Cómo citar este artículo: A. I. Burgos-Naranjo, D. S. Vásquez-Játiva, D. O. Navarrete-Chávez, "Aplicaciones de la industria 4.0 en la estandarización del proceso productivo de las mermeladas" Revista Ingeniería, Investigación y Desarrollo, vol. 21 (1), pp. 39-46, enero. 2021.

DOI: https://doi.org/10.19053/1900771X.v21.n1.2021.13513

APLICACIONES DE LA INDUSTRIA 4.0 EN LA ESTANDARIZACIÓN DEL PROCESO PRODUCTIVO DE LAS MERMELADAS

Applications of Industry 4.0 in the standardization of jams' production process

Ángel Isaac Burgos Naranjo[1], Daniel Sebastián Vásquez Játiva[2], Danny Orlando Navarrete Chávez[3]
[1-3]Universidad San Francisco de Quito, Departamento de Ingeniería Industrial, Ecuador.
E-mail: [1]angelburgosnaranjo@gmail.com, [2]dase266@gmail.com, [3]dnavarrete@usfq.edu.ec

(Recibido febrero 10 de 2021 y aceptado junio 17 de 2021)

Resumen

El presente artículo tiene como objetivo ilustrar una de las tantas aplicaciones de la Industria 4.0 mediante el uso de procedimientos analíticos multivariados y modelos de aprendizaje automático multirrespuesta, como un camino para analizar, modelar y estandarizar las relaciones entre las distintas variables de entrada y de salida que gobiernan la formulación de las mermeladas. Este trabajo de investigación es llevado a cabo en una compañía dedicada a la producción y comercialización de productos agropecuarios, describe la metodología de estudio utilizada que permitió hallar los rangos de valores para los niveles de azúcar (°Bx) y acidez (pH) que satisfacen matemática y estadísticamente los parámetros de liberación de producto terminado definidos por la misma compañía.

Palabras clave: *consistencia, estándares, grados Brix, mermeladas, modelos, pH, variables.*

Abstract

This article aims to illustrate one of the many applications of Industry 4.0 through the use of multivariate analytical procedures and multi-response machine learning models, as a way to analyze, model and standardize the relationships between the different input and output variables that drive jams' formulation. This research work is accomplished in a company dedicated to the production and commercialization of agricultural products, it describes the methodology study used that helped to find the ranges of values for the levels of sugar (°Bx) and acidity (pH) that satisfy mathematics and statistically the finished product release parameters defined by the own company.

Key words: *consistency, standards, Brix degrees, jams, models, pH, variables.*

1. INTRODUCCIÓN

Las confituras, jaleas y mermeladas son productos alimenticios de consistencia gelatinosa derivados de la cocción de frutas, azúcares, ácidos comestibles y pectinas [1]. Surgieron como un esfuerzo por conservar los alimentos para su consumo fuera de temporada gracias al incremento de su acidez y su contenido de sólidos solubles [2]. De hecho, estos dos últimos ingredientes son de gran importancia para la calidad de dichos alimentos.

Tal es así, que se afirma que la gelificación de las mermeladas con pectinas de alto metoxilo (AM) únicamente se logra trabajando con pequeños rangos de pH (2.8–3.5) y bajos contenidos de azúcar (~600–800 g/kg) [3]. Por ello, con el fin de minimizar el efecto de la variabilidad de la pectina natural presente en las frutas, es común añadir pectina comercial (entre 0 y ~10 g/kg del producto terminado) en la producción industrial de este tipo de conservas [3].

Los niveles mínimos y máximos permitidos de sólidos solubles y agentes gelificantes en las mermeladas dependen de la legislación y las entidades de regulación y control sanitario de cada país [1]. En el caso de Ecuador, por ejemplo, en [4] se indica que el límite mínimo de contenido de sólidos solubles en producto terminado se debe encontrar entre el 60% y el 65%.

Es por este tipo de requerimientos que la estandarización en la formulación de los alimentos es uno de los desafíos más retadores que comúnmente sufre el sector alimenticio, y esta industria no es la excepción [3]. La compañía ABC, firma dedicada a la producción y comercialización de productos agropecuarios, enfrenta ciertos retos para lograr niveles uniformes de consistencia, acidez y azúcares en sus cinco sabores de mermeladas: mora, frutilla, piña, guayaba y frutimora.

