Modelo Predictivo Sobre el Precio del Whisky en Iowa.

Leiva R,. Alejandra. maleivar@eafit.edu.co
Castaño M, Juan Pablo. jpcastano@eafit.edu.co
Londoño. B, Juan Pablo. jplondonob@eafit.edu.co

I. ENTENDIMIENTO DEL NEGOCIO

I-A. ANTECEDENTES

El sector de bebidas alcohólicas en Estados Unidos es una industria de gran relevancia económica. En 2023, este mercado generó ingresos por aproximadamente 470.696,5 millones de dólares, y se proyecta que alcance los 1.006.255,3 millones de dólares para el año 2030 [1].

Históricamente, la producción y consumo de alcohol han estado arraigados en la cultura estadounidense desde la época colonial. Durante el siglo XVII, el ron y el whisky comenzaron a popularizarse ampliamente. Sin embargo, el sector enfrentó desafíos significativos, como la Ley Seca entre 1920 y 1933, que prohibió la producción y venta de alcohol, afectando gravemente la economía y fomentando el mercado ilegal [2].

Tras la anulación de la Prohibición, la industria se reestructuró y experimentó un crecimiento sostenido. En particular, el whisky estadounidense ha ganado reconocimiento internacional y ha sido un motor clave en la expansión del mercado de bebidas. La importancia económica del sector también se refleja en su contribución al empleo y a los ingresos fiscales.

I-B. OBJETIVO DEL NEGOCIO

El objetivo principal de este proyecto es predecir los precios de diferentes tipos de whisky en el estado de Iowa. Esta predicción permite a distribuidores y comerciantes ajustar sus precios en función de variables clave, mejorando la competitividad y los márgenes de ganancia. Asimismo, facilita la identificación de patrones de consumo y tendencias en las ventas, apoyando la toma de decisiones estratégicas.

I-C. PREGUNTAS DE NEGOCIO

- ¿Qué variables muestran la mayor correlación con el precio del whisky?
- ¿En qué condado el modelo ofrece su mejor predicción y en cuál la más imprecisa?
- ¿Para qué número de botellas por caja (pack) el modelo alcanza su mayor precisión y para cuál su menor precisión?
- ¿Puede el modelo alcanzar un MAE inferior a un dólar?

I-D. CRITERIOS DE ÉXITO

El proyecto se considerará exitoso si cumple con los siguientes criterios:

- Determinar las variables más influyentes en la formación de precios del whisky.
- Lograr un error promedio absoluto (MAE) menor un dollar.
- Presentar los resultados de forma clara para facilitar su interpretación.

I-E. INVENTARIO DE RECURSOS

Fuente de datos: conjunto de datos público del estado de Iowa (https://data.iowa.gov/Sales-Distribution/Iowa-Liquor-Sales/m3tr-qhgy).

Versión utilizada: filtrado de 8.389.606 filas y 11 columnas. Más adelante se detalla cómo se realiza el filtrado.

Herramientas: Python, pandas, scikit-learn, matplotlib, seaborn y el entorno de Jupyter Notebook.

Hardware: computadores personales de los integrantes del equipo.

I-F. REQUISITOS

- Informe final en formatos `.docx` y `.pdf`.
- Dos notebooks: uno para análisis, preparación, modelado y evaluación; otro para despliegue.
- Uso de una misma semilla aleatoria para todos los modelos.
- Incluir visualizaciones claras y resultados explicables.
- Código bien estructurado y comentado.

I-G. SUPUESTOS

Los precios están en dólares estadounidenses. No existen factores externos significativos que alteren los precios y no estén registrados en el dataset.

I-H. LIMITACIONES

El volumen del dataset original obligó a realizar un filtrado previo para facilitar el procesamiento.

Se trabajó en equipos personales con capacidad limitada para grandes volúmenes de datos.

I-I. RIESGOS Y CONTIGENCIAS

Datos inválidos o inconsistentes:

- Contingencia: limpieza de datos y validación cruzada.

