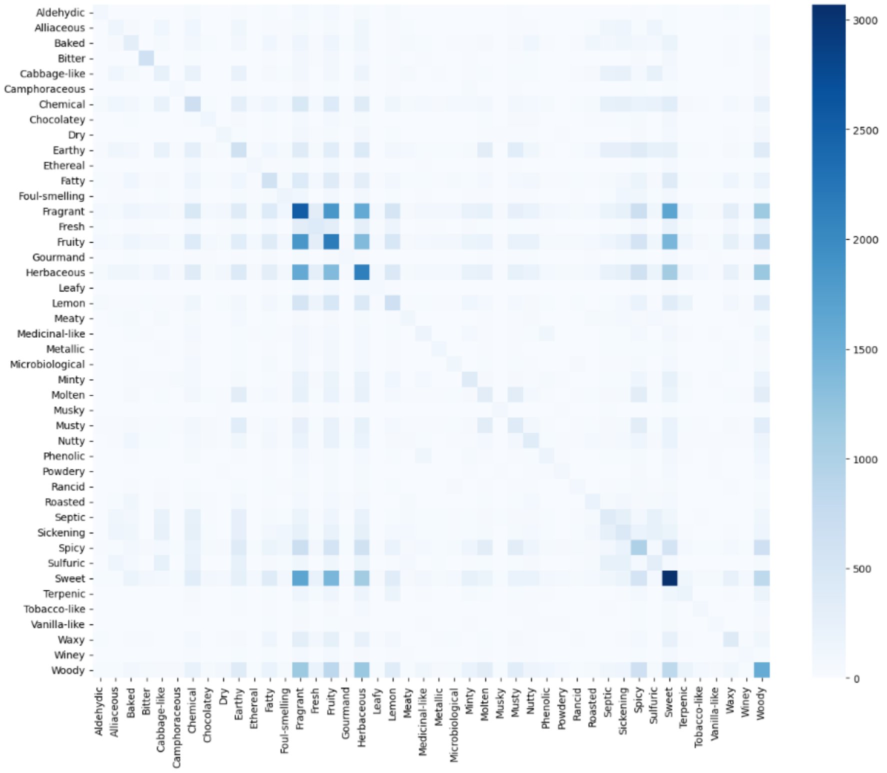
Label Frequency Distribution Histogram before MLSMOTE



Furthermore, stronger relationships between frequently occurring labels were also highlighted using a label co-occurrence matrix shown as a heatmap in Graph 3.

Figure 4

Heatmap: Label co-occurrence matrix



The co-occurrence matrix shows strong relationships between broad descriptors such as Fruity, Sweet, Fragrant, and Fresh, as previously mentioned, which frequently appear together across molecules. In contrast, descriptors like Sulfuric or Septic show low co-occurrence, indicating specificity.

The co-occurrence matrix shows strong relationships between broad descriptors such as Fruity, Sweet, Fragrant, and Fresh, reinforcing the results of the label imbalance analysis. This not only affects individual label frequencies but also shapes how often descriptors appear together.

To address the imbalance, MLSMOTE was applied to generate synthetic samples and balance the label distribution. A total of 2,000 synthetic samples were generated to enhance the dataset's representation within the classification task. As a result, the Imbalance Ratio per Label (IRLbl) (Annex 1) and Mean Imbalance Ratio (MeanIR) was reduced to 3.7633 (Table 1), which made it possible to continue creating machine learning models with a more evenly distributed set of labels. This is also shown on the Frequency Distribution Histogram in Graph 4.

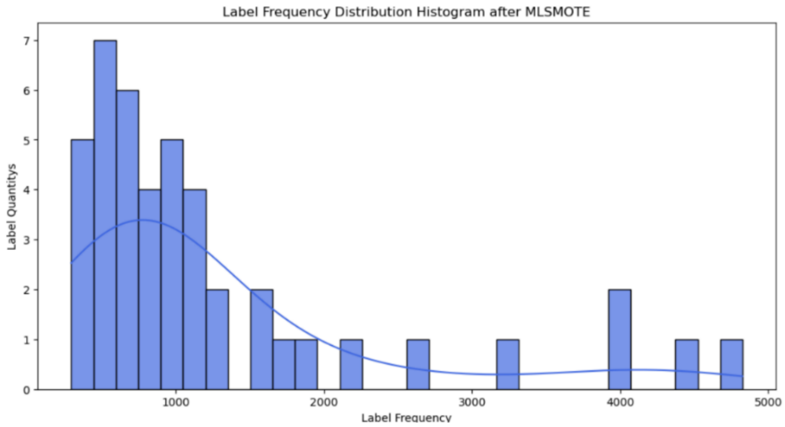
Table 1:

Results of MeanIR metrics before and after applying MLSMOTE to the datasets

	MeanIR
Before MLSMOTE	6.1005
After MLSMOTE	3.7633

Figure 5

Label Frequency Distribution Histogram after MLSMOTE



It’s important to mention that to evaluate the impact of SMOTE on model performance, a test run was conducted for each model using both the first original dataset (R2B247) and the first dataset processed with MLSMOTE (R2B247_MLSMOTE).

Dataset division.

In machine learning, the division of datasets into training, validation and test is crucial for developing models that generalize unseen data. One of the best practices is to use the 80-10-10 ratio, allocating 80% for the training set, 10% for the validation, and 10% for the test, preventing issues like overfitting and having more reliable performance metrics (Sumalatha et al., 2024). Since the dataset on this research with the imbalanced ratio couldn't satisfy that all labels were evenly split on all the sections, Keller et al. uses the 80-20 subsets, being 80% (6,284 labels) for the training and 20% (1,571 labels) for the testing, and this also works perfectly for the model and becomes effective because it won't leave each subset too small, and also maximizes the training data (2017). It is important to mention that since a fivefold stratified cross-validation technique will be used for developing a better model, the validation process is no longer necessary because it is already integrated within the training phase.

Machine Learning Models.