La variabilidad propia de la materia prima, así como la estacionalidad y el estado de madurez de las frutas, inciden directamente en la calidad de esta clase de conservas.

Por este hecho, el presente trabajo tiene como objetivo el utilizar procedimientos analíticos multivariados y modelos de aprendizaje automático multirrespuestas, como un camino para estudiar y modelar las relaciones entre las distintas variables de entrada y de salida que gobiernan la formulación de las mermeladas.

2. ANTECEDENTES

En lo que refiere a la literatura, distintos procedimientos estadísticos han sido comúnmente requeridos para describir y estudiar los factores que influencian las propiedades químicas, físicas, sensoriales y microbiológicas de los alimentos [5]. Entre algunas de las aplicaciones estadísticas más importantes en esta industria, se destacan el uso de métodos descriptivos e inferenciales, modelos de regresión y correlación, el control estadístico de la calidad y el diseño de experimentos [6].

Procedimientos analíticos menos comunes como el análisis multivariado y el aprendizaje automático, sin duda, también se han hecho presentes. Por ejemplo, en [7] se hizo uso del análisis múltiple de varianza (MANOVA), el análisis de componentes principales (PCA) y el análisis canónico de discriminantes (CDA) con la finalidad de estimar los efectos de la variabilidad de los componentes químicos de los vinos en sus distintas respuestas sensoriales.

En [8] afirman que las publicaciones de modelos de aprendizaje automático en el estudio de los alimentos han registrado un crecimiento acelerado desde el año 2010. En [8] señalan, además, que existe un énfasis particular en la publicación de trabajos que utilizan modelos de ensamble o de conjunto, los cuales emplean múltiples algoritmos, simultáneamente, con el fin de lograr un mejor rendimiento predictivo. El bosque aleatorio, o *random forest* (RF) por su nombre en inglés, es uno de los más utilizados.

En [9] se hizo uso de *random forest*, en colaboración con PCA, para clasificar el tequila tradicional del mezcal. Más aún, en [10] se hizo uso del algoritmo de máquinas de vectores de soporte (SVM) y el de los k-vecinos más cercanos (kNN) para clasificar distintas muestras de cacao con base en su país de procedencia.

3. METODOLOGÍA

Se llevó a cabo una adaptación metodológica del proceso de modelado estadístico descrito en [11]. Esta contempla una serie genérica de actividades que comprenden desde la definición del problema y los objetivos del estudio (mismos que se detallaron en la Introducción de este trabajo), hasta el análisis y la interpretación de los resultados. El contar con una metodología de estudio garantiza que los modelos sean construidos sistemática y eficientemente [12].

3.1. Diseño del Plan de Recolección de Datos

Debido a la limitación de recursos disponibles para llevar a cabo el muestreo de la información, se hizo

uso de los datos históricos del proceso de producción de mermeladas. Estos fueron entregados por la misma compañía, ABC. El origen de los datos parte de fuentes secundarias, puesto que no fueron directamente diseñados para efectos del estudio [12].

El alcance de la información disponible comprende de enero a septiembre del año 2020, cubriendo así ocho meses de operación. Los datos disponibles reportaron una dimensionalidad de 574 filas y 22 columnas. Sin embargo, muchas de ellas cumplían únicamente con un rol informativo o referencial. Por este motivo las variables de entrada y salida de interés se resumen en las Tablas 1 y 2, respectivamente.

Tabla 1. Variables de entrada.

No.	Variable	Notación
1	Sólidos solubles materia prima	bx_{mmpp}
2	pH materia prima	pH_{mmpp}

Tabla 2. Variables de salida.

No.	Variable	Notación
1	Sólidos solubles producto terminado	bx
2	pH producto terminado	pH
3	Consistencia producto terminado	const

Los grados Brix, o porcentaje de sólidos solubles, son útiles para cuantificar el contenido de azúcar tanto en las frutas como en las mermeladas, y son medidos con un refractómetro [13]. Por otro lado, el pH indica el grado de acidez o alcalinidad del producto, y se gradúa con un potenciómetro [14]. La consistencia, por su parte, es una medida del grado de cohesión de las partículas que constituyen las mermeladas, y se la obtiene gracias a un consistómetro [15].