Volumen excesivo de datos:

- Contingencia: filtrado del dataset y procesamiento por bloques.

Posibles valores atípicos y datos faltantes:

- Contingencia: aplicación de técnicas estadísticas de imputación y eliminación selectiva.

I-J. TERMINOLOGÍA

- `Precio del whisky`: valor total pagado por una unidad de venta (pack) del producto. Variable objetivo a predecir.
- `Zip Code`: código postal de cinco dígitos donde se encuentra la tienda.
- `County Number`: identificador oficial de condado dentro del estado de Iowa.
- `Pack`: número de botellas contenidas en cada unidad de venta.
- `Bottles Sold`: cantidad total de botellas compradas en la transacción.
- `MAE (Mean Absolute Error)`: promedio del error absoluto entre el valor predicho y el valor real.
- `MAPE (Mean Absolute Percentage Error)`: error porcentual promedio de las predicciones.
- `Random Forest`: técnica de aprendizaje automático basada en múltiples árboles de decisión.

I-K. COSTO

- Tiempo estimado por integrante distribuido en recolección, limpieza, análisis y documentación.
- Uso de recursos computacionales personales: almacenamiento, energía, conectividad.
- Tiempo dedicado a pruebas y validaciones por el tamaño del dataset.

I-L. BENEFICIOS

- Identificación de los factores clave que influyen en el precio del whisky.
- Desarrollo de un modelo predictivo preciso y robusto.
- Aplicabilidad del modelo a otros tipos de bebidas o regiones.

I-M. OBJETIVO ANALÍTICO

El objetivo es construir un modelo supervisado de regresión que permita predecir el precio del whisky con base en variables relacionadas con el producto, la tienda y las características de la venta. Se evaluarán modelos como árboles de decisión, random forest y gradient boosting, variando hiper parámetros y subconjuntos de variables.

I-N. CRITERIOS DE ÉXITO DESDE LA ANALÍTICA

El modelo será considerado exitoso si:

- Alcanza un `MAE` inferior a un dollar.
- Muestra consistencia entre los conjuntos de validación y prueba (sin overfitting).

I-Ñ. PLAN DE PROYECTO

El desarrollo del proyecto se estructuró según las fases de la metodología CRISP-DM: entendimiento del negocio, entendimiento de los datos, preparación, modelado, evaluación y despliegue. Las tareas se distribuyeron entre los miembros del equipo de forma equitativa y se realizaron ajustes iterativos cuando fue necesario.

I-O. EVALUACIÓN INICIAL DE HERRAMIENTAS Y TÉCNICAS

Desde el inicio se seleccionaron herramientas estándar en ciencia de datos como Python, pandas, scikit-learn y Jupyter Notebook. Se definió un enfoque sistemático para entrenar al menos 125 modelos de árboles de decisión, random forest y gradient boosting, usando ciclos automatizados de prueba de hiperparámetros y análisis de desempeño.

I-P. DICCIONARIO DEL DATASET UTILIZADO

A continuación se describe cada uno de los campos presentes en el conjunto de datos utilizado en este proyecto, especificando el tipo de dato y el tamaño en megabytes de cada campo.

- **Date**: Fecha en la que se realizó la orden de licor.
  - **Tipo de dato**: objeto
  - **Tamaño**: 472.06 MB
- **Store Number**: Número único asignado a la tienda que realizó la compra del licor.
  - **Tipo de dato**: entero
  - **Tamaño**: 64.01 MB
- **Zip Code**: Código postal de la ubicación de la tienda que realizó el pedido.
  - **Tipo de dato**: objeto
  - **Tamaño**: 321.03 MB
- **County Number**: Código del condado en el estado de Iowa donde se encuentra la tienda que hizo el pedido.
  - **Tipo de dato**: decimal
  - **Tamaño**: 64.01 MB
- **Pack**: Número de botellas que contiene cada caja del licor solicitado.
  - **Tipo de dato**: entero
  - **Tamaño**: 64.01 MB
- **Bottle Volume (ml)**: Volumen en mililitros de cada botella de licor incluida en el pedido.
  - **Tipo de dato**: entero
  - **Tamaño**: 64.01 MB
- **State Bottle Cost**: Precio que la División de Bebidas Alcohólicas (entidad estatal responsable de comprar, distribuir y fijar precios al por mayor) del estado pagó por cada botella de licor.
  - **Tipo de dato**: decimal
  - **Tamaño**: 64.01 MB
- **Vendor Number**: Identificador único del proveedor o empresa que distribuye la marca de licor comprada por el estado.