For the machine learning algorithms, three algorithms were decided to analyze. The first one is TabNet, a deep learning architecture designed to work with tabular data proposed by Sercan O. Arik and Tomas Pfister in 2019. This model, unlike others from deep learning, uses sequential attention to choose which features to reason from at each decision step, enabling interpretability and more efficient learning as the learning capacity is used for the most salient features (Arik & Pfister, 2021). One of the principal advantages of TabNet, and the reason it was selected for this study, is its ability to outperform neural network and decide three variants on a wide range of unsaturated tabular datasets in performance. Arik & Pfister (2021) mention that this model provides interpretable feature attributions and offers detailed insight into the overall behavior of TabNet. Another important contribution from this model is focused on tabular data, where the demonstration of self-supervised learning is significantly improving performance when a large amount of unlabeled data is available.

Complex multi-label classification issues also can be handled using the Random Forest algorithm, a machine learning method based on an ensemble of several decision trees. In this method a random sample of the dataset is used to train each tree, and the final predictions are derived by voting or averaging each tree separately. One Random Forest model can predict several classes at once in multi-label issues without having to handle each label independently. By enabling each sample to belong to many categories and assigning labels based on the sum of the predictions from separate trees, Random Forest effectively manages multi-label classification (Clare & King, 2001).

The latest model in use, XGBoost (Extreme Gradient Boosting), is a machine learning model based on decision trees that has been improved for speed and efficiency. It adopts a boosting technique in which several trees are trained one after the other, with each new tree focusing on fixing the mistakes of the one before it. Techniques including early pruning and effective management of missing data are used to increase accuracy and decrease overfitting (Sanz, 2024). Because it manages complex relationships between labels, XGBoost is helpful for these kinds of multi-label classification tasks. It is also quicker than the earlier mentioned Random Forest or Neural Network techniques due to its parallelization and optimization features. Because of its balance between accuracy, interpretability, and computational economy, XGBoost was utilized in this model.

The performance of Random Forest and XGBoost were evaluated using n=100 and n=300 for each dataset while maintaining the same default hyperparameters. This approach ensures consistent conditions across both models, allowing for a more accurate performance assessment.

Evaluation Metrics.

Proceeding with the model's performance metrics, it is essential to choose the most appropriate since it will evaluate and understand the model's behavior, especially in multi-label classification tasks. Archaya in his article talks that accuracy is a more comprehensive performance evaluation and represents the proportion of correctly predicted instances out of the total instances, in other words, it is used to know the ability of a model to correctly classify data points, regardless of the prediction performance by class or label. Recall is another metric used to assess the model's capacity to detect all relevant instances, demonstrating how well it captures positive cases. In multi-label cases, recall is calculated for each label and summed, indicating the model's ability to retrieve all relevant labels across occurrences (2024). Hamming Loss on the other side is a measure of the proportion of labels that are improperly predicted, measuring the percentage of misclassified labels, indicating how many incorrectly label assignments occur on average; a lower Hamming Loss suggests better performance. F1-Score is the fourth evaluation metric in use to measure the model's accuracy, combining and balancing precision and recall, so with a high F1-Score means it is good at both identifying relevant instances and minimizing false positives (Acharya, 2024). The performance metric Area Under the ROC Curve (AUROC) evaluates a model's capacity to differentiate between classes, in this case aromatic profiles. The True Positive Rate (TPR) and False Positive Rate (FPR) at different decision thresholds are compared using the Receiver Operating Characteristic (ROC) curve as its foundation. AUROC is notably beneficial in issues with imbalanced classes since it does not rely on a set classification threshold. Multi-label classification can be adapted using strategies such as One-vs-Rest (OvR), which calculates a ROC curve for each label relative to the rest, or One-vs-One (OvO), which compares each pair of labels individually; the latter strategy was utilized in this study. The lasts evaluation metrics is the Top-5 TSS and Top-2 TSS, meaning Top-k True Skill Statistics (TSS) evaluate the model's performance based on its ability to correctly predict the true labels within its top-k predictions (Yoon & Lee, 2022). Top-5 TSS assesses whether the true label is among the model's top five predicted labels, while Top-2 TSS checks within the top two predictions. This last evaluation was included since in the challenge the authors are referencing, they include this metric, and it was decided to use for a comparison between a global challenge on this topic and the current situation, with the opportunity to improve the evaluation metrics' results. The formulas of each metric can be found in Annex 6.

Stratified Cross-Validation: Evaluating Generalizability and Robustness.

To evaluate the performance of Random Forest, XGBoost and TabNet algorithms with another verification method, stratified cross-validation was used. According to, Szeghalmy & Fazekas (2023), the stratified cross-validation (SCV) is a robust version of the common k-folds cross-validation where the labels are randomly splited in to the folds, but the variability in the distribution can affect the strength of the validation, so the best way to solve this is using the SCV where in every fold follows a similar distribution to the original distribution. This validation helps to determine which of the models is having better performance by evaluating it acrros different folds with similar conditions. Although, the number of folds (k-value) goes from 5-10, for this study a 5-folds SCV was selected, to ensure a consistent proportion in all the sets of training and testing, and give a more consistent evaluation of the models. (Prusty et al., 2022).

Hypothesis and means test.

To evaluate the effects of the database and machine learning model on performance, a two-way ANOVA was conducted. This statistical method was chosen because it allows for the examination of both main effects of radius on molecular representation and model and their interaction effect, determining whether the model's performance varies depending on the dataset used. The stratified cross-validation process provided the data used in this analysis. A five-fold cross-validation was performed for every combination of database type and machine learning model, and the outcomes of F1-score of each fold were saved. As a result, thirty data points in total were examined. The F1-score metric was selected because it is particularly useful in multi-label classification tasks, it provides a balanced evaluation of both precision and recall. This balance is essential when a model must correctly identify not just whether labels are present, but also how many and which ones, making it well-suited for complex multi-label problems (Bénédict et al., 2021). The data and combinations used for the ANOVA analysis can be found in Annex 2.