3.2. Análisis Exploratorio de Datos

En primer lugar, se normalizaron los datos disponibles con el fin de contar con una observación por registro, una variable por columna y una sola unidad de información por tabla [16]. Este proceso facilitó la generación de una base de datos independiente para cada uno de los cinco sabores de mermeladas. Los datos considerados por la compañía como atípicos y reprocesos fueron retirados, al igual que los registros duplicados.

Así también, los valores nulos o faltantes fueron rellenados con la mediana de sus respectivas variables, puesto que en [17] se argumenta que este criterio de imputación impacta mínimamente en la distribución original de los datos, y mantiene en gran medida sus propiedades tanto de tendencia central como de dispersión. Un resumen del análisis exploratorio de datos se muestra en la Tabla 3. Por su parte, la Tabla 4, muestra la cantidad total de datos iniciales y finales, discriminados por sabor de mermelada.

Tabla 3. Resumen del análisis exploratorio de datos.

Sabor	Duplicado	Reproceso	Atípico
Frutilla	31	10	2
Frutimora	15	0	3
Mora	6	1	8
Piña	26	8	9
Guayaba	13	4	4

Tabla 4. Cantidad de datos iniciales y finales por sabor de mermelada.

Sabor	Datos Iniciales	Datos Finales
Frutilla	166	123
Frutimora	95	77

Mora	111	96
Piña	104	61
Guayaba	91	70

Gracias a un análisis de dispersión realizado con diagramas de caja y bigote, se encontró que el producto terminado no es aprobado debido a la poca uniformidad en su consistencia. Esto concuerda con la revisión de la literatura y la motivación de este trabajo, y se aprecia en la Figura 1, para el caso del sabor de mermelada de frutilla.

Por otro lado, mediante la prueba de Shapiro-Wilk, se encontraron variables de respuesta no normales, las cuales se transformaron usando el método de potencias de Tukey. Sin embargo, no todas las distribuciones pudieron ser normalizadas por este ni por otros métodos, por lo que también se incorporaron técnicas no paramétricas para su estudio como se discutirá en los siguientes apartados.

3.3. Selección de Variables, Modelos o Algoritmos

En cuanto a los tipos de algoritmos utilizados, estos fueron clasificados en tres grandes categorías: modelos estadísticos multivariados, modelos lineales generalizados (GLM), y modelos de aprendizaje automático multirrespuesta. Los procedimientos multivariados fueron empleados debido a la presencia de correlaciones moderadas ($0.30 < r < 0.60$) entre las variables de respuesta, mismas que se observan en la Tabla 5, para el caso de la mermelada de frutilla [18].

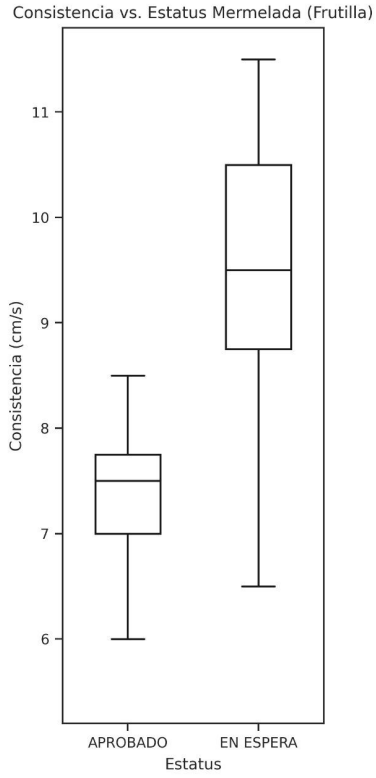

Figura 1. Diagrama de caja y bigote (mermelada de frutilla).

Tabla 5. Correlaciones entre las variables de salida para el caso del sabor de mermelada de frutilla.