  - **Tipo de dato**: decimal
  - **Tamaño**: 64.01 MB
- **Bottles Sold**: Cantidad total de botellas de licor compradas por la tienda.
  - **Tipo de dato**: entero
  - **Tamaño**: 64.01 MB
- **Sale (Dollars)**: Costo total de la compra.
  - **Tipo de dato**: decimal
  - **Tamaño**: 64.01 MB
- **Category Name**: Categoría del licor solicitado.
  - **Tipo de dato**: objeto
  - **Tamaño**: 542.81 MB
- **Volumen del dataset**: 8.389.606 filas × 11 columnas
- **Tamaño total del dataset**: Aproximadamente 1.3 GB

Campo	Tamaño (MB)	Tipo de dato	Descripción
Date	472.06	objeto	Fecha en la que se realizó la orden de licor.
Store Number	64.01	entero	Número único asignado a la tienda que realizó la compra del licor.
Zip Code	321.03	objeto	Código postal de la ubicación de la tienda que realizó el pedido.
County Number	64.01	decimal	Código del condado en el estado de Iowa donde se encuentra la tienda que hizo el pedido.
Pack	64.01	entero	Número de botellas que contiene cada caja del licor solicitado.
Bottle Volume (ml)	64.01	entero	Volumen en mililitros de cada botella de licor incluida en el pedido.
State Bottle Cost	64.01	decimal	Precio que la División de Bebidas Alcohólicas pagó por cada botella de licor.
Vendor Number	64.01	decimal	Identificador único del proveedor o empresa que distribuye la marca de licor comprada por el estado.
Bottles Sold	64.01	entero	Cantidad total de botellas de licor compradas por la tienda.
Sale (Dollars)	64.01	decimal	Costo total de la compra.
Category Name	542.81	objeto	Categoría del licor solicitado.

Figura 1: Diccionario.

II. ENTENDIMIENTO DE DATOS

II-A. CARGA DE DATOS

La base de datos original contenía 31,550,621 registros. Debido a que el conjunto de datos era muy grande y contenía múltiples categorías de bebidas alcohólicas, decidimos filtrar el dataset para incluir únicamente los registros correspondientes a whisky. Este filtrado se realizó sobre la columna `Category Name`, que indica el tipo de bebida alcohólica, seleccionando solo aquellas filas cuyo valor correspondía a alguna de las siguientes categorías de whisky:

- Canadian Whiskies
- Straight Bourbon Whiskies
- Tennessee Whiskies
- Scotch Whiskies
- Blended Whiskies
- Straight Rye Whiskies
- Irish Whiskies
- Single Barrel Bourbon Whiskies
- Corn Whiskies
- Japanese Whisky

Esta decisión se tomó con el fin de enfocar el análisis en un tipo de bebida específico, facilitando obtener conclusiones más claras y precisas sobre el comportamiento de venta en este segmento particular.

Para iniciar el proyecto, se realizó la carga correcta del conjunto de datos filtrado, asegurando que el archivo se importara sin errores y en el formato esperado.

```
filtered_data = pd.read_csv("filtered_data.csv")
filtered_data
```

Figura 2: Proceso de descarga del conjunto de datos.

Posteriormente, se realizó una selección cuidadosa de las columnas que se consideraron más influyentes y relevantes para la construcción del modelo. De este modo, se excluyeron variables que no aportaban información significativa o que podían generar redundancia.