However, before applying ANOVA, its assumptions (normality, homogeneity of variances, and independence of residuals) were tested to ensure the reliability of statistical conclusions. The initial study showed violations of normality and homoscedasticity so, to ensure the validity of the ANOVA assumptions, a Box-Cox transformation was applied to the response variable. The estimated lambda (λ) value was 3.27, which was rounded to λ = 3 for practical purposes. The Box-Cox transformation is commonly used to correct non-normal distributions in parametric tests by applying an optimal power transformation, ensuring that ANOVA assumptions are met (Sureiman & Mangera, 2020). After the transformation, residual analysis confirmed improved normality and variance homogeneity, ensuring the validity and reliability of ANOVA results. The Hypothesis for the Two-Way Anova is shown in Table 2.

Table 2

Hypotheses for Two-Way ANOVA

Hypotesis Type	**Null Hypotesis (H_0)**	**Alternative Hypotesis (H_1)**
Effect of radius on molecular representation	The radius on molecular representation does not affect performance.	The radius on molecular representation affects performance.
Effect of Model	The model does not affect performance.	The model affects performance.

Interaction (Radius * Model)	No interaction between the model and the radius on molecular representation.	Interaction exists between the model and the radius on molecular representation.

Table 1 summarizes the null and alternative hypotheses tested in the two-way ANOVA, including the effects of the database, the model, and their interaction. The significance level for each hypothesis was set at α = 0.05. If the p-value is less than or equal to 0.05, the null hypothesis is rejected, indicating a statistically significant effect. On the contrary, if the p-value is greater than 0.05, there is not enough evidence to reject the null hypothesis.

If the two-way ANOVA returns statistically significant results, a Tukey's Honestly Significant Difference (HSD) test will be conducted as a post hoc analysis to identify which specific group means differ from each other. The Tukey HSD test determines the differences between each pair of means and compares it to a critical value. This method effectively controls the family-wise error rate, ensuring that the probability of making at least one Type I error remains within the predefined significance level (α = 0.05). This makes it possible to compare all pairwise group means in a meaningful way, which provides a better understanding of the major differences (Nanda et al., 2021). The hypothesis for the Tukey's Honestly Significant Difference is shown in Table 3

Table 3

Hypotheses for Tukey's Honesty Significant Difference

Hypotesis Type	**Null Hypotesis (H_0)**	**Alternative Hypotesis (H_1)**
Pairwise Mean Comparison	There is no significant difference between the group means.	There is a significant difference between at least two group means.

RESULTS

The performance of the models was evaluated by using multiple evaluation metrics, including Hamming Loss, AUROC, Top-2 and Top-5 True Skill Statistics (TSS), Precision, Recall, and F1-score. The models were tested with and without SMOTE to evaluate the impact of data augmentation on classification performance. Additionally, in Random Forest and XGBoost models, different tree counts (n=100 and n=300) were analyzed as hyperparameters and to examine their effect on model robustness and predictive stability. In the following subsections, we present the experimental results.

Tab Net Model

SMOTE Tab Net.

Table 4

Comparison of the performance of the first dataset without SMOTE (R2B2048) and the first dataset with SMOTE (R2B2048_MLSMOTE) in the Tab Net model.

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R2B2048	0.2361	0.5752	0.2134	0.3873	0.2482	0.3500	0.2787
R2B2048_MLSMOTE	0.1498	0.7548	0.3026	0.5114	0.5892	0.2929	0.3913

The model was evaluated using two databases, both with a radius of two, one applying SMOTE and the other without. Results from TabNet demonstrated significant improvements across most metrics when SMOTE was applied. Specifically, the AUROC increased from 0.5752 to 0.7548, precision improved from 0.2482 to 0.5892, and the Hamming loss decreased from 0.2361 to 0.1498, indicating that the model predicted fewer incorrect labels compared to the true labels.

Performance Tab Net.

Table 5

*Performance metrics of the Tab Net model on both datasets with MLSMOTE (*R2B2048_MLSMOTE and R3B2048_MLSMOTE*)*

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R2B2048_SMOTE	0.1498	0.7548	0.3026	0.5114	0.5892	0.299	0.3913
R3B2048_SMOTE	0.1551	0.7642	0.2998	0.5088	0.6031	0.2656	0.3688

The two databases (radius two and three) were evaluated in the Tab Net algorithm. The difference in the metrics between both was minimum, because they had a similar development. However, in the metrics of Hamming loss, Top 2 TSS, Top 5 TSS, Recall and F1-Score that the database with radius two showed slightly better results, making this one the best to use in Tab Net model.

Stratified Cross Validation Tab Net.

Table 6

Average performance ± standard deviation of the five folds conducted through cross-validation in the Tab Net model.

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R2B2048_SMOTE	0.155±0.001	0.662±0.029	0.287±0.013	0.610±0.009	0.601±0.016	0.1633±0.047	0.2528±0.057
R3B2048_SMOTE	0.158±0.002	0.688±0.018	0.293±0.007	0.448±0.021	0.602±0.004	0.169±0.047	0.261± 0.057

The performance of the Tab Net model under five-fold stratified cross-validation revealed similar outcomes across the two datasets, R2B2048 and R3B2048, with the latter demonstrating a slight advantage. Stratified Cross Validation results showed an improved performance in the metrics of AUROC, Top 2 TSS, Precision, Recall and F1-score for the

R3B2048 dataset. Overall, the evaluation metrics suggest that the model exhibits a relatively balanced predictive behavior (F1-Score 0.261± 0.057 and Precision 0.602±0.004), however, the recall metric had a low value (0.169±0.047) and this means that the model have a conservative tendency in identifying true positive instances, which could limit the effectiveness of the model in cases where sensitivity is critical.

Random Forest Model

SMOTE Random Forest.

Table 7

Comparison of the performance of the first dataset without SMOTE (R2B2048) and the first dataset with SMOTE (R2B2048_MLSMOTE) in the Random Forest model.

