

Instituto Tecnológico y de Estudios Superiores de Monterrey

Maestría en Inteligencia Artificial Aplicada - MNA

Proyecto Integrador

**Avance 0**

**Propuesta de proyecto y firma de convenios**

**Análisis de tendencias y sistema de recomendación de suplementos dietéticos mediante aprendizaje supervisado: aplicación de datos reales al proyecto de emprendimiento en Guatemala**

GLOBULIVE, S.A.

**Equipo 39**

A01321600 Ignacio Moreda Rovirosa

A01795095 Enma Beatriz Galindo Arango

A00826769 Gustavo Adolfo Benitez Leonés

**Índice**

[Antecedentes 3](#_2o037ewqubvj)

[Entendimiento del negocio 4](#_oz8uy9gd1s8v)

[Formulación del problema 4](#_mbh128g9386j)

[Contexto 4](#_ll2h1endszfb)

[Objetivos 4](#_7cn73x9l2mfg)

[Preguntas clave 5](#_sm9zsi4fvilz)

[Involucrados 5](#_qczhm5r06o8h)

[Entendimiento de los datos 7](#_6bm1tn5cyhw9)

[Descripción de los datos 7](#_58m3zagpq2wj)

[Técnica de Machine Learning 7](#_2srdypse1g23)

[Identificación de las variables 8](#_jo0fbc5rvt29)

[Anexos 9](#_4z9mk1c8p36u)

# 

# **Antecedentes**

Globulive, S.A. es un emprendimiento guatemalteco constituido legalmente bajo la razón social de Distribuidora Globulive, Sociedad Anónima, con domicilio en Ciudad de Guatemala. Su objeto social incluye la comercialización y distribución de suplementos dietéticos y alimenticios bajo la marca registrada MICROGUD®.

Cada formulación de la línea MICROGUD® se diseña con el objetivo de incorporar un probiótico, debido a la evidencia científica cada vez más amplia acerca de sus posibles beneficios. Según el National Center for Complementary and Integrative Health, los probióticos contribuyen a la conservación de una comunidad de microorganismos sanos en el intestino, con cada cepa teniendo efectos diferentes. Asimismo, crear formas farmacéuticas combinadas que sean capaces de incorporar en una única presentación varios principios activos (vitaminas, minerales y probióticos) [(NIH, 2019)](https://www.zotero.org/google-docs/?1qDPEO). Esta estrategia aborda el problema de la adherencia: cuando el cliente tiene acceso a una opción "all-in-one", no es fácil para él tomar cinco cápsulas diferentes al día. Según [Marcum & Gellad (2012](https://www.zotero.org/google-docs/?4pnswV)), “la complejidad del régimen terapéutico es uno de los factores más importantes que contribuyen a la falta de adherencia”. Estos hallazgos sustentan la propuesta de Globulive de apostar por combinaciones funcionales que optimicen la eficacia y al mismo tiempo promuevan la adherencia del consumidor [(Marcum & Gellad, 2012)](https://www.zotero.org/google-docs/?z2AH7l).

El modelo de comercialización se basa en "máquinas expendedoras" que están ubicadas en lugares con un alto tráfico, como centros comerciales, gimnasios y universidades. Este canal proporciona beneficios operativos: un costo reducido de espacio físico, costos operativos más bajos al no necesitar personal en el lugar y la disponibilidad ininterrumpida (24/7). Además, el costo medio de alquiler para colocar una máquina es notablemente menor que el alquiler por metro cuadrado de un establecimiento comercial.

De manera integral, el emprendimiento tiene como objetivo respaldar su estrategia de comercialización y producto mediante el análisis de datos auténticos obtenidos de reseñas verificadas de clientes que consumen vitaminas y suplementos. La finalidad es detectar oportunidades de mercado, patrones de satisfacción y factores de preferencia que puedan aplicarse a la línea MICROGUD® utilizando modelos supervisados en el conjunto de datos "Amazon Customer Reviews - Health & Personal Care". Así, Globulive podrá fundamentar sus decisiones de formulación, posicionamiento y comunicación en datos empíricos, lo que le permitirá reforzar la competitividad de su marca en el mercado regional y distinguirse entre los demás.

