Tecnológico de Monterrey

INTELIGENCIA ARTIFICIAL AVANZADA
PARA LA CIENCIA DE DATOS II

TC3007C

Reto - Arca Continental

Autores:
Daniela Jiménez Téllez - A01654798
Lautaro Gabriel Coteja - A01571214
Andrés Villarreal González - A00833915
Héctor Hibran Tapia Fernández - A01661114

Índice

1. Introducción

Arca Continental es la segunda embotelladora más grande de América Latina, y una de las más relevantes a nivel global, con una gran trayectoria en la producción y distribución de bebidas de The Coca-Cola Company, así como de botanas saladas en países como México, Ecuador y Estados Unidos. Con marcas como Bokados, Inalecsa y Wise, la compañía ofrece una variedad en su oferta, y tiene un posicionamiento sólido en distintas áreas de consumo.

En este contexto, contar con una comprensión de cómo se comportan los clientes en sus compras es de suma importancia para la empresa, especialmente cuando se trata del lanzamiento de nuevos productos. La forma en la que responde el mercado a estos lanzamientos puede variar dependiendo de varios factores, como lo son el perfil de clientes, y sus hábitos de compra y venta en torno a otros productos similares. Identificar patrones que hay entre los clientes que tienden a comprar o ignorar productos nuevos es un aspecto importante para crear estrategias comerciales más efectivas, las cuales permitan minimizar el riesgo de fracaso y optimizar la inversión en estos productos para así obtener mejores resultados.

El propósito de este proyecto es abordar los puntos anteriormente mencionados a través del análisis de una base de datos, proporcionada por Arca Continental, que contiene información relevante sobre las características de los clientes, su historial de compras, y su interacción con diferentes productos. Al explorar estos datos, se busca crear un perfil de cada consumidor, con la intención de crear y probar diferentes modelos de Machine Learning para predecir el éxito que tendrían futuros productos de lanzamiento con cada cliente, basado en comportamientos previos. Habiendo hecho esto, se pretende brindar a Arca Continental una lista de los clientes que probablemente comprarían estos productos, la cual no solo aporta valor en términos de conocer a sus clientes, sino que también ofrece una ventaja competitiva en el mercado.

2. Objetivos

A continuación se presentan los objetivos que guiarán el análisis para obtener resultados prácticos y precisos que apoyen la toma de decisiones en el lanzamiento de productos nuevos:

1. Desarrollar modelos predictivos empleando técnicas de Machine Learning e Inteligencia Artificial para identificar clientes con alta probabilidad de compra de productos de lanzamiento.

2. Identificar los segmentos de clientes más dirigidos a comprar productos nuevos, considerando factores como historial de compra y venta, y preferencias de consumo.

3. Optimizar los perfiles de cliente para aumentar las oportunidades de venta de los productos de lanzamiento.

4. Evaluar el impacto de factores demográficos y de infraestructura en el rendimiento de los productos lanzados, identificando patrones que contribuyan a su éxito o fracaso en el mercado.

3. Antecedentes

3.1. Inteligencia Artificial en el Entorno Empresarial

La Inteligencia Artificial ha transformado significativamente el panorama empresarial en las últimas décadas, proporcionando herramientas que optimizan procesos, mejoran la toma de decisiones y potencian la innovación. Desde su conceptualización, la IA ha evolucionado hacia aplicaciones prácticas que permiten a las empresas abordar desafíos complejos y aprovechar oportunidades en tiempo real. Este impacto es evidente en organizaciones que buscan liderar en mercados altamente competitivos, donde la capacidad de respuesta rápida y el aprovechamiento de datos son esenciales para el éxito estratégico.

Un ejemplo de estas aplicaciones es el uso de modelos como la regresión logística y las redes neuronales, como se presenta en el artículo de **Muñoz Muñoz (2015), titulado Aplicación de redes neuronales y regresión logística para predecir el éxito de la compra de deuda de una entidad financiera**. Este trabajo analiza cómo estas herramientas pueden predecir eventos binarios, como el éxito o fracaso de un producto, a través del análisis de datos complejos. La regresión logística permite modelar probabilidades en función de variables predictoras clave, mientras que las redes neuronales capturan relaciones no lineales y patrones ocultos en los datos.

3.2. Arca Continental

Arca Continental es una de las principales embotelladoras y distribuidoras de Coca-Cola en América Latina y Estados Unidos. La empresa destaca por su compromiso con la sostenibilidad, la innovación y la satisfacción del cliente. Su mercado objetivo abarca consumidores de bebidas carbonatadas, jugos, snacks y productos

lácteos, a quienes ofrece soluciones adaptadas a sus necesidades específicas, desde jóvenes hasta deportistas.