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R2B2048	0.067	0.790	0.503	0.718	0.660	0.433	0.523
R2B2048_MLSMOTE	0.068	0.912	0.395	0.608	0.795	0.640	0.709

Model performance was significantly improved by using MLSMOTE, as shown by a comparison of the dataset without and with SMOTE. In particular, the AUROC metric grew from 0.790 to 0.912, improving the model's capacity to distinguish between various odor descriptors. The F1-score also increased from 0.523 to 0.709, suggesting a more equitable trade-off between recall and accuracy. On the other hand, the Top-5 TSS and Top-2 TSS decreased from 0.718 to 0.608, indicating that although the model's overall classification accuracy increased, its capacity to accurately rank the real label among the top five and two predictions were somewhat weakened.

Performance Random Forest.

Table 8

*Performance metrics of the Random Forest model on the first dataset with MLSMOTE (*R2B2048_SMOTE*) with n=100 and n=300*

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R2B2048_SMOTE n=100	0.086	0.912	0.395	0.608	0.795	0.640	0.709
R2B2048_SMOTE n=300	0.085	0.915	0.399	0.608	0.797	0.642	0.711

Table 9

*Performance metrics of the Random Forest model on the second dataset with MLSMOTE (*R3B2048_SMOTE*) with n=100 and n=300*

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R3B2048_SMOTE n=100	0.085	0.912	0.400	0.605	0.804	0.643	0.714
R3B2048_SMOTE n=300	0.084	0.916	0.402	0.607	0.807	0.646	0.717

The effect of increasing the number of trees from 100 to 300 was analyzed for both R2B2048_SMOTE and R3B2048_SMOTE datasets. The results showed a marginal improvement in AUROC scores across both datasets, while the F1-score increased slightly, indicating a minor yet consistent enhancement in classification performance. Meanwhile, there were very slight changes in Top-2 and Top-5 TSS, indicating that the influence of a larger tree count on ranking accuracy was minimal.

Stratified Cross Validation Random Forest.

Table 10

Average performance ± standard deviation of the five folds conducted through cross-validation in the Random Forest model.

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R2B2048_SMOTE n=300	0.084±0.002	0.917±0.002	0.398±0.012	0.610±0.009	0.802±0.013	0.640±0.004	0.712±0.005
R3B2048_SMOTE n=300	0.082±0.002	0.916±0.004	0.396±0.013	0.606±0.009	0.817±0.011	0.653±0.007	0.726±0.008

The five-fold stratified cross-validation results show the Random Forest model's stability and strong classification performance across different dataset partitions. With n = 300, both datasets (R2B2048_SMOTE and R3B2048_SMOTE) showed high AUROC scores and low Hamming Loss, suggesting a low percentage of incorrectly categorized labels and potent discriminating potential.

XGBoost Model

SMOTE XGBoost.

Table 11

Comparison of the performance of the first dataset without SMOTE (R2B2048) and the first dataset with SMOTE (R2B2048_MLSMOTE) in the XGBOOST model.

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R2B2048	0.066	0.833	0.508	0.728	0.667	0.464	0.547
R2B2048_MLSMOTE	0.086	0.932	0.398	0.616	0.795	0.643	0.711

For the XGBoost model, all the metrics got a better score, except for the Top 2, Top 5 TSS and hamming loss indicating that its ability to correctly place the actual label in the top five and two forecasts were a little affected. The model with data augmentation got an improvement from 0.833 to 0.932 for AUROC and from 0.547 to 0.711 for F1-score, meaning that precision and recall are having a good balance.

Performance XGBoost.

Table 12

*Performance metrics of the XGBoost model on the first dataset with MLSMOTE (*R2B2048_SMOTE*) with n=100 and n=300*

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R2B2048_SMOTE n=100	0.086	0.932	0.398	0.616	0.795	0.643	0.711
R2B2048_SMOTE n=300	0.081	0.936	0.393	0.611	0.788	0.694	0.738

Table 13

*Performance metrics of the XGBoost model on the second dataset with MLSMOTE (*R3B2048_SMOTE*) with n=100 and n=300*

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R3B2048_SMOTE n=100	0.083	0.933	0.404	0.606	0.813	0.652	0.724
R3B2048_SMOTE n=300	0.080	0.938	0.397	0.599	0.799	0.697	0.745

About the performance of the model with the different tree count, it can be appreciated that on both datasets, the best results were gotten with 300 trees, meaning that the model got to understand and predict better metrics with more trees. The improved metrics were hamming loss, AUROC, recall and F1-score which are the most important ones. For Top 2, Top 5 TSS and precision were slightly lower, with a minimal difference. And as an overall for all the metrics, the highest one is AUROC on both datasets with 300 trees, having 0.936 for R2B2048_SMOTE, and 0.938 for R3B2048_SMOTE.

Stratified Cross Validation XGBoost.

Table 14

Average performance ± standard deviation of the five folds conducted through cross-validation in the XGBOOST model.

	Hamming Loss	AUROC	Top 2 TSS	Top 5 TSS	Precision	Recall	F1-score
R2B2048_SMOTE n=300	0.081±0.001	0.953±0.001	0.377±0.012	0.547±0.009	0.793±0.011	0.683±0.006	0.734±0.003
R3B2048_SMOTE n=300	0.080±0.002	0.954±0.002	0.375±0.001	0.548±0.007	0.803±0.009	0.691±0.004	0.742±0.006

Finally, the stratified cross validation with 5-fold was applied to the dataset with the bests results, which were both with 300 trees. All the performance metrics improved their results because the model had several attempts to learn the data. The only metrics with lower score were AUROC on R3B2048_SMOTE, Top 5 TSS on R2B2048_SMOTE, and Recall and F1-score on both datasets, but the standard deviation is relatively low, so the results remain very confident.

Two-Way ANOVA

Table 15

Two-Way ANOVA Results for Model and Database Performance.

