Sensory Characterization of Mezcal Using Free Choice Profiling and Machine Learning Tools

Antonieta Martínez-Velasco[1], Sergio Erick García Barrón[2], Claudia Ariadna Acero Ortega[2], Socorro Josefina Villanueva Rodríguez[3] and Enrique Herrera López[4]

[1]Facultad de Ingeniería, Universidad Panamericana, México
[2]ESDAI, Universidad Panamericana, México
[3]Unidad de Tecnología Alimentaria, Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco, A.C., Jalisco, México
[4]Laboratorio para la Innovación en Bioelectrónica e Inteligencia Artificial, LINBIA, Unidad de Biotecnología Industrial, Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco, A.C., México

amartinezv@up.edu.mx (corresponding author)
sergiogarcia@up.edu.mx
cacero@up.edu.mx
svillanueva@ciatej.mx
eherrera@ciatej.mx

Abstract: This study characterized the sensory profiles of four regional mezcals by integrating free-choice profiling with machine learning. A total of 1,148 consumers across four Mexican cities described the samples using their own vocabulary and rated attribute intensities. Machine learning analysis revealed significant correlations between sensory descriptors and demographic variables—such as age, gender, and origin—identifying detailed consumer preference patterns. The findings enable producers to customize products for specific market segments, enhancing their commercial strategies. By supporting traditional producers, this research also promotes sustainable economic growth in rural communities, aligning with the objectives of SDG 8.

Keywords: Consumer description, Agave spirits, Intelligent systems, Sensory evaluation

1. Introduction

The sensory characteristics of a product are decisive for both its quality and its identity, particularly when it comes to traditional beverages (García-Barrón et al. 2021). These characteristics are the result of the production process, raw materials, handling, and storage, among others (Lücke et al. 2019). On the other hand, there are non-sensory factors that influence the perception of these characteristics, such as familiarity with the product, consumer sensitivity, and the sociocultural environment (Köster, 2009). The type and number of words used to describe the odor, aroma, mouthfeel, texture, and appearance reflect different factors such as the culture to which consumers belong (Chen and Antonelli, 2020).

There are different methodologies available for carrying out the sensory characterization of a product. Among these tools is the Free Choice Profile (FCP) (Williams and Langron, 1984; Varea and Ares, 2014). The FCP is based on the idea that consumers can describe the sensory characteristics of a product using their own vocabulary (Varela and Ares, 2014). Through FCP, it is possible to understand how consumers describe and differentiate food and beverages, as intensity data is obtained, providing quantitative information and allowing different aspects related to the product to be quantified, in addition to gathering information on cognitive perception directly from consumers (Tarrega and Tarancón, 2014).

However, one of the main drawbacks of the FCP is the complexity of analyzing and interpreting sensory vocabulary, differences in how participants use the scale, which increases interindividual differences among participants and, therefore, the difficulty of analyzing them (Varela and Ares, 2014). In this regard, machine learning (ML) tools have begun to be used to analyze the sensory characteristics of different alcoholic beverages, such as wine. These tools are well-suited for extracting useful information from noisy, uncertain, and nonlinear dynamic sensory data (Wang et al., 2022).

On the other hand, mezcal is a traditional alcoholic beverage made from the fermentation and distillation of agave juices, which has great cultural and economic importance (García-Barrón et al., 2017). Its production and consumption have increased significantly, particularly since it obtained the Denomination of Origin, where sensory characteristics play a significant role (Lazo et al. 2025). In this regard, several studies have been conducted on the sensory characterization of mezcal using trained judges (Garcia-Barrón et al. 2012; Mosqueda-

Balderas et al. 2018; Lazo et al. 2025; Vazquez-Lecona et al., 2025). However, while these studies are important for ensuring product quality and identity, there is little information on consumer perception and level of liking (García-Barrón, 2023).In this regard, the objective of this study was to analyze the description of sensory attributes obtained through the free-choice profile and processed with ML tools. Additionally, the descriptive data were correlated with the level of liking.

2. Methodology

This study employed a quantitative research design. The methodological framework was structured into three main phases: (1) data acquisition and preprocessing, (2) classification, and (3) prediction using machine learning techniques. This approach integrates established sensory science protocols with advanced computational tools to extract meaningful insights from complex consumer data.