Para lograrlo, Arca Continental emplea estrategias avanzadas de segmentación basadas en el análisis de datos sobre hábitos de consumo y comportamiento, lo que le permite personalizar campañas de marketing, mejorar la lealtad del cliente y aumentar las ventas. Este proyecto refuerza esos objetivos mediante el uso de herramientas de análisis de datos y modelos predictivos, que no solo optimizan la asignación de productos, sino que también facilitan la predicción del éxito de nuevos lanzamientos en mercados diversificados. En línea con su sólido Código de Ética, estas acciones también contribuyen a promover la transparencia, el respeto a los derechos humanos y la sostenibilidad ambiental, fortaleciendo su liderazgo en el sector.

3.3. Cuestiones Éticas y Normativas en la I.A.

El análisis de datos y el uso de Inteligencia Artificial implican importantes cuestiones éticas y normativas que fueron consideradas cuidadosamente durante el desarrollo del proyecto. La adopción de modelos avanzados como Redes Neuronales, Random Forest y Regresión Logística se realizó con un enfoque ético, alineado con estándares internacionales, garantizando la integridad del proceso y la confianza en los resultados. A continuación, se detallan los principales aspectos éticos abordados:

- **Privacidad de los datos:** El manejo de la información proporcionada por Arca Continental se realizó cumpliendo estrictamente con normativas como el **Reglamento General de Protección de Datos (GDPR)** y la **Ley Federal de Protección de Datos Personales en Posesión de los Particulares**, asegurando que los datos utilizados fueran únicamente para los fines establecidos en el proyecto.

- **Transparencia en el uso de la IA:** Se adoptaron modelos como Redes Neuronales, Random Forest, Regresión Logística, para garantizar que las decisiones basadas en datos pudieran ser interpretadas y validadas por la empresa.

- **Sesgo y Equidad:** Se evitaron los sesgos en los modelos que se realizaron para evitar el trato injusto hacia ciertos grupos. Para esto, se utilizaron tecnicas para mitigar dichos sesgos en los datos de entrenamiento y de prueba.

- **Explicabilidad:** Los modelos realizados contienen comentarios para que sean interpretables y comprensibles para la parte interesada. Los metodos, la toma de decisiones, todo para asegurar un uso etico.

- **Supervision Humana:** Durante la ejecucion y visualizacion de los modelos, siempre hubo supervision humana, para asegurar que los resultados dados por los modelos fueran correctos, o posibles.

4. Herramientas y Recursos

Para abordar el problema de manera efectiva, se emplearon las siguientes herramientas y recursos:

1. **Bases de Datos:** En el proyecto de Arca Continental, se utilizaron tres bases de datos principales: ventas, productos y clientes, cada una diseñada para proporcionar información clave sobre diferentes aspectos del negocio. A continuación, se describe cada una:

    - **Ventas:** Esta base contiene registros detallados de las transacciones realizadas por los clientes (tiendas). Cuenta con un total de 2,347,110 filas y 4 columnas. Igualmente, tiene el identificador único de cada cliente, así como sus productos vendidos y la fecha de venta de los mismos.
    - **Productos:** Esta base proporciona información detallada sobre las características de cada producto. Cuenta con un total de 793 filas y 22 columnas, dentro de las cuales se encuentra el código único de cada producto, así como la descripción y detalles del mismo.
    - **Clientes:** Esta base contiene información sociodemográfica y contextual sobre las tiendas (clientes) que compran los productos. Cuenta con 2041 filas y 207 columnas, las cuales incluyen datos relacionados con la ubicación, entorno y características de consumo de cada cliente, así como su número de identificador.

2. **Python:** Lenguaje base elegido por su flexibilidad y su ecosistema de bibliotecas orientadas a la ciencia de datos.

    - **Pandas / Numpy:** Manipulación y transformación de datos mediante DataFrames y operaciones numéricas optimizadas.
    - **Plotly Express:** Visualización interactiva para análisis exploratorio.
    - **Scikit-Learn:** Algoritmos de machine learning y herramientas de evaluación de modelos.

- **TensorFlow:** Entrenamiento y construcción de redes neuronales avanzadas.
- **Google Colab / Jupyter Notebooks:** Entornos colaborativos para ejecución y documentación de código.

3. **Hardware:** Para garantizar un desempeño óptimo en el entrenamiento de redes neuronales y el análisis de grandes volúmenes de datos, se utilizó un equipo con las siguientes especificaciones:

   - **Procesador:** Apple M1 Max con 10 núcleos de CPU.
   - **GPU:** 24 núcleos de procesamiento gráfico.
   - **Memoria:** Entre 32 GB de memoria unificada, con un ancho de banda de 400 GB/s, optimizada para procesamiento simultáneo.