Effects	Valor p
Radius	0.000
Model	0.000
Data Base*Model	0.001
Error	
Total	

The two-way ANOVA results indicate a significant main effect of the radius on molecular representation on model performance (p = 0.001), a significant main effect of the model (p =

0.001), and a significant interaction effect between radius on molecular representation and model (p = 0.001). These results suggest that the choice of database and model significantly influence performance, and that the effect of the model varies depending on the radius on molecular representation used.

Tukey's Results for Model and Database Interaction.

A Tukey post-hoc test was conducted to determine which model and database combinations exhibited significant differences in performance. It's important to mention that in Table 16 'Mean' represents the average F1-score obtained from the cross-validation folds, while the 'Groups' indicate that combinations sharing the same letter are not significantly different, whereas those with different letters show significant differences.

Table 16

Tukey's Results for Model and Database Interaction with a confidence level of 95%.

Radius*Model	Mean	Groups		
R3B2048_SMOTE XGBoost	0.742460	A		
R3B2048_SMOTE Random Forest	0.726152		B	
R2B2048_SMOTE Random Forest	0.712678		B	
R2B2048_SMOTE XGBoost	0.712331		B	
R3B2048_SMOTE Tab Net	0.272458			C
R2B2048_SMOTE Tab Net	0.244475			C

Means that do not share a letter are significantly different.

According to the results, the XGBoost model trained on the R3B2048 SMOTE database had the greatest mean performance (0.742460) and is in Group A, meaning it is superior to and substantially different from the other model– radius on molecular representation combinations. This analysis suggests that the performance is significantly affected by both model and database selection, with certain models being more sensitive to changes in the dataset than others.

Group B includes Random Forest (both R3B2048 and R2B2048 SMOTE) and XGBoost on R2B2048 SMOTE. While there are no statistically significant differences between them, they all perform noticeably worse than XGBoost on R3B2048 SMOTE.

However, regardless of whether it was trained on R3B2048 or R2B2048 SMOTE, TabNet performs the worst and is in Group C, which is quite different from Groups A and B.

Testing with Cacao Database

The results obtained from the validation of robustness and predictive capability of the best-performing model (XGBoost with R3B2048_SMOTE) are presented in Table 17:

Table 17

*Performance metrics of the XGBoost model with the Cacao database with MLSMOTE (*R3B2048_SMOTE*) with n=300 considered just for testing*

	Hamming Loss	**AUROC**	**Top 2 TSS**	**Top 5 TSS**	**Precision**	**Recall**	**F1-score**
R3B2048 n=300	0.099	0.944	0.164	0.118	0.042	0.758	0.079

In the Cacao database, the Hamming Loss value of 0.099 is low, indicating that most predictions do not contain individual label errors. The AUROC of 0.944 is as good as the other datasets, suggesting that the model effectively distinguishes between positive and negative classes in this new dataset. However, the Top-2 TSS (0.164) and Top-5 TSS (0.118) values are quite low, indicating that the model does not correctly assign the most probable aromas in the first prediction positions. The low precision (0.042) suggests that the model's predictions are not accurate in a significant number of cases. Despite this, the recall (0.758) is high, indicating that the model can identify a significant number of important labels. This imbalance suggests that the model is conservative in assigning labels when it is very confident, which can be problematic in applications that require a broader coverage of the aromatic profiles. Lastly, the F1-score (0.079) indicates weak overall performance, reflecting a poor balance between precision and recall in multi-label classification.

Table 18

*Performance metrics of the XGBoost model with the Cacao database with MLSMOTE (*R3B2048_SMOTE*) with n=300 considered for training and testing.*

	Hamming Loss	**AUROC**	**Top 2 TSS**	**Top 5 TSS**	**Precision**	**Recall**	**F1-score**
R3B2048 n=300	0.005	0.663	0.248	0.251	1.000	0.155	0.269

To further explore the model's behavior, the cacao database was also used as both the training and testing set under the same model and parameters (R3B2048_SMOTE, n=300), with the results shown in Table 18. Compared to the scenario where the cacao data was used solely for testing, some key differences emerged. The Hamming Loss dropped significantly from 0.099 to 0.005, indicating fewer errors per label. Precision improved drastically to 1.000, meaning that all predicted labels were correct, though this came at the expense of recall, which dropped to 0.155. This shift suggests the model became extremely cautious, only predicting labels when it was highly certain. Despite a slight improvement in Top-2 TSS (0.248) and Top-5 TSS (0.251), the F1-score increased only modestly to 0.269, highlighting the tradeoff between precision and recall. These results reinforce the observation that while training on the same dataset improves certain metrics, it can also reduce the model's ability to generalize and predict a broader range of relevant labels, limiting its robustness in real world applications.

DISCUSSION

The present study investigated the efficacy of machine learning models in predicting aromatic profiles based on molecular characteristics. Three different algorithms: TabNet, Random Forest, and XGBoost, as well as thorough preprocessing, data augmentation strategies (MLSMOTE), detailed feature extraction using Morgan Fingerprints with varying radius were all part of the evaluation.

Initially, the dataset showed a significant label imbalance, with a Mean Imbalance Ratio (MeanIR) of 6.1, indicating severe imbalance mostly as a result of the "Sweet" descriptor being overrepresented. With the implementation of MLSMOTE, this imbalance was successfully decreased to a MeanIR of 3.7633, improving the predictive performance for all models. This corroborates Sukhwani's (2021) assertion that improving generalization in imbalanced multi-label scenarios requires the production of synthetic data. All models showed better metrics such as AUROC and F1-score, proving that MLSMOTE is a helpful method for addressing class imbalance issues.

Among the three models, XGBoost achieved the highest mean performance metrics, according to the two-way ANOVA and the Tukey HSD test. It also showed higher overall performance, particularly on the R3B2048 dataset with MLSMOTE. This outcome supports earlier research that demonstrated XGBoost's durability and efficiency in processing high-dimensional chemical datasets because of its efficient handling of complex relationships between molecular structure and sensory descriptors (Kou et al., 2023).