Las variables elegidas cubren diferentes aspectos fundamentales del negocio y del proceso de venta, garantizando una visión completa de las condiciones bajo las cuales se realizan las órdenes de whisky. Las variables ultilizadas fueron:

- `Date`
- `Store Number`
- `Zip Code`
- `County Number`
- `Pack`
- `Bottle Volume (ml)`
- `State Bottle Cost`
- `Vendor Number`
- `Bottles Sold`
- `Sale (Dollars)`

Esta selección permite que el modelo tome en cuenta aspectos temporales, espaciales, comerciales y de producto, lo que contribuye a un análisis integral y a la obtención de predicciones más representativas del comportamiento real del mercado.

Esta base de datos, con las variables mencionadas y debidamente filtrada y preparada, fue utilizada para realizar todo el análisis y desarrollo del proyecto.

II-B. ANÁLISIS DESCRIPTIVO DE LOS DATOS

Al observar las primeras cinco filas del dataset, se puede ver que los datos contienen información de ventas de whisky, incluyendo columnas como la fecha de la transacción, el número de tienda, código postal, número de condado, cantidad de botellas por paquete, volumen de botella en mililitros, costo estatal por botella, número de botellas vendidas, número de proveedor, el monto de venta en dólares y la categoría del whisky. Las fechas en el ejemplo varían entre los años 2012 y 2015. Las variables numéricas como `Store Number`, `Pack`, `Bottle Volume (ml)` y `Bottles Sold` aparecen en formato entero, mientras que otras variables relacionadas con costos y ventas (`State Bottle Cost`, `Sale (Dollars)`) están en formato decimal. Las categorías de whisky se presentan en formato texto y reflejan distintas clases, siendo una columna clave para segmentar el análisis.

```
data_elegida_reset.head()
```

	Date	Store Number	Zip Code	County Number	Pack	Bottle Volume (ml)	State Bottle Cost	Bottles Sold	Vendor Number	Sale (Dollars)	Category Name
0	03/27/2012	4222	50707	7.0	10	600	4.96	10	115.0	74.40	CANADIAN WHISKIES
1	03/06/2012	2591	50022	15.0	12	750	5.23	12	115.0	94.08	CANADIAN WHISKIES
2	04/09/2015	2806	52732	23.0	6	1750	13.71	6	259.0	127.14	STRAIGHT BOURBON WHISKIES
3	01/30/2013	4197	51054	97.0	12	750	13.54	12	85.0	243.72	TENNESSEE WHISKIES
4	03/18/2014	4131	50322	77.0	12	750	12.25	12	260.0	220.56	SCOTCH WHISKIES

Figura 3: Head de la base de datos.

El resumen estadístico de las columnas numéricas muestra que el dataset cuenta con 8,389,606 registros, con una gran variedad de valores únicos en varias columnas, por ejemplo, 3492 fechas únicas y 1527 códigos postales diferentes. En cuanto a las variables numéricas, se observa que el volumen de las botellas tiene una media cercana a 893 ml, lo que coincide con tamaños comunes en botellas de whisky. El número de paquetes vendidos varía entre 1 y 48, con una media de alrededor de 12 botellas por paquete. Los costos estatales por botella muestran una amplia dispersión, con valores mínimos en cero y máximos muy elevados, lo que podría indicar la presencia de outliers o errores en los datos. Las ventas en dólares presentan también valores negativos mínimos, lo que podría reflejar errores de captura, y valores máximos muy altos, lo que indica que algunas ventas son excepcionalmente grandes. La columna de categoría presenta 10 valores únicos, siendo `Canadia Whiskies` la categoría más frecuente.

```
data_elegida_reset.describe(include="all")
```