   Estas herramientas se seleccionaron por su capacidad de manejar grandes volúmenes de datos, proporcionar análisis detallados y construir modelos predictivos robustos.

5. Metodología

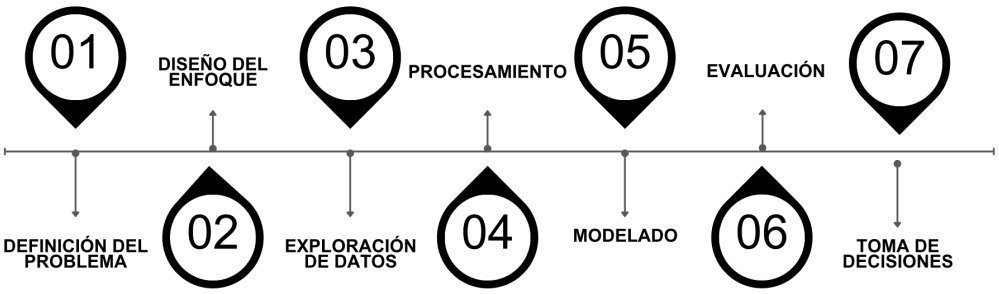

Figura 1: Diagrama del proceso para determinar si un producto será exitoso o no.

El diagrama presentado en la Figura 1 resume las etapas clave del proceso metodológico empleado. Este enfoque estructurado se diseñó para optimizar y garantizar la precisión en la predicción del éxito de un producto, que va desde la recolección de datos hasta la evaluación de resultados.

5.1. Definición del Problema

Como se mencionó anteriormente, en el sector embotellador, el lanzamiento de nuevos productos representa una inversión significativa. La predicción del éxito o fracaso de estos productos antes de su lanzamiento es muy importante para reducir riesgos y optimizar recursos. En esta sección se identificó el reto central, el cual es predecir la probabilidad de que un cliente compre un producto de lanzamiento, basándose en patrones históricos de compra. Este problema fue una tarea de clasificación binaria.

Definiciones clave proporcionadas por Arca Continental:

- **Productos de lanzamiento:** Son aquellos identificados por la primera fecha en que aparecen en el historial de ventas, excluyendo los productos lanzados en septiembre de 2019.
- **Productos exitosos:** Un producto se considera exitoso si, al excluir los dos primeros meses posteriores a su lanzamiento (para eliminar efectos iniciales como promociones o campañas de introducción), el cliente adquiere el producto durante tres meses consecutivos.

Estas definiciones delimitan el alcance del análisis y permiten una evaluación precisa del éxito de los productos lanzados.

5.2. Diseño del Enfoque

Para poder llevar a cabo el proyecto, se planteó un plan de acción que abarcó desde la recopilación y análisis exploratorio de los datos, hasta el desarrollo de modelos predictivos. Este enfoque incluyó tanto técnicas clásicas de Machine Learning, como lo son **Regresión Logística, Random Forest y Redes Neuronales**, además de tratar el problema como un **problema de clasificación.**

5.3. Exploración de Datos

En esta sección se analizó la calidad y estructura de las tres bases de datos proporcionadas: **productos**, **clientes** y **ventas**. Este análisis incluyó visualizaciones descriptivas y métricas estadísticas para comprender mejor las características clave de los datos, como tendencias de compra, distribución de clientes y propiedades de los productos. Igualmente, se identificaron datos faltantes en algunas variables, así como columnas que no proporcionaban información crucial para el análisis.

Dataset de ventas

	CustomerId	material	calmonth	uni_box
count	2.347110e+06	2.347110e+06	2.347110e+06	2.347110e+06
mean	5.027978e+08	4.056356e+03	2.020960e+05	8.035520e+00
std	4.425087e+06	4.392404e+03	9.677544e+01	4.638741e+01
min	4.999201e+08	1.000000e+00	2.019090e+05	-2.007783e+01
25%	5.001125e+08	4.520000e+02	2.020080e+05	8.805000e-01
50%	5.002994e+08	2.253000e+03	2.021060e+05	2.113400e+00
75%	5.099778e+08	9.086000e+03	2.022040e+05	5.072400e+00
max	5.108391e+08	1.736900e+04	2.022120e+05	7.279920e+03

Figura 2: Datos estadísticos del Dataset de Ventas

Dataset de productos

	Material	Productos_Por_Empaque	MLSize	Ncb
count	793.000000	793.000000	793.000000	793.000000
mean	6884.563682	10.223203	892.046658	0.382093
std	5461.796220	7.428573	785.294099	0.486206
min	1.000000	1.000000	0.000000	0.000000
25%	1956.000000	6.000000	355.000000	0.000000
50%	5063.000000	8.000000	600.000000	0.000000
75%	14074.000000	12.000000	1200.000000	1.000000
max	17369.000000	40.000000	5000.000000	1.000000