On the other hand, TabNet demonstrated much worse performance metrics in both radius-based datasets. Although its sequential attention processes and interpretability benefits potentially improve feature usage (Arik & Pfister, 2021), its predictive efficacy was limited in this specific sensory prediction scenario. Moreover, stratified cross-validation indicated that

TabNet barely improved predictive metrics like AUROC and accuracy, but it had significant recall issues, which would limit its usefulness in applications where sensitivity to true positives is crucial.

Another significant contribution of the present research involved evaluating the impact of molecular feature extraction parameters on prediction accuracy. Minimal differences were seen when comparing Morgan Fingerprints with radius 2 and 3. However, the slightly improved metrics with radius 2 indicate that this parameter choice could offer a better compromise between computational efficiency and informational content, resonating with findings by Zhou and Skolnick (2024).

Cross-validation analysis confirmed the stability and robustness of Random Forest and XGBoost models, consistently demonstrating high predictive performance and reliability across data partitions. Additionally, the 5-fold stratified cross-validation (SCV) validated their generalizability by maintaining consistent label distributions, confirming the suitability of these models for practical applications in the fragrance and food industries (Szeghalmy & Fazekas, 2023).

The results of using the external cacao database to validate the optimum model (XGBoost with R3B2048_SMOTE) were not quite consistent. Its accuracy in diverse classification tasks was supported by a low Hamming Loss (0.099) and high AUROC (0.944), which demonstrated great overall discriminating capabilities. Nevertheless, low Top-2 TSS (0.164) and Top-5 TSS (0.118) values revealed the model's limitations in accurately ranking the most probable aromatic descriptors. Additionally, although having a high recall (0.758), the poor accuracy (0.042) demonstrated a cautious approach to label assignment, which limited its usefulness. Last but not least, a low F1-score (0.079) revealed a poor recall-precision balance, suggesting that multi-label categorization scenarios might need a lot more work.

Limitations & Recommendations

Several limitations must be acknowledged. The results were limited since the technical equipment utilized to run the simulations lacked the computing capacity and processors necessary to handle more complicated models. Due to time constraints, SMILES was represented textually; however, the introduction of molecular graphs or learnt embeddings might enhance the representation of molecules, enabling the model to capture more intricate structural interactions that impact the perception of odor. As a result, it's possible that certain crucial structural elements of olfactory senses are not fully represented.

Another limitation was the datasets. Using pre-existing databases raises the possibility of biases since they can naturally include overrepresented groups of chemical structures or aromatic descriptions. Also, the size of the dataset, which, for efficient purposes, required the application of MLSMOTE to introduce artificial data and balance the dataset. Finally, external validation with the cacao database revealed considerable limitations in precise ranking accuracy (Top-2 and Top-5 TSS) and overall precision, highlighting the model's limited transferability and generalization capacity to unseen, specialized datasets.

In order to eliminate the necessity for artificial data augmentation techniques, future research could focus on creating bigger, naturally balanced datasets from the beginning. Furthermore, this study only included the 90% of the relevant descriptors, resulting in 44 descriptors out of the original 476; therefore, future research should include a wider variety of descriptors. Increasing the number of descriptors and chemical structures might improve model training, boost performance metrics, and improve prediction accuracy.

Verifying the statistical power of an ANOVA is essential, as low power can make results less robust and increase the risk of Type II errors. According to Alade et al. (2024), the number of imputations, effect size, and missing data all have an impact on power in a two-factor ANOVA. Although the study results were encouraging, we were unable to verify the power of our ANOVA analysis due to time constraints. Therefore, power analysis should be included in future research planning.

Overall, the present study effectively demonstrated the potential of machine learning methods in predicting sensory profiles. It highlighted the importance of data preprocessing, effective feature extraction, and robust algorithmic approaches, notably exemplified by the XGBoost model. Despite existing limitations, continuous methodological improvement and integration of larger datasets have great potential to increase predictive accuracy and practical use in the aromatic profile prediction process.

CONCLUSIONS

This study demonstrated viability of using machine learning techniques (Random Forest, TabNet, and XGBoost) for the prediction of aroma descriptors from the molecular structures (SMILES) after using an encoding of Morgan Fingerprints. Aditionally, among the tested models with the different databases, XGBoost with the R3B2048 database balanced with MLSMOTE techniques, showed the best predictive performance.

Data augmentation with the creation of synthetic data using MLSMOTE helped mitigate the imbalance in the data sets, improving the models' generalizability and predictive stability, demonstrated in some metrics such as AUROC, Precision, and F1-Score in all the models of machine learning. Additionally, the stratified cross-validation confirmed the reliability and the robustness of the XGBoost model, over the other models of Random Forest and TabNet that were also tested. This emphasizes the high capability of this algorithm to make a correct multi-label classification in odor prediction.

Furthermore, the Two-Way ANOVA helped to have solid statistical validation, confirming that there is a significant interaction between the databases with different parameters of the Morgan fingerprints radius (two and three) with the model type, showing that it's necessary to select carefully a database and machine learning algorithm. Moreover, the Tukey's test gave statistical validation that the best model to predict is the XGBoost model.

Finally, validation with the test of an independent database (cacao database), allowed to prove the utility and applicability of the XGBoost model in other areas. This helped to determine what is working in the algorithm and what else its necessary to improve, to be able to apply this as an efficient alternative of odor prediction and expand it to use to more industries, like fragrances and flavors.

BIBLIOGRAPHY

Charte, F. & Rivera Rivas, Antonio & Del Jesus, María José & Herrera, Francisco. (2015). MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation. Knowledge-Based Systems. -. 10.1016/j.knosys.2015.07.019.