# **Entendimiento del negocio**

## Formulación del problema

La empresa Globulive, S.A., dedicada a la comercialización y distribución de suplementos dietéticos/alimenticios bajo la marca MICROGUD®, enfrenta el reto de establecer estrategias para el desarrollo y la colocación del producto en un mercado competitivo y en continua evolución. Dado que la compañía es una empresa nueva, no tiene información histórica de ventas ni datos empíricos acerca de los gustos de los clientes locales. Esta escasez de información restringe la capacidad de decidir acerca de las formulaciones más atractivas, los mensajes comunicacionales y el sector de clientes potenciales.

## Contexto

Resolver esta problemática es fundamental para reducir la incertidumbre en la introducción de nuevos productos y orientar el crecimiento de la marca sobre bases analíticas sólidas. El proyecto sugiere, en vez de crear datos simulados, emplear información auténtica de reseñas de clientes que se encuentran en el conjunto de datos Amazon Customer Reviews – Health & Personal Care. Este último incluye opiniones y valoraciones sobre productos relacionados, así como vitaminas y suplementos.

El objetivo es encontrar los elementos que determinan la satisfacción del cliente y las tendencias de comportamiento que podrían trasladarse al contexto de Globulive, usando modelos supervisados. Esta perspectiva hará posible que se analicen las tendencias del mercado, se pronostique la aceptación de diferentes formulaciones y se dirija la estrategia de comercialización y desarrollo de la línea MICROGUD®.

## Objetivos

*Objetivo general:* Aplicar modelos supervisados sobre datos reales de reseñas de suplementos para analizar patrones de preferencia y satisfacción del consumidor, con el fin de orientar las decisiones de producto y mercado del emprendimiento Globulive, S.A.

*Objetivos específicos:*

1. Extraer y procesar información del conjunto de datos Amazon Customer Reviews – Health & Personal Care para obtener variables relevantes sobre satisfacción y comportamiento del consumidor.
2. Implementar técnicas de ingeniería de características y aprendizaje supervisado (ej. regresión, Random Forest) para identificar los principales factores que influyen en las reseñas positivas.
3. Analizar las tendencias y patrones de consumo derivados de los modelos para orientar decisiones sobre formulación, posicionamiento y comunicación de la línea MICROGUD®.
4. Proveer un marco analítico que permita a Globulive anticipar preferencias del mercado y diseñar estrategias basadas en evidencia.

## Preguntas clave

1. ¿Qué factores influyen en la satisfacción y valoración positiva de los suplementos según los consumidores?
2. ¿Qué patrones o características del producto se asocian con mejores reseñas y mayor probabilidad de recompra?
3. ¿Qué tipo de formulaciones o categorías presentan mayor aceptación en el mercado global de suplementos?
4. ¿Cómo pueden estos hallazgos orientar las decisiones de producto, formulación y marketing para la línea MICROGUD® en Guatemala?

## Involucrados

* Equipo de proyecto (alumnos): responsables del diseño metodológico, preprocesamiento de datos, ingeniería de características, construcción y validación de modelos supervisados. Además, analizan los resultados para identificar patrones de comportamiento y satisfacción del consumidor que puedan aplicarse a las decisiones estratégicas de Globulive.
* Área de comercialización de Globulive: participa en la interpretación de los hallazgos del modelo para definir estrategias de posicionamiento, comunicación y diferenciación de los productos de la línea MICROGUD®.
* Área de desarrollo de producto de Globulive: aporta la visión técnica sobre formulaciones y combina los resultados analíticos con el diseño de nuevas combinaciones de suplementos, vitaminas y probióticos orientadas a las preferencias del mercado.