Figura 3: Datos estadísticos del dataset de Productos

Dataset de clientes

	CustomerId	pc_agr_300m	pc_comercial_300m	pc_generales_300m	pc_habitacional_300m	pc_habitacional_mixta_300m	pc_industrial_300m	pc_minero_300m	pc_mixta_300m	pc_negocios_300m
count	2.041000e+03	2041.000000	2041.000000	2041.000000	2041.000000	2041.000000	2041.000000	2041.000000	2041.000000	2041.000000
mean	5.045471e+08	0.048996	2.638111	7.142613	61.222468	19.483435	2.711553	0.102474	3.844304	0.284439
std	5.077644e+06	2.213495	9.322655	14.075650	30.494866	19.805396	11.069627	2.419347	9.834340	3.468600
min	4.999201e+08	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	5.001543e+08	0.000000	0.000000	0.000000	38.270000	1.680000	0.000000	0.000000	0.000000	0.000000
50%	5.004064e+08	0.000000	0.000000	1.700000	64.650000	14.070000	0.000000	0.000000	0.000000	0.000000
75%	5.104164e+08	0.000000	0.450000	7.880000	88.570000	31.450000	0.000000	0.000000	2.320000	0.000000
max	5.108467e+08	100.000000	100.000000	100.000000	100.000000	94.340000	100.000000	72.510000	82.430000	82.470000

velocidad_hora_16	velocidad_hora_17	velocidad_hora_18	velocidad_hora_19	velocidad_hora_20	velocidad_hora_21	velocidad_hora_22	velocidad_hora_23	accesibilidad	industry_customer_size
1104.000000	1104.000000	1104.000000	1104.000000	1104.000000	1104.000000	1104.000000	1104.000000	1968.000000	2041.000000
25.654363	24.783288	24.028699	24.288119	24.985386	25.245758	26.150332	26.488753	0.249359	2.266536
14.091818	14.082192	13.993819	13.723700	13.638142	13.493612	13.969840	13.940245	0.104189	1.330083
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.046641	1.000000
15.000000	14.000000	14.000000	14.000000	15.000000	15.000000	16.000000	17.000000	0.176717	1.000000
23.000000	21.000000	20.100000	21.000000	22.000000	22.000000	23.600000	24.000000	0.231652	2.000000
39.000000	34.000000	32.000000	32.000000	33.150000	35.000000	40.000000	40.000000	0.296014	3.000000
90.000000	90.000000	90.000000	90.000000	90.000000	88.000000	88.000000	90.000000	0.865606	5.000000

Figura 4: Datos estadísticos del data de clientes

5.4. Procesamiento de Datos

Antes de desarrollar los modelos, fue necesario llevar a cabo un exhaustivo procesamiento de los datos. Este proceso incluyó una reducción de las 235 columnas iniciales a 25 columnas finales, enfocándose en eliminar variables irrelevantes, consolidar información redundante y seleccionar las características más relevantes para los objetivos del análisis, para seguimos los siguientes pasos:

5.4.1. Limpieza de Datos

Durante el proceso de limpieza de datos, se identificaron columnas con un alto porcentaje de valores faltantes, lo que podía afectar negativamente el análisis y los modelos predictivos. En este caso, se decidió eliminar las columnas relacionadas con la movilidad y la velocidad promedio por horas, ya que presentaban un volumen significativo de datos faltantes que no podían ser imputados de manera confiable.

En el siguiente paso del proceso de limpieza de datos, se abordaron aquellas columnas que contenían pocos valores faltantes. En lugar de eliminarlas, se optó por imputar los valores faltantes utilizando la media de cada columna agrupada por la categoría correspondiente de la variable sub_canal_comercial.

5.4.2. Codificación de Variables Categóricas

Se usó One-Hot Encoding para transformar variables categóricas, convirtiéndolas en variables binarias que los algoritmos pueden interpretar efectivamente. Variables como ProductType, Flavor, y categoria_instalaciones, se representaron en este formato. Esto permitió mantener la diversidad de categorías mientras se preserva la capacidad del modelo para generalizar.

5.4.3. Estandarización de Variables Numéricas

La estandarización se aplicó a las variables numéricas para que tuvieran una media de 0 y una desviación estándar de 1. Esto asegura que las variables tengan la misma escala, evitando que magnitudes mayores dominen el modelo, siguiendo las recomendaciones de IBM sobre estandarización. Entre las variables estandarizadas se incluyeron indicadores económicos como gasto_promedio_300m e indicadores de éxito como success_ratio.