Variable	const	bx	pH
const	1.00	-0.64	-0.35
bx	-0.64	1.00	0.42
pH	-0.35	0.42	1.00

Estos modelos evalúan las pruebas de hipótesis sobre la existencia de términos significativos en los algoritmos considerando dichas asociaciones. En este sentido, concretamente, se decidió hacer uso de modelos de regresión multivariada múltiple (MMR) debido a ciertas ventajas sobre otros procedimientos analíticos multivariados, como lo es su eficiencia computacional y facilidad de interpretación [18].

Para el caso de las variables de respuesta que no lograron ser transformadas a una distribución normal, se decidió emplear modelos lineales generalizados (GLM), los cuales son capaces de modelar distribuciones distintas a la normal, como las de Poisson, gamma, binomial, entre otras. Esto es algo que no es posible con las técnicas paramétricas convencionales [19].

Finalmente, se llevaron a cabo modelos de aprendizaje automático multirrespuesta como una poderosa herramienta predictiva y de gran utilidad para la compañía. Fueron dos los tipos de modelos seleccionados por su gran adaptabilidad a pequeños conjuntos de datos, tal y como lo señalan en [20]. El primero fue el modelo de árboles de decisión (Decision Tree o DT), mientras que el segundo fue el algoritmo k vecinos más cercanos (kNN).

3.4. Formulación, Evaluación y Validación de Modelos o Algoritmos

Previo a la implementación de los distintos tipos de modelos, se crearon predictores adicionales (ver Tabla 6) con la finalidad de capturar las relaciones no lineales entre las dos variables independientes disponibles. Lo anterior, cuidando que los términos añadidos no resulten en un posible sobreajuste de los modelos.

Tabla 6. Nuevas variables añadidas.

2° grado	3° grado	4° grado
pH_{mmpp}^2	pH_{mmpp}^2* bx_{mmpp}	pH_{mmpp}^2* bx_{mmpp}^2
bx_{mmpp}^2	pH_{mmpp}* bx_{mmpp}^2	-
pH_{mmpp}* bx_{mmpp}	-	-

En [21] se detalla que para modelos de regresión de mínimos cuadrados parciales (OLS) de efectos fijos, como lo es la regresión multivariada múltiple, no es necesario que los modelos sean lineales en las covariables; estos mantienen sus propiedades siempre y cuando sean lineales en los parámetros.

4. RESULTADOS

Se utilizó MMR para estudiar cada variable de respuesta normal, considerando todos los sabores de mermeladas. Con la ayuda de MANOVA se identificaron los términos estadísticamente significativos en cada modelo, así como sus efectos expresados en términos de coeficientes. El rendimiento de los modelos fue evaluado con el estadístico R^2, y su validez se corroboró con un análisis de residuales. Los resultados obtenidos se observan en las ecuaciones (1) a (6).

4.1. Frutilla

$$pH = 0.59ph_{mmpp} + 1.40 \times (R^2=0.25) \quad (1)$$

4.2. Frutimora

$$pH = -0.25bx_{mora}+0.015bx^2_{mora} + 4.11 \times (R^2 = 0.08) \quad (2)$$

4.3. Piña

$$pH = -5.44ph_{mmpp} + 0.87ph^2_{mmpp} + 10.89 \times (R^2 = 0.28) \quad (3)$$

$$const = 65.76ph_{mmpp} - 9.54ph^2_{mmpp} - 98.47 \times (R^2 = 0.29) \quad (4)$$

4.4. Guayaba

$$pH = 0.35ph_{mmpp} - 0.012bx_{mmpp} + 2.27 \times (R^2 = 0.39) \quad (5)$$

$$brix = 1.37ph_{mmpp} - 0.25bx_{mmpp} + 66.14 \times (R^2 = 0.33) \quad (6)$$

Para las variables de respuesta con distribuciones distintas a la normal, aún después de haber sido transformadas, se hizo uso de GLM. Para estimar su rendimiento se utilizó el Criterio de Información de Akaike (AIC) y la Regla de Bondad de Ajuste (GF) propuesta en [22]. Los resultados obtenidos se observan en las ecuaciones (7) a (8).