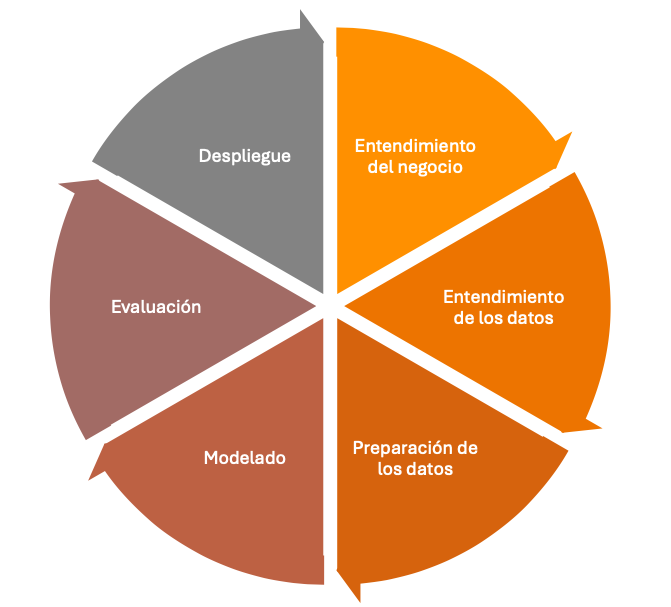


Figura 1. Etapas de CRISP-ML con la fase de “Entendimiento del negocio”.

1. Entendimiento del negocio: definir los objetivos del proyecto y la problemática desde la perspectiva de Globulive, estableciendo cómo los patrones de satisfacción del consumidor pueden orientar decisiones de producto y marketing.
2. Entendimiento de los datos: analizar las características del dataset Amazon Customer Reviews – Health & Personal Care, evaluando su calidad, estructura y variables relevantes (por ejemplo, calificación, texto de reseña, producto, y verificación de compra).
3. Preparación de los datos: realizar limpieza, codificación y transformación de variables para obtener un conjunto de datos balanceado y representativo, aplicando técnicas de ingeniería de características.
4. Modelado: entrenar modelos supervisados para identificar los factores que predicen la valoración positiva de los productos y explorar la posibilidad de recomendaciones personalizadas.
5. Evaluación: verificar el desempeño de los modelos mediante métricas apropiadas (precisión, F1-score, MAE) y analizar su coherencia con los objetivos estratégicos de Globulive.
6. Despliegue: presentar los resultados en un formato visual y analítico que permita a Globulive aplicar las conclusiones a sus decisiones de producto, formulación y posicionamiento en el mercado guatemalteco.

# **Entendimiento de los datos**

## Descripción de los datos

Dado que Globulive, S.A. es un emprendimiento en etapa de lanzamiento y no cuenta con información histórica propia, el proyecto se apoya en un conjunto de datos real y público proveniente de reseñas de consumidores de productos de salud y suplementos en línea. En particular, se utilizará el dataset “Amazon Customer Reviews – Health & Personal Care”, que recopila millones de valoraciones y comentarios verificados sobre vitaminas, probióticos, minerales y productos afines.

El conjunto incluye variables cualitativas, como el texto de la reseña, el título del producto o el indicador de compra verificada, así como cuantitativas, como la evaluación por estrellas, la cantidad total de reseñas o la cifra de votos útiles. Estas cualidades posibilitan examinar la satisfacción del cliente, los elementos que definen una evaluación positiva y las tendencias de consumo en el ámbito de los suplementos.

De este conjunto de datos se crearán nuevas variables derivadas, como la longitud del texto de la crítica, el porcentaje de letras en mayúscula, la presencia de signos de exclamación o interrogación y la proporción de votos útiles. Estas variables ampliarán los datos disponibles y harán más fácil encontrar patrones que sean significativos para Globulive.

## Técnica de Machine Learning

El proyecto aplicará un enfoque de aprendizaje supervisado, específicamente en el dominio de clasificación binaria, donde la variable objetivo será la probabilidad de recibir una reseña positiva (≥4 estrellas). Este tipo de modelado permitirá predecir, a partir del contenido textual y las características de las reseñas, qué atributos o combinaciones de producto generan mayor aceptación entre los consumidores.