5.4.4. Transformación de Nuevas Variables

Se crearon nuevas variables para mejorar el análisis, capturando el comportamiento histórico de los clientes y su relación con los productos. Estas incluyen indicadores de éxito y métricas basadas en características específicas, como sabor, tamaño y tipo de contenedor.

Variable	Descripción
successful	Variable binaria que indica si un producto de lanzamiento fue considerado exitoso para un cliente. Un producto es exitoso si, excluyendo los primeros dos meses de venta, el cliente lo adquiere al menos tres veces consecutivas. Definición proporcionada por el socio formador.
categoria_instalaciones	Variable categórica que clasifica las instalaciones cercanas a cada cliente en función de su cantidad: - **Ni una**: Sin instalaciones (0). - **Pocas**: Entre 0.2 y 3 instalaciones. - **Medio**: Entre 3 y 6 instalaciones. - **Muchas**: Más de 6 instalaciones.
success_ratio	Porcentaje de productos exitosos que un cliente ha comprado históricamente dentro de un tipo específico de producto.
success_ratio_flavor	Porcentaje de productos exitosos que un cliente ha comprado históricamente dentro de un tipo de sabor específico.
success_ratio_mlsize	Porcentaje de productos exitosos que un cliente ha comprado históricamente dentro de un tamaño de producto específico.
success_ratio_container	Porcentaje de productos exitosos que un cliente ha comprado históricamente dentro de un tipo de contenedor específico.
success_ratio_total	Porcentaje de productos exitosos que un cliente ha comprado históricamente.

5.4.5. Selección de Variables

Para garantizar un modelo equilibrado y representativo, se seleccionaron las variables buscando un balance entre las características de la tienda y del producto. Este enfoque evita que las características de una categoría (ya sea tienda o producto) predominen sobre las demás, asegurando que ambas dimensiones tengan una influencia equitativa en las predicciones. De esta manera, se busca capturar tanto el contexto del cliente como las propiedades del producto, mejorando la capacidad del modelo para generalizar y realizar predicciones precisas.

Variable	Descripción
Características de la tienda	
categoria_instalaciones	Clasifica las instalaciones cercanas según su cantidad.
pc_comercial_300m	% de área comercial en 300m de la tienda.
pc_negocios_300m	% de área de negocios en 300m.
pc_turismo_300m	% de área de turismo en 300m.
gasto_promedio_300m	Gasto promedio por hogar en 300m.
ingreso_promedio_300m	Ingreso promedio por hogar en 300m.
accesibilidad	Facilidad de acceso a la tienda.
sub_canal_comercial	Tipo de canal comercial de la tienda.
pob_ab_300m	% de población nivel A/B en 300m.
pob_cmas_300m	% de población nivel C+ en 300m.
pob_c_300m	% de población nivel C en 300m.
pob_cmen_300m	% de población nivel C- en 300m.
pob_dmas_300m	% de población nivel D+ en 300m.
pob_d_300m	% de población nivel D en 300m.
pob_e_300m	% de población nivel E en 300m.
success_ratio	% de productos exitosos por tipo de producto.
success_ratio_flavor	% de productos exitosos por sabor.
success_ratio_mlsize	% de productos exitosos por tamaño (ml).
success_ratio_container	% de productos exitosos por contenedor.
Características del producto	
Flavor	Sabor del producto.
MLSize	Tamaño en mililitros.
ProductType	Tipo de producto.
Size	Tamaño del empaque.
Returnability	Indicador de si es retornable.
Container	Tipo de contenedor (botella, lata).

5.4.6. División de Datos

Los datos se dividieron en dos conjuntos principales: entrenamiento y prueba. El conjunto de prueba incluyó las observaciones correspondientes al rango de los seis meses previos al mes de diciembre de 2022, incluyendo dicho mes como límite superior. El conjunto de entrenamiento abarcó todas las observaciones antes de este rango, es decir, aquellas anteriores al inicio del período de prueba.

5.5. Modelado

Después de la selección de variables, se entrenaron y evaluaron tres modelos utilizando el mismo conjunto de variables predictoras. Las técnicas aplicadas incluyeron:

5.5.1. Regresión Logística

El modelo de **Regresión Logística** fue utilizado como una referencia base debido a su simplicidad y capacidad para modelar relaciones lineales entre las variables predictoras y el objetivo.