2.1 Data Acquisition and Preprocessing

Sensory data were collected using the Free Choice Profiling (FCP) protocol, selected for its ability to capture authentic consumer vocabulary without the constraints of a predefined lexicon. A total of 1,748 consumers, all regular mezcal drinkers, were recruited for the panel. Each participant evaluated four distinct mezcal samples. For each sample, consumers generated their own descriptors and rated the perceived intensity of each attribute on a 9-point scale. This process yielded a raw data matrix in which rows corresponded to consumers and columns to an extensive, idiosyncratic set of sensory descriptors.

Given the high dimensionality and variability of FCP data, preprocessing was essential to ensure analytical robustness. Data cleaning involved identifying and addressing missing values, duplicates, outliers, irrelevant features, and inconsistent formats. These procedures are critical in machine learning applications, as they reduce noise, optimize feature representation, and mitigate the risks of model overfitting or underfitting.

The raw dataset was initially examined through descriptive statistics and exploratory visualizations to assess its structure and detect anomalies. Rows or columns with substantial missing values were considered for elimination, provided their removal did not compromise dataset representativeness. The goal of this stage was to transform the raw data into a structured and reliable dataset suitable for downstream analysis.

Subsequently, all odor, flavor, and sensation descriptors were standardized and grouped into predefined semantic categories using a keyword dictionary. Non-matching descriptors were classified as “other,” and a record was maintained for further analysis. Finally, mezcal samples were differentiated by brand, enabling consumer behavior to be analyzed within a product-specific sensory framework.

As part of the data preprocessing, the descriptors underwent a standardization process. Three researchers conducted an independent analysis of the descriptors to ensure consistency in their classification. Subsequently, through consensus, the descriptors used in the final analysis were defined. Those with a frequency of less than 2% or that could not be classified appropriately were eliminated.

2.2 Classification

The second stage involved the classification of sensory descriptors into relevant categories in order to simplify analysis and reduce dimensionality. After preprocessing, descriptors were clustered into odor, flavor, and sensation groups, each representing a higher-level construct of consumer perception. This classification step provided a structured input for subsequent machine learning models.

To evaluate classification performance, supervised machine learning algorithms were implemented. Among these, Support Vector Machines (SVM), Decision Trees (DT), and k-Nearest Neighbors (k-NN) were applied, as they are well-suited to handling nonlinear relationships and high-dimensional datasets. Cross-validation (k-fold, with $k = 10$) was employed to assess model generalizability and to prevent overfitting.

Model performance was evaluated using accuracy, precision, recall, and F1-score, ensuring a balanced assessment of classification effectiveness across descriptor categories. The outcomes of this step provided a clearer understanding of how consumers differentiate sensory attributes across mezcal brands.

2.3 Prediction

The third phase focused on predicting consumer satisfaction levels, operationalized as the target variable “level of liking.” Using the cleaned and categorized dataset, machine learning regression models were trained to identify the most influential sensory attributes and to forecast consumer preferences.

Multiple algorithms were tested, including:

Support Vector Regression (SVR): chosen for its robustness in handling nonlinear relationships.

Random Forest Regression (RFR): effective in capturing complex variable interactions and reducing variance.

Multiple Linear Regression (MLR): included as a baseline model for comparison.

The predictive models were trained on 70% of the dataset and validated on the remaining 30%, using stratified sampling to preserve representativeness across brands and consumer profiles. Model performance was assessed using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 (coefficient of determination).

The predictive framework allowed us not only to identify the sensory attributes most strongly associated with consumer liking but also to generate practical insights into how producers can optimize sensory profiles to meet consumer expectations.

2.4 Evaluation Metrics

For regression tasks, evaluation focused on the magnitude of prediction errors. Two primary indicators were applied:

Mean Absolute Error (MAE): the average of the absolute differences between predicted and actual values.

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

Mean Squared Error (MSE): the average of squared differences between predicted and actual values, penalizing larger errors more heavily.

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

These metrics provided a robust assessment of predictive accuracy and enabled comparative evaluation across alternative regression models.

3. Results and Discussion

The final dataset comprised responses from 1,148 participants. A detailed overview of the key demographic characteristics of the sample is presented below.

The sample was predominantly male, with men representing nearly two-thirds of respondents (65.85%), compared to women (34.15%) (Fig. 1).

In terms of age distribution, participants were primarily young adults, with almost 80% falling between 18 and 34 years of age. Specifically, 38.33% were between 18–24 years, 40.42% between 25–34 years, 11.50% between 35–44 years, 5.92% between 45–54 years, and only 3.83% were aged 55 years or older (Fig. 2).

Regarding educational attainment, the largest group of respondents held a Bachelor's degree (37.63%), followed by those with a high school diploma (33.10%). Participants with a Master's degree represented 14.63%, those with a technical degree 9.06%, and those with middle school education 3.14%. A small proportion held a doctoral degree (2.44%).