REFERENCES

Acharya, N. (2024). Understanding precision, recall, F1-score, and support in machine learning evaluation. *Medium*. https://medium.com/%40nirajan.acharya777/understanding-precision-recall-f1-score-and-support-in-machine-learning-evaluation-7ec935e8512e

Alade, S. P., Okolo, A., Akinrefon, A., & Dike, I. J. (2024). *Estimating statistical power for a two-factor ANOVA design with missing data through multiple imputation*. FUDMA Journal of Sciences, 8(4), 291–295. https://doi.org/10.33003/fjs-2024-0804-2669

Arik, S. Ö., & Pfister, T. (2021). TABNET: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(8), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

Arık, S. Ö., & Pfister, T. (2021). *TabNet: Attentive interpretable tabular learning*. Association for the Advancement of Artificial Intelligence. https://arxiv.org/abs/1908.07442

Bénédict, G., Koops, V., Odijk, D., & de Rijke, M. (2021). SigmoidF1: A smooth F1 score surrogate loss for multilabel classification. *arXiv preprint arXiv:2108.10566*.

Camilo, L. J. J., Ramirez, J., Amaya-Gómez, R., & Nicolas, R. R. (2023). Gin's Aromatic Palette: a dataset of 160 botanicals, volatile compounds, and aromatic descriptors. *Mendeley Data*. https://doi.org/10.17632/h6y25vxxwd.1

Clare, A., & King, R. D. (2001). *Knowledge discovery in multi-label phenotype data*. In L. De Raedt & A. Siebes (Eds.), *PKDD 2001, LNAI 2168* (pp. 42–53). Springer-Verlag. https://doi.org/10.1007/3-540-44794-6_4

Firmenich (2021). AIcrowd | Learning to Smell | Challenges. https://www.aicrowd.com/challenges/learning-to-smell

FlavorDB2. (2025). https://cosylab.iiitd.edu.in/flavordb2/entity_details?id=283

Gqamana, P. P. (2024). Entry-level cheminformatics of Bile Acids via PubChemPy searches and RDKit analysis in Python. Medium. https://medium.com/%40putuma.gqamana/entry-level-cheminformatics-of-bile-acids-via-pubchempy-searches-and-rdkit-analysis-in-python-007e17491f43

Keller, A., Gerkin, R. C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., Mainland, J. D., Ihara, Y., Yu, C. W., Wolfinger, R., Vens, C., Schietgat, L., De Grave, K., & Meyer, P. (2017). Predicting human olfactory perception from chemical features of odor molecules. Science, 355(6327), 820–826. https://doi.org/10.1126/science.aal2014

Kou, X., Shi, P., Gao, C., Ma, P., Xing, H., Ke, Q., & Zhang, D. (2023). *Data-driven elucidation of flavor chemistry*. *Journal of Agricultural and Food Chemistry, 71*(18), 6789–6802. https://doi.org/10.1021/acs.jafc.3c00909

Kundu, R. (2022). F1 Score in Machine Learning: Intro & Calculation. *V7*. https://www.v7labs.com/blog/f1-score-guide

Nanda, A., Mohapatra, B. B., Mahapatra, A. P. K., Mahapatra, A. P. K., & Mahapatra, A. P. K. (2021). *Multiple comparison test by Tukey's honestly significant difference (HSD):*

Do the confident level control type I error. International Journal of Statistics and Applied Mathematics, 6(1), 59–65. https://doi.org/10.22271/maths.2021.v6.i1a.636

Noorhalim, N., Ali, A., & Shamsuddin, S. M. (2019). Handling imbalanced ratio for class imbalance problem using SMOTE. In L.-K. Kor, A. R. Ahmad, Z. Idrus, & K. Mansor (Eds.), Proceedings of the third international conference on computing, mathematics and statistics (iCMS2017). Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-7279-7_3

Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Frontiers in Nanotechnology, 4, 972421. https://doi.org/10.3389/fnano.2022.972421&#8203;:contentReference[oaicite:0]{index=0}

Pyrfume. (2024). *GitHub - pyrfume/pyrfume-data: Provenance and tracking for Pyrfume data sources*. GitHub. https://github.com/pyrfume/pyrfume-data/tree/main

Saini, K., & Ramanathan, V. (2022). Predicting odor from molecular structure: a multi-label classification approach. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-18086-y

Sánchez-Lengeling, B., Wei, J. N., Lee, B. K., Gerkin, R. C., Aspuru-Guzik, A., & Wiltschko, A. B. (2019). Machine learning for scent: Learning generalizable perceptual representations of small molecules. arXiv preprint. https://arxiv.org/abs/1910.10685

Sanz, F. (2024, July 6). Cómo funciona el algoritmo XGBoost en Python. The Machine Learners. https://www.themachinelearners.com/xgboost-python/#comment-49

Sisson, L., Barsainyan, A. A., Sharma, M., & Kumar, R. (2024). Olfactory label prediction on aroma-chemical pairs. CSIR-Central Scientific Instruments Organisation. https://arxiv.org/abs/2312.16124

Sukhwani, N. (2020, June 15). *Handling data imbalance in multi-label classification (MLSMOTE)*. Medium. https://medium.com/thecyphy/handling-data-imbalance-in-multi-label-classification-mlsmote-531155416b87

Sumalatha, U., Prakasha, K. K., Prabhu, S., & Nayak, V. C. (2024). Enhancing Finger Vein Recognition with Image Preprocessing Techniques and Deep Learning Models. *IEEE Access*, *12*, 173418–173440. https://doi.org/10.1109/access.2024.3498601

Sureiman, O., & Mangera, C. M. (2020). Conceptual framework on conducting two-way analysis of variance. *Journal of the Practice of Cardiovascular Sciences, 6*(3), 207-215. https://doi.org/10.4103/jpcs.jpcs_75_20

Szeghalmy, S., & Fazekas, A. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. Sensors, 23(4), 2333. https://doi.org/10.3390/s23042333

Weininger, D. *(2022). Daylight Theory: SMILES*. Chemical Information Stystems https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html?utm_source=chatgpt.com