Hiperparámetros Seleccionados:

```
from sklearn.linear_model import LogisticRegression

model = LogisticRegression(penalty='l2', C=1.0, solver='lbfgs'
    , max_iter=2000, class_weight='balanced', random_state=42)
```

- `penalty='l2'`: Regularización Ridge, utilizada para reducir el sobreajuste penalizando coeficientes grandes.
- `C=1.0`: Factor de regularización; valores más altos disminuyen la regularización, otorgando mayor flexibilidad al modelo.
- `solver='lbfgs'`: Algoritmo de optimización eficiente para problemas de clasificación binaria con regularización.
- `max_iter=2000`: Incremento del número máximo de iteraciones para garantizar la convergencia del modelo.
- `class_weight='balanced'`: Ajuste automático de pesos basado en la frecuencia de las clases, útil en casos de desbalance de datos.
- `random_state=42`: Semilla para asegurar la reproducibilidad de los resultados.

5.5.2. Random Forest

Se utilizó **Random Forest**, ajustado mediante GridSearchCV, para encontrar la combinación óptima de hiperparámetros que maximizan su desempeño.

Hiperparámetros Seleccionados:

```
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n_estimators=600, max_depth=50,
    min_samples_split=10, min_samples_leaf=4, max_features='
   sqrt', class_weight='balanced', random_state=42)
```

- `n_estimators = 600`: Número de árboles en el bosque.
- `max_depth = 50`: Profundidad máxima.
- `min_samples_split = 10`: Muestras mínimas necesarias para dividir un nodo.
- `min_samples_leaf = 4`: Muestras mínimas requeridas en hojas terminales.
- `max_features = 'sqrt'`: Selección de características basada en la raíz cuadrada del total.
- `class_weight = 'balanced'`: Pesos ajustados automáticamente según la frecuencia de clases.

tabularx booktabs

5.5.3. Redes Neuronales

Para capturar relaciones no lineales complejas en los datos, se utilizó una **Red Neuronal de Alimentación Directa** (*Feedforward Neural Network*). El modelo fue diseñado con múltiples capas ocultas y técnicas de regularización para prevenir el sobreajuste.

Parámetros Seleccionados	
`Capa de entrada`	64 neuronas con función de activación `ReLU`
`Primera capa oculta`	32 neuronas con `ReLU`, seguida de *Batch Normalization* y *Dropout* (50 %)
`Segunda capa oculta`	16 neuronas con `ReLU`, seguida de *Batch Normalization* y *Dropout* (50 %)
`Capa de salida`	1 neurona con función de activación `Sigmoid`
`Batch Normalization`	Aplicada para estabilizar y acelerar el entrenamiento
`Dropout`	Tasa del 50 % para reducir el sobreajuste
`Función de pérdida`	`Binary Crossentropy`
`Optimizador`	`Adam`
`Número de épocas`	50
`Tamaño de lote`	32

Detalles Técnicos Adicionales:

- **Funciones de Activación:**
  - `ReLU` (*Rectified Linear Unit*):
    $$f(x) = \max(0, x)$$
  - `Sigmoid`:
    $$\sigma(x) = \frac{1}{1 + e^{-x}}$$
- **Regularización:**
  - *Batch Normalization* ayuda a acelerar el entrenamiento y estabilizar la distribución de las activaciones.
  - *Dropout* con una tasa del 50 % previene el sobreajuste al desactivar aleatoriamente neuronas durante el entrenamiento.
- **Optimizador `Adam`:** Combina las ventajas de los algoritmos `AdaGrad` y `RMSProp` para una optimización eficiente y adaptativa.

5.6. Evaluación

Para evaluar el desempeño de los modelos desarrollados, se utilizaron métricas clave como precisión, recall, F1-score, y accuracy. Estos indicadores proporcionan una visión integral sobre la capacidad predictiva y la eficiencia de cada modelo en clasificar correctamente los productos como exitosos o no exitosos.

Modelo	**Precisión Clase 0**	**Precisión Clase 1**	**Recall (Macro Avg)**	**F1-Score (Macro Avg)**	**Accuracy**
Regresión Logística	0.98	0.67	0.90	0.85	0.88
Random Forest	0.99	0.82	0.96	0.93	0.94
Red Neuronal	0.77	0.79	0.52	0.47	0.77

Aunque dos de los modelos mostraron un desempeño sólido, el que obtuvo los mejores resultados fue el modelo *Random Forest*, destacándose significativamente sobre las demás opciones evaluadas. Este logró una exactitud promedio del 94 % (proporción de predicciones correctas entre el total de predicciones realizadas), indicando que fue altamente efectivo al identificar correctamente los productos exitosos y no exitosos. Su recall promedio fue del 96 % (proporción de casos positivos correctamente identificados entre todos los casos positivos reales), lo que reflejó su capacidad para captar casi todos los casos positivos de productos exitosos. Esto es crucial, ya que minimizar falsos negativos ayuda a no ignorar posibles oportunidades. Además, el F1-score de 93 % (promedio armónico entre precisión y recall, que equilibra ambas métricas) demostró un balance óptimo entre precisión y recall, asegurando un rendimiento equilibrado. La red neuronal, en cambio, no demostró un buen desempeño en comparación con los otros modelos.