4.5. Frutilla

$$brix = 478ph_{mmpp} + 49 \times (AIC = 444.96, GF = 2.11) \quad (7)$$

4.6. Guayaba

$$const = 0.57bx_{mmpp} - 2.25 \times (AIC = 221.09, GF = 1.43) \quad (8)$$

Como una de las aplicaciones de la Industria 4.0, se corrieron modelos de aprendizaje automático multirrespuesta. Los algoritmos de árboles de decisión y kNN fueron elegidos como una herramienta útil para poder predecir los resultados de nuevas o futuras observaciones en cada uno de los cinco (05) sabores de mermeladas. Su rendimiento se aprecia en la Tabla 7, en términos del error absoluto medio (MAE).

Tabla 7. Error absoluto medio por cada sabor de mermelada.

Sabor	MAE (DT)	MAE (kNN)
Frutilla	0.91	0.77
Frutimora	0.71	0.58
Mora	0.63	0.55
Piña	0.47	0.59
Guayaba	0.53	0.51

Finalmente, se realizó un análisis de sensibilidad con los modelos MMR y GLM resultantes. Con el objetivo de identificar qué rangos de las variables de entrada satisfacen los parámetros de liberación de producto terminado, entregados por la compañía. Los rangos encontrados para el pH y los grados Brix de materia prima por cada sabor de mermelada se ilustran en la Tabla 8.

Tabla 8. Correlaciones entre las variables de salida para el caso del sabor de mermelada de frutilla.

Sabor	Variable de salida	Rango pH_{mmpp}	Rango bx_{mmpp}
Frutilla	pH	[3.22-3.90]	-
	bx	[3.32-4.16]	-
Frutimora	pH	-	[8.34-12.26]
Guayaba	pH	[2.36-4.65]	-
	const	-	[4.82-10.09]
Piña	pH	[3.46-3.91]	-

5. CONCLUSIONES

Gracias al trabajo realizado se logró modelar la formulación de cuatro de los cinco sabores de mermeladas. Lo que significó, encontrar los factores significativos en cada uno de los modelos; así también, se estimó su efecto o importancia relativa en términos de coeficientes o parámetros. Para las tres grandes categorías de algoritmos, adicionalmente, se evaluó su rendimiento mediante el cálculo de distintas métricas.

Mediante un análisis de sensibilidad, estos modelos fueron de gran utilidad para encontrar rangos de las variables de materia prima que satisfacen los parámetros de liberación de producto terminado establecidos por la compañía. Dichos rangos obedecen a las relaciones que las regresiones obtenidas lograron capturar, y deben ser validados a nivel de planta. Por ello se sugiere realizar corridas piloto entre dichos valores, y monitorear sus respuestas en la línea del proceso.

Los modelos multivariados correspondientes a los sabores de frutilla, piña y guayaba, así como los modelos lineales generalizados, revelan un ajuste relativamente bueno a las distintas variables de respuesta bajo estudio. Esto se concluye gracias a los resultados de los coeficientes de determinación y la regla de bondad de ajuste propuesta en [22].

En [23] y [24] definen que los valores recomendados de R2 deben ser iguales o superiores a 0.13 y 0.33, respectivamente, para poder argumentar que la varianza total de una variable de respuesta explicada por un modelo de regresión es la adecuada. En este sentido, se concluye que los coeficientes obtenidos se encuentran en estos límites. Más aún, las pruebas de bondad de ajuste realizadas (GF) no superan considerablemente la unidad, tal y como lo sugieren en [22].

Finalmente, como seguimiento a este trabajo, se propone a la compañía implementar un estudio a mediano plazo relacionado con una metodología estadística experimental conocida como operación evolucionaria

[25]. Esta consiste en un método de mejora continua y monitoreo para operaciones a larga escala, por lo que se ajusta de manera oportuna a las necesidades de ABC.

Cabe resaltar que este método busca garantizar la integridad y la calidad de los productos fabricados en planta, al no realizar cambios, suficientemente, drásticos en los distintos niveles de las variables de entrada del proceso. Este tipo de diseño quiere obtener relaciones causa-efecto en el estudio de la consistencia de las mermeladas, sin afectar negativamente los productos ni incurrir en grandes costos para las empresas

REFERENCIAS

[1] V. Fuster, (2004). Mermeladas y confituras. En P. López, J. Boatella, y R. Codony, Química y bioquímica de los alimentos II (pág. 105). Barcelona: Edicions Universitat Barcelona.