* Algoritmos candidatos: Logistic Regression, Random Forest Classifier y Gradient Boosting (XGBoost).
* Métricas de evaluación: Exactitud (Accuracy), F1-Score y ROC-AUC, para medir el desempeño del modelo en la identificación de reseñas positivas frente a negativas.

## Identificación de las variables

| **Tipo de variable** | **Nombre de la variable** | **Descripción** | **Ejemplo de valores** |
| --- | --- | --- | --- |
| Cualitativa nominal | marketplace | Identifica el país o región donde se realizó la compra. | "US" |
| Cualitativa nominal | product\_category | Categoría general del producto en Amazon. | "Health & Personal Care" |
| Cualitativa nominal | product\_title | Nombre o título del producto reseñado. | "Vitamin C Gummies 500mg" |
| Cualitativa nominal | vine | Indica si la reseña proviene de un usuario del programa Amazon Vine. | "Y", "N" |
| Cualitativa nominal | verified\_purchase | Indica si la compra fue verificada por Amazon. | "Y", "N" |
| Cualitativa nominal | review\_headline | Título o resumen breve que el usuario asigna a su reseña. | "Excelente suplemento!" |
| Cualitativa textual | review\_body | Texto completo de la reseña escrita por el consumidor. | "Me encanta el sabor y me siento con más energía." |
| Cuantitativa discreta | star\_rating | Calificación asignada al producto en una escala de 1 a 5. | 1, 2, 3, 4, 5 |
| Cuantitativa discreta | helpful\_votes | Número de usuarios que marcaron la reseña como útil. | 0, 5, 32 |
| Cuantitativa discreta | total\_votes | Número total de votos recibidos por la reseña. | 0, 10, 48 |
| Cuantitativa discreta | year *(derivada)* | Año en que se publicó la reseña, extraído de review\_date. | 2019, 2020, 2021 |
| Cuantitativa discreta | month *(derivada)* | Mes de publicación de la reseña. | 1, 7, 12 |
| Cuantitativa discreta | weekday *(derivada)* | Día de la semana en que fue publicada la reseña. | 0 (lunes) – 6 (domingo) |
| Cuantitativa continua | help\_ratio *(derivada)* | Proporción de votos útiles sobre el total de votos. | 0.25, 0.80 |
| Cuantitativa discreta | len\_body\_words *(derivada)* | Longitud del texto de la reseña en número de palabras. | 25, 132 |
| Cuantitativa discreta | len\_headline *(derivada)* | Longitud del título de la reseña. | 4, 12 |
| Cuantitativa continua | upper\_ratio *(derivada)* | Proporción de letras en mayúscula sobre el total del texto (indicador de énfasis o emoción). | 0.03, 0.12 |
| Cualitativa binaria | is\_positive *(objetivo)* | Variable objetivo: 1 si la reseña es positiva (≥4 estrellas), 0 si es negativa (≤3). | 1, 0 |

**Tabla 1.** Variables de entrada y salida para el modelo de predicción de demanda en Globulive

**Literatura citada**

1. [Marcum, Z. A., & Gellad, W. F. (2012). Medication Adherence to Multi-Drug Regimens. *Clinics in Geriatric Medicine*, *28*(2), 287–300. https://doi.org/10.1016/j.cger.2012.01.008](https://www.zotero.org/google-docs/?mo87FX)
2. [NIH. (2019). *Probiotics: Usefulness and Safety*. NCCIH. https://www.nccih.nih.gov/health/probiotics-usefulness-and-safety](https://www.zotero.org/google-docs/?mo87FX)

# **Anexos**

En el caso del presente proyecto integrador no será necesario establecer un convenio de confidencialidad, ya que no se maneja información sensible ni datos confidenciales de la empresa. La información que se utilizará corresponde a datos de carácter público, insumos bibliográficos y escenarios simulados para el modelado predictivo, lo cual garantiza la transparencia del proceso. De este modo, se cuenta con el permiso y el acceso suficiente para desarrollar el proyecto sin requerir acuerdos adicionales.