5.7. Toma de Decisiones

Se seleccionó **Random Forest** como el modelo final debido a su desempeño superior en la clasificación de productos exitosos. Este modelo demostró ser efectivo para ingresar un producto de lanzamiento y generar un listado con el top 10 de clientes más propensos a adoptar dicho producto con éxito. Los resultados obtenidos facilitarán decisiones estratégicas al priorizar clientes clave y optimizar recursos, maximizando así las probabilidades de éxito comercial de los nuevos productos.

La elección del modelo se fundamentó en un análisis integral basado en métricas de desempeño como recall, F1-score y precisión. Dado que la clase mayoritaria en el conjunto de datos era "no exitosos", muchos modelos tendían a clasificar cualquier producto como un fracaso. Sin embargo, Random Forest sobresalió por su capacidad para identificar correctamente los productos exitosos, cumpliendo con el objetivo principal del proyecto: garantizar que la clase exitosa fuera calculada con precisión para decisiones informadas y efectivas.
Además, se espera que este modelo brinde valor significativo a Arca Continental, ayudándole a la **toma de decisiones** estratégicas para detectar con antelación si un producto será exitoso o no. Esto permitirá a la organización ajustar sus estrategias de marketing, distribución y producción de manera proactiva, minimizando riesgos e incrementando las probabilidades de que los nuevos lanzamientos alcancen el éxito en el mercado.

6. Resultados

Tras la implementación y evaluación de los modelos predictivos, se obtuvo un análisis detallado del desempeño de cada uno. El modelo Random Forest destacó en la identificación de productos exitosos, superando las limitaciones observadas en la Regresión Logística y las Redes Neuronales. Los hallazgos principales son los siguientes:

Clientes con Mayor Probabilidad de Éxito para el Producto: COLAS REGULAR

ID del Cliente	Tipo de Producto	Probabilidad de Éxito
510115930	COLAS REGULAR	95.50%
499923595	COLAS REGULAR	94.33%
510771444	COLAS REGULAR	94.00%
510797075	COLAS REGULAR	94.00%
510795583	COLAS REGULAR	93.00%
500023855	COLAS REGULAR	93.00%
510795583	COLAS REGULAR	93.00%
510825031	COLAS REGULAR	92.67%
510758682	COLAS REGULAR	92.50%
500411797	COLAS REGULAR	91.00%
500192457	COLAS REGULAR	90.67%
500101080	COLAS REGULAR	90.50%
510786517	COLAS REGULAR	90.33%

Figura 5: Clientes con mayor probabilidad de éxito para Colas Regulares

En esta figura se puede observar una lista con los clientes del conjunto de prueba que tienen mayor probabilidad de comprar un producto por su categoría. En este

caso, se escogió Colas Regular, y para poder comprobar los resultados del modelo, se hizo la siguiente gráfica:

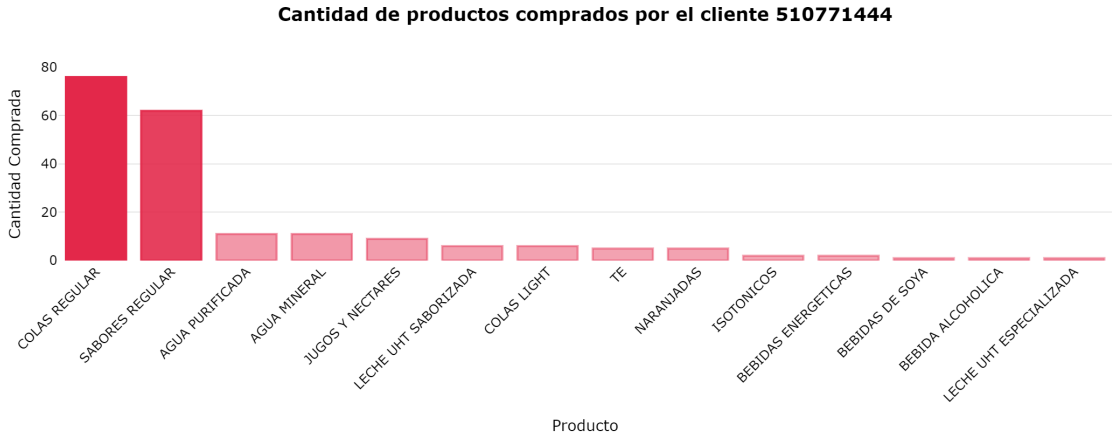

Figura 6: Productos comprados por el cliente 510771444

En este histograma se observa la cantidad de productos por categoría que compró el cliente 510771444. De acuerdo a su historial, el tipo de producto que más compró de Septiembre de 2019 a Diciembre 2022 fueron las Colas Regulares. Esto quiere decir que el modelo logró darle una probabilidad coherente ante un tipo de producto, con respecto a su historial de compra. Igualmente, cabe recalcar que el código para generar esta gráfica puede ser modificado para hacer experimentos con más clientes y tipos de productos.