[2] R. Baker, D. Barrett, N. Berry y Y. Hui, (2005). Fruit preserves and jams. En D. Barrett, L. Somogyi, y H. Ramaswamy, Processing fruits: science and technology (p. 113). Boca Raton: CRC Press LLC.

[3] J. Garrido, D. Genovese y J. Lozano, (2015). Effect of formulation variables on rheology, texture, color, and acceptability of apple jelly: Modelling and optimization. *Food Science and Technology*, 325-332.

[4] Instituto Ecuatoriano de Normalización INEN. (2013). Norma para las confituras, jaleas y mermeladas. Disponible en: https://www.normalizacion.gob.ec/buzon/normas/nte-inen-2825.pdf

[5] V. De Araújo Calado, D. Granato y B. Jarvis, (2014). Observations on the use of statistical methods in food science and technology. *Food Research International*, 137-149.

[6] J. Bower, (2013). Statistical methods for food science: Introductory procedures for the food practitioner. New Jersey: John Wiley & Sons, Inc.

[7] I. Arvanitoyannis, S. Kallithraka, M. Katsota, E. Psarra y E. Soufleros, (1999). Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). *Trends in Food Science & Technology*, 321-336.

[8] G. Bagur, L. Cuadros, A. González y A. Jiménez, (2019). Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review. *Food Research International*, 25-39.

[9] S. Martinez, A. Moreno, D. Cazares & R. Winkler, (2017). Automated chemical fingerprinting of Mexican spirits derived from agave (tequila and mezcal) using direct-injection electrospray ionization (DIESI) and low-temperature plasma (LTP) mass spectrometry. Analytical Methods.

[10] F. Botchway, F. Han, X. Huang y E. Teye,(2014). Discrimination of cocoa beans according to geographical origin by electronic tongue and multivariate algorithms. *Food Analysis Methods*, 360-365.

[11] G. Shmueli, (2010). To explain or to predict? Statistical Science, 289-310.

[12] G. Shmueli y O. Koppius, (2006). Predictive analytics in information systems research. Paphos: Conference on Information Systems and Technology.

[13] W. Graham y A. MacGillivray, (1969). Brix Determination. Proceedings of The South African Sugar Technologists' Association, 215-2018.

[14] E. Álzate, R. Escobar y J. Montes (2012). Acondicionamiento del sensor de pH y temperatura para realizar titulaciones potenciométricas. *Scientia Et Technica*, vol. XVII, núm. 51, agosto, 20212, pp. 188-196. Universidad Tecnológica de Pereira. Pereira, Colombia.

[15] P. Jordano, (2000). Fruits and frugivory. En M. Fenner, Seeds: The ecology of regeneration in plant communities (pp. 125-166). Wallingford: CABI Publ.

[16] H. Wickham, (2014). Tidy data. *Journal of Statistical Software*, 1-24.

[17] P. Bruce, P. Gedeck, N. Patel y G. Shmueli, (2020). Data mining for business analytics. Hoboken: John Wiley & Sons.

[18] P. Dattalo, (2013). Analysis of multiple dependent variables. New York: Oxford University Press.

[19] P. McCullagh y J. Nelder, (1989). Generalized linear models. Boca Raton: Chapman & Hall/CRC.

[20] J. Prakash, (2018). Breaking the curse of small datasets in machine learning. Available in: Towards Data Science: https://towardsdatascience.com/breaking-the-curse-of-small-datasets-in-machine-learning-part-1-36f28b0c044d

[21] B. Kenkel, (2016). Higher order terms. Available in: Reintroduction to linear regression: http://bkenkel.com/psci8357/notes/04-higher-order.html

[22] D. Montgomery, E. Peck y G. Vining, (2015). Introduction to linear regression analysis. New Jersey: John Wiley & Sons.

[23] W. Chin, (1998). The partial least squares approach for structural equation modeling. En G. Marcoulides, Modern Methods for Business Research (pp. 295-236). London: Lawrence Erlbaum Associates.

[24] J. Cohen, (1998). Statistical power analysis for the behavioral sciences. New York: Lawrence Erlbaum Associates Publishers.

[25] Douglas C. Montgomery, (2013). Design and analysis of experiments. New York: Wiley.