Participants were distributed across four principal urban centers. The largest proportion resided in Oaxaca (33.80%), followed by Durango (24.04%), Mexico City (CDMX) (21.25%), and the Guadalajara Metropolitan Area (ZMG) (20.91%) (Fig. 3).

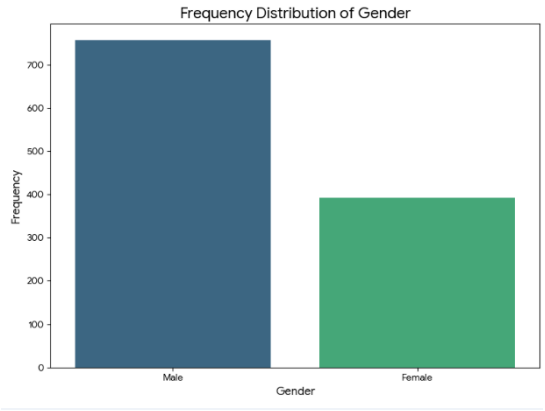

Figure 1: Distribution by Gender

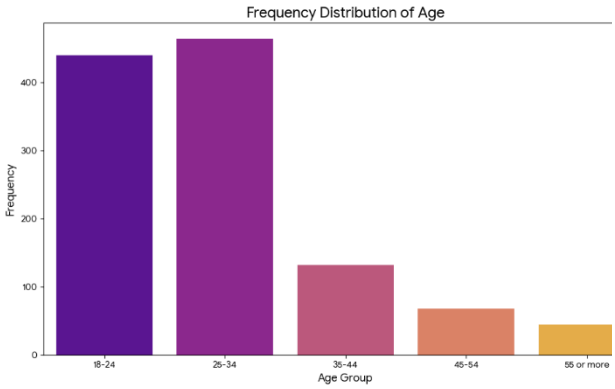

Figure 2: Distribution by Age group

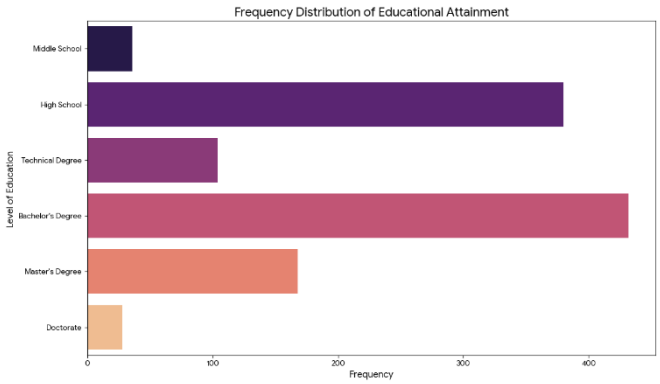

Figure 3: Distribution by Educational level

This study presents the sensory analysis of four mezcal brands. The analysis began with the estimation of Spearman's rank correlation coefficients to explore the associations between sensory descriptors and consumer preferences. These correlations revealed consistent relationships between specific aromas and overall consumer taste, providing a statistical basis for subsequent predictive modeling. Correlations between the variables age, gender, and origin were also measured with respect to the level of linking feature. The correlations show that the feature most strongly correlated with the target variable is origin, and the gender variable has a negative correlation with the target variable (Table 1).

Building on these findings, the study addressed the machine learning classification task by systematically comparing the performance of three widely used algorithms: Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) (Table 2). Each model was trained to classify consumer responses based on sensory descriptors, allowing for an objective assessment of their predictive capacity.

The results demonstrated that SVM consistently achieved superior performance, reflecting its robustness in handling high-dimensional and nonlinear data structures typical of Free Choice Profiling datasets. Random Forest also performed well, benefiting from its ensemble structure and ability to capture complex attribute interactions, though with slightly lower accuracy compared to SVM. In contrast, Logistic Regression, while offering interpretability and simplicity, exhibited the weakest predictive power, underscoring the limitations of linear models for this type of heterogeneous sensory data.

Overall, these results highlight the potential of advanced machine learning approaches to complement traditional sensory analysis by uncovering latent patterns and improving classification accuracy. By demonstrating that nonlinear models outperform classical linear approaches, this work provides a methodological basis for applying machine learning in the study of traditional beverages, where consumer perceptions are highly variable and multidimensional.