Finalmente, se probó el modelo con un producto completamente nuevo de la categoría “Colas Regulares”, con las características que se ven en la siguiente imagen:

```
ProductType: COLAS REGULAR
Flavor: MANGO
MLSize: 500
Container: LATA
Returnability ('Returnable' o 'Non-Returnable'): NO RETORNABLE
Size (e.g., 'FAMILIAR', 'INDIVIDUAL'): INDIVIDUAL

Top 10 clientes más propensos al éxito del nuevo producto:
      CustomerId  probabilidad_exito
1942   510632319            0.591504
334    500101560            0.588114
789    500370987            0.586594
1127   510100589            0.585586
1021   500488633            0.585158
349    500104639            0.585029
1793   510275595            0.584990
679    500283229            0.584958
1758   510020475            0.584824
1051   500517459            0.584565
```

Figura 7: Ejemplo de Nuevo Producto a Random Forest.

Los resultados indicaron una baja probabilidad de éxito, lo que valida la capacidad del modelo para generalizar a nuevos casos y no sesgarse ante tipo de productos que son muy comprados. En la imagen, se puede observar cómo Random Forest asigna probabilidades de éxito para cada cliente, lo que permite generar un ranking con los 10 clientes más propensos a adoptar nuevos productos. Esto proporciona una herramienta estratégica para priorizar esfuerzos de marketing.

7. Conclusión

El desarrollo de este proyecto para Arca Continental, ha demostrado el potencial de la inteligencia artificial como herramienta clave para la toma de decisiones estratégicas en el sector empresarial. El uso del modelo Random Forest, seleccionado por su capacidad para manejar datos complejos y ofrecer una alta precisión, permitio identificar a los clientes más propensos a adoptar nuevos productos, así optimizando recursos y reduciendo riesgos asociados a los lanzamientos. Este enfoque predictivo refuerza las estrategias comerciales al maximizar la eficiencia y la probabilidad de éxito de los nuevos productos en un mercado altamente competitivo.

Además de los logros técnicos, el proyecto resalta la importancia de las normativas y consideraciones éticas en la aplicación de la inteligencia artificial. Se garantizo la protección de los datos sensibles mediante el cumplimiento de regulaciones como el GDPR y la Ley Federal de Protección de Datos Personales en Posesión de los Particulares. La implementación de modelos transparentes y explicables asegura que las decisiones puedan ser comprendidas y validadas por las partes interesadas, fo-

mentando la confianza en los resultados. Asimismo, la supervisión humana a lo largo del proceso evitó el sesgo algorítmico y garantizo una equidad en el tratamiento de los datos, reflejando un compromiso con la integridad y el respeto a los derechos humanos.

En este contexto, el proyecto no solo destaca por su impacto en la optimización de las estrategias de negocio, sino también por su enfoque ético y responsable en el uso de tecnologías avanzadas. Al integrar prácticas normativas y éticas robustas con herramientas de inteligencia artificial, Arca Continental refuerza su liderazgo en el sector, demostrando que es posible innovar sin comprometer los valores fundamentales que promueven un entorno empresarial sostenible y equitativo.

Por ende, este proyecto sirve como un modelo ejemplar de cómo la inteligencia artificial puede ser utilizada de manera efectiva y responsable para abordar problemas complejos en el ámbito corporativo, ofreciendo una guía para futuras implementaciones que busquen equilibrar la innovación tecnológica con el cumplimiento normativo y la ética empresarial.

8. Referencias

[1] Arca Continental. (s.f.). *Nuestra compañía*. Recuperado de `https://www.arcacontal.com/nuestra-compa%C3%B1%C3%ADa.aspx`

[2] Fernandes, A. A. T., Figueiredo Filho, D. B., Rocha, E. C. da, & Nascimento, W. da S. (2020). Read this paper if you want to learn logistic regression. *Revista de Sociologia e Política*, 28(74), e006. 10.1590/1678-987320287406en

[3] Kufel, J., Bargieł-Łaczek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., . . . , & Gruszczyńska, K. (2023). What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine. *Diagnostics*, 13(15), 2582. 10.3390/diagnostics13152582

[4] Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, 20(1), 3–29. 10.1177/1536867X20909688

9. Repositorio de GitHub

Este repositorio contiene el código necesario para revisar los modelos.
`https://github.com/Lautaro000/TC3007C.101-Equipo-5`