Yoon, S., & Lee, W. (2022). Application of true skill statistics as a practical method for quantitatively assessing CLIMEX performance. *Ecological Indicators*, *146*, 109830. https://doi.org/10.1016/j.ecolind.2022.109830

ANNEXES

ANNEX 1: LINK ONE DRIVE CODES

<u>Tesis 2025</u>

**The step-by-step codes are on the "Data Bases" folder*

ANNEX 2: LINK DATABASES PYRFUME GITHUB

https://github.com/pyrfume/pyrfume-data/tree/main/

ANNEX 3: RESULTS OF THE IRLBL METRIC BEFORE AND AFTER THE DATA AUGMENTATION

IRLbl_with_MLSMOTE		IRLbl_without_MLSMOTE	
Vanilla-like	48.698413	Musky	15.934211
Winey	39.844156	Vanilla-like	14.163743
Leafy	39.333333	Leafy	12.294416
Camphoraceous	36.963855	Aldehydic	12.019851
Gourmand	36.52381	Camphoraceous	10.716814
Tobacco-like	36.094118	Tobacco-like	10.622807
Musky	35.674419	Gourmand	9.707415
Aldehydic	32.294737	Powdery	9.460938
Powdery	30.68	Winey	9.071161
Ethereal	30.68	Metallic	8.807273
Rancid	28.943396	Microbiological	8.238095
Chocolatey	28.146789	Terpenic	8.0599
Metallic	27.890909	Meaty	7.876423
Microbiological	25.147541	Alliaceous	7.676704

Meaty	23.782946	Roasted	7.592476
Dry	23.6	Chocolatey	7.545171
Alliaceous	21.158621	Dry	7.384146
Phenolic	19.417722	Rancid	6.718447
Foul-smelling	19.175	Phenolic	6.315515
Medicinal-like	18.593939	Cabbage-like	6.290909
Terpenic	16.147368	Ethereal	5.606481
Roasted	14.75	Sulfuric	5.358407
Cabbage-like	12.783333	Medicinal-like	5.293989
Sulfuric	10.125413	Waxy	5.142251
Baked	9.5875	Fresh	5.040583
Molten	9.268882	Baked	4.968205
Nutty	8.617978	Bitter	4.844
Minty	8.382514	Foul-smelling	4.435897
Septic	8.314363	Minty	4.395644
Musty	8.116402	Molten	4.245399
Waxy	7.826531	Nutty	4.143713
Fresh	7.575309	Septic	3.993405
Sickening	6.941176	Musty	3.766719
Fatty	5.12187	Lemon	3.021834
Earthy	4.932476	Sickening	3.008696
Bitter	4.869841	Fatty	2.771167
Lemon	4.763975	Earthy	2.615551
Chemical	4.606607	Chemical	2.218965
Spicy	3.105263	Spicy	1.849561
Woody	1.951654	Woody	1.505283
Herbaceous	1.437002	Fruity	1.232884
Fruity	1.413825	Herbaceous	1.216169
Fragrant	1.203609	Fragrant	1.098911

Sweet	1	Sweet	1

ANNEX 4: COMBINATIONS AND DATA USED FOR ANOVA ANALYSIS

Model	Data Base	Fold number	F1 Score MICRO
Random Forest	R2B2048	1	0.7068
Random Forest	R2B2048	2	0.7158
Random Forest	R2B2048	3	0.7176
Random Forest	R2B2048	4	0.7059
Random Forest	R2B2048	5	0.7171
Random Forest	R3B2048	1	0.7169
Random Forest	R3B2048	2	0.7326
Random Forest	R3B2048	3	0.7232
Random Forest	R3B2048	4	0.7214
Random Forest	R3B2048	5	0.7363
Tab Net	R2B2048	1	0.2138
Tab Net	R2B2048	2	0.2451
Tab Net	R2B2048	3	0.1764
Tab Net	R2B2048	4	0.1522
Tab Net	R2B2048	5	0.3407
Tab Net	R3B2048	1	0.2559
Tab Net	R3B2048	2	0.1819
Tab Net	R3B2048	3	0.3456
Tab Net	R3B2048	4	0.2235
Tab Net	R3B2048	5	0.2959
XGBoost	R2B2048	1	0.7107
XGBoost	R2B2048	2	0.7123
XGBoost	R2B2048	3	0.7175
XGBoost	R2B2048	4	0.7094

XGBoost	R2B2048	5	0.7117
XGBoost	R3B2048	1	0.7370
XGBoost	R3B2048	2	0.7433
XGBoost	R3B2048	3	0.7502
XGBoost	R3B2048	4	0.7356
XGBoost	R3B2048	5	0.7460

ANNEX 5: LINK DATABASE CHEMICAL COMPONENTS AND AROMA PROFILES

https://github.com/ggvvmm/Aroma-Profiles

ANNEX 6: FORMULAS OF EVALUATION METRICS

Performance Metric	Formula	Leyend
Hamming Loss	$\frac{mismatch\_count}{N * L}$	$for\ i = 1\ to\ N\ and\ for\ j = 1\ to\ L\ if\ {y^{(i)}}_j \neq {\hat{y}^{(i)}}_j\ then\ mismatch\_count = mismatch\_count + 1$
AUROC	$\int TPR(FPR)\ dFPR$	TPR: True Positive Rate FPR: False Positive Rate
Top 2 TSS	$\frac{NS_{Top\ 2}}{TS}$	$NS_{Top\ 2}$: Number of samples where the top 2 predictions include at least one true label TS: Total number of samples
Top 5 TSS	$\frac{NS_{Top\ 5}}{TS}$	$NS_{Top\ 5}$: Number of samples where the top 5 predictions include at least one true label TS: Total number of samples

Precision	$\frac{TP}{TP + FP}$	TP: True Positive FP: False Positive
Recall	$\frac{TP}{TP + FN}$	TP: True Positive FP: False Negative
F1-Score	$2 \times \frac{Precision\ \times\ Recall}{Precision +\ Recall}$	-