Table 1: Spearman´s correlations for target variable (Level of linking)

Feature 1	Feature 2	Spearman´s Correlation
Agave odor	level of linking	0.148
Wood odor	level of linking	0.11
Vegetable odor	level of linking	0.098
Fruity flavor	level of linking	0.075
Gender	level of linking	-0.117
Origin	level of linking	**0.055**
Age	level of linking	0.029

The target variable "level of liking" was selected, as it directly reflects consumer preference. To evaluate predictive performance, three experiments were conducted, each employing a different algorithm: Support Vector Machine (SVM), Random Forest (RF), and Linear Regression (LR). The results of the classification task show that SVM achieved the best performance, yielding the lowest Mean Squared Error (MSE) value, and thus demonstrating superior predictive accuracy compared to the other models.

Table 2: Models Evaluation by means MSE

Model	MSE
SVM	4.238
Random Forest	4.576
Linear Regression	6.480

The developed classification model proved to be highly effective in predicting the consumer's level of liking. This variable was assessed using the 9-point hedonic scale, a well-established instrument for quantifying degrees of pleasure or displeasure in sensory and consumer research. The scale provides a standardized framework that allows for sensitive and reliable measurement of consumer preferences, with verbal anchors ranging from "dislike extremely" to "like extremely" (Table 3).

Analysis of the response distribution revealed that the majority of participants reported ratings clustered around the mid-to-high range of the scale, indicating a generally positive perception of the evaluated mezcal samples. Lower scores were less frequent, suggesting that negative evaluations were comparatively rare within the sample. This pattern reinforces the predictive capacity of the model, as it successfully identified the sensory attributes most strongly associated with higher hedonic scores.

Table 3: Hedonic standardized scale (9 points)

Hedonic Scale	Level of linking
9	Like Extremely
8	Like Very Much
7	Like Moderately
6	Like Slightly
5	Neither Like nor Dislike
4	Dislike Slightly
3	Dislike Moderately
2	Dislike Very Much
1	Dislike Extremely

Further analysis revealed that the maximum value of the target variable was predicted under a multivariate condition, rather than through the independent effect of single predictors. Specifically, the highest levels of liking were observed when an optimal configuration of the most influential variables was present simultaneously. This complex interaction highlights the model's ability to capture nuanced patterns in consumer data and to define the precise conditions under which consumer engagement is maximized.

Based on the best-performing model (SVM), a prediction was generated to visualize the values that key variables must take in order to achieve the corresponding level of liking. To emphasize the influence of the variables most strongly correlated with the target outcome, the mean values for Agave, Wood, and Vegetable odors, along with Fruity flavor, are presented in Table 4.

Table 4: Average values for the variables most correlated with level of linking.

Level of linking	Agave odor	Wood odor	Vegetable odor	Fruity Flavor
9	2.210	2.259	2.445	0.703
8	2.015	0.616	1.779	0.664
7	1.591	0.943	1.174	0.584
6	2.561	0.000	0.752	0.684
5	1.448	0.154	1.288	0.546
4	1.982	1.247	1.609	0.518
3	0.000	0.325	0.857	0.000
2	0.722	0.000	1.667	0.000
1	0.000	0.171	0.000	0.000

4. Conclusions

The results of this study provide compelling evidence that odors exert a stronger influence than flavors or sensations in shaping consumer preferences for mezcal. Spearman's correlation analysis revealed that agave and wood aromas are the most positively associated with consumer satisfaction, followed by vegetable odor, whereas fruity flavor showed the weakest correlation. This pattern underscores the central role of olfactory cues in consumer perception of mezcal, highlighting those preferences are not driven by single attributes in isolation but rather by the optimal interaction of key sensory variables.

The use of the Free Choice Profile (FCP) method was critical in capturing consumers' spontaneous vocabulary, thus providing an authentic view of the consumption experience. However, the inherent variability and subjectivity of FCP data required the integration of machine learning (ML) techniques to extract meaningful insights. Among the models evaluated, Support Vector Machines (SVM) demonstrated the highest predictive accuracy, confirming their suitability for handling nonlinear and high-dimensional sensory data. This highlights the potential of ML tools to complement traditional sensory analysis by revealing latent trends and multivariate interactions that are otherwise difficult to capture.

From an applied perspective, these findings carry practical implications for producers. By emphasizing sensory attributes most strongly linked to consumer liking—particularly agave and wood aromas—mezcal producers can strategically tailor their products to meet market preferences and differentiate themselves in a competitive landscape. Such approaches not only improve product positioning but also strengthen commercial strategies targeted at specific consumer segments. Importantly, this research also supports sustainable economic development in rural communities, aligning with the objectives of Sustainable Development Goal 8 (Decent Work and Economic Growth) by providing traditional producers with actionable insights for enhancing product value.

Finally, this work contributes a novel methodological framework by integrating classical sensory analysis protocols with advanced ML techniques. This hybrid approach offers a replicable model for investigating other traditional beverages and foods, enabling researchers and producers to systematically explore consumer perception while leveraging data-driven tools for innovation and sustainability.

Ethics statement: Ethical review and approval were not required for this study in accordance with the local legislation and institutional requirements.

AI statement: The authors certify that no generative artificial intelligence (AI) tools, such as large language models or image generators, were used at any stage of manuscript preparation.

References

Abdalla, S. M., Rosenberg, S. B., Maani, N., Contreras, C. M., Yu, S., & Galea, S. (2025). Income, education, and the clustering of risk in cardiovascular disease in the US, 1999–2018: an observational study. *The Lancet Regional Health–Americas*, *44*.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

Chen, P. J., & Antonelli, M. (2020). Conceptual models of food choice: influential factors related to foods, individual differences, and society. *Foods*, *9*(12), 1898.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

García-Barrón, S. E. (2023) Caracterización sensorial del mezcal: desde los aromas hasta las preferencias. Eds. Camacho-Ruiz, R.M, Gutiérrez-Mora, A., Gschaedler-Mathis, A. *LOS AGAVES Y SUS DERIVADOS: TENDENCIAS CIENTÍFICAS, USO SOSTENIBLE Y PATRIMONIO*, 1er ed. Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco, Jalisco, Mex.

García-Barrón, S. E. 2012. Effect of the region of origin of the agave and fermentation conditions on the aromatic profile. México: Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco (CIATEJ), MSc thesis.

García-Barrón, S. E., de Jesús Hernández, J., Gutiérrez-Salomón, A. L., Escalona-Buendía, H. B., & Villanueva-Rodríguez, S. J. (2017). Mezcal y Tequila: análisis conceptual de dos bebidas típicas de México. *Revista Iberoamericana de Viticultura, Agroindustria y Ruralidad*, *4*(12), 138-162.

García-Barrón, S. E., Guerrero, L., Vázquez-Elorza, A., & Lazo, O. (2021). What turns a product into a traditional one?. *Foods*, *10*(6), 1284.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Kheirollahi, H., Chahardowli, M., & Simjoo, M. (2022). A new method of well clustering and association rule mining. Journal of Petroleum Science and Engineering, 214, 110479.

Köster, E. P. (2009). Diversity in the determinants of food choice: A psychological perspective. *Food quality and preference*, *20*(2), 70-82.

Lazo, O., García-Ortíz, A. L., Pardo, J., & Guerrero, L. (2025). Mezcal Characterization Through Sensory and Volatile Analyses. *Foods*, *14*(3), 402.

Lücke, F. K., Tannhäuser, K., Sharma, A., & Fritz, V. (2019). Development of food products with addition of rapeseed presscake fermented by Rhizopus: Sensory properties and consumer acceptance. *British Food Journal*, *121*(10), 2351-2364.

Mozqueda-Balderas, R., Delgado-Alvarado, A., Herrera-Cabrera, B. E., & Vargas-López, S. (2018). Evaluación sensorial del mezcal de la localidad de Totomochapa, Tlapa de Comonfort, Guerrero, México. *Agro Productividad*, *11*(10).

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.

Tárrega, A., & Tarancón, P. (2014). *Free-choice profile combined with repertory grid method* (pp. 157-174). CRC Press, Boca Raton, FL.

Varela, P., & Ares, G. (Eds.). (2014). *Novel techniques in sensory characterization and consumer profiling*. CRC Press, Boca Raton, FL..

Vázquez-Lecona, H. U., Ramírez-Rivera, E. J., López-Espíndola, M., Hernández-Martínez, R., & Herrera-Corredor, J. A. (2025). Development of sensory lexicon for aromas of espadin mezcal (*Agave angustifolia*) based on Analytical Hierarchy Process with trained panellists and mezcal masters. *International Food Research Journal*, *32*(1).

Wang, A., Zhu, Y., Zou, L., Zhu, H., Cao, R., & Zhao, G. (2022). Combination of machine learning and intelligent sensors in real-time quality control of alcoholic beverages. Food Science and Technology, 42, e54622.

Williams, A. A., & Langron, S. P. (1984). The use of free-choice profiling for the evaluation of commercial ports. *Journal of the Science of Food and Agriculture*, *35*(5), 558-568.