	Date	Store Number	Zip Code	County Number	Pack	Bottle Volume (ml)	State Bottle Cost	Bottles Sold	Vendor Number	Sale (Dollars)	Category Name
count	8389606	8.389606e+06	8367943.0	8.376930e+06	8.389606e+06	8.389606e+06	8.389606e+06	8.389606e+06	8.389602e+06	8.389606e+06	8389606
unique	3492	NaN	1527.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	10
top	12/27/2022	NaN	50010.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	CANADIAN WHISKIES
freq	5357	NaN	120289.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3012895
mean	NaN	4.008763e+03	NaN	5.696741e+01	1.207741e+01	8.938504e+02	1.253402e+01	9.794379e+00	2.135503e+02	1.687672e+02	NaN
std	NaN	1.430776e+03	NaN	2.734576e+01	7.990236e+00	5.250608e+02	1.651954e+01	2.628807e+01	1.464432e+02	5.954848e+02	NaN
min	NaN	2.106000e+03	NaN	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	-6.480000e+02	1.000000e+01	-4.801680e+03	NaN
25%	NaN	2.626000e+03	NaN	3.100000e+01	6.000000e+00	7.500000e+02	6.810000e+00	3.000000e+00	8.500000e+01	4.047000e+01	NaN
50%	NaN	3.943000e+03	NaN	6.200000e+01	1.200000e+01	7.500000e+02	1.045000e+01	6.000000e+00	2.590000e+02	9.240000e+01	NaN
75%	NaN	4.868000e+03	NaN	7.700000e+01	1.200000e+01	1.000000e+03	1.599000e+01	1.200000e+01	2.600000e+02	1.770000e+02	NaN
max	NaN	1.058500e+04	NaN	9.900000e+01	4.800000e+01	1.890000e+05	1.843600e+04	1.195200e+04	9.780000e+02	2.795573e+05	NaN

Figura 4: Describe de la base de datos.

La estructura del dataset muestra 11 columnas con un total de 8,389,606 filas. Tres columnas están en formato objeto: la fecha (`Date`), el código postal (`Zip Code`) y la categoría (`Category Name`). Cuatro columnas están en formato entero (`Store Number`, `Pack`, `Bottle Volume (ml)`, `Bottles Sold`), mientras que otras cuatro están en formato de punto flotante (`County Number`, `State Bottle Cost`, `Vendor Number`, `Sale (Dollars)`).

```
data_elegida_reset.info(verbose=True)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8389606 entries, 0 to 8389605
Data columns (total 11 columns):
 #   Column              Dtype
---  ------              -----
 0   Date                object
 1   Store Number        int64
 2   Zip Code            object
 3   County Number       float64
 4   Pack                int64
 5   Bottle Volume (ml)  int64
 6   State Bottle Cost   float64
 7   Bottles Sold        int64
 8   Vendor Number       float64
 9   Sale (Dollars)      float64
 10  Category Name       object
dtypes: float64(4), int64(4), object(3)
memory usage: 704.1+ MB
```

Figura 5: Info de la base de datos.

II-C. ANÁLISIS DE LAS DIMENSIONES DE CALIDAD DE DATOS

Para asegurar la confiabilidad y validez del análisis, se realizó un estudio cuidadoso de las principales dimensiones de calidad de datos en el dataset. Cada dimensión fue evaluada para identificar posibles problemas que pudieran afectar el modelado y las conclusiones.

Completitud: Se identificaron datos faltantes importantes en dos columnas clave: 21,663 valores nulos en la columna `Zip Code` y 2,012,676 valores nulos en `County Number`.

Conformidad : Se verificó que las fechas estuvieran en el formato correcto, encontrándose que todas cumplían con el estándar esperado.

Consistencia: Respecto a la columna `County Number`, se validó que sus valores fueran válidos según las reglas del estado de Iowa, es decir, que estuvieran entre 1 y 99. No se encontraron valores inválidos ni nulos.

Para `Zip Code`, se revisó que cada código tuviera cinco dígitos (zip codes en Estados Unidos tienen 5 dígitos) y que correspondiera a códigos postales válidos en Iowa, que comienzan con “50” o “52” y están en el rango de 50001 a 52809. Se detectaron 3,435,123 códigos postales mal formateados.

Precisión/Exactitud: Se examinó que no existieran números negativos en columnas donde estos carecen de sentido lógico, tales como `Store Number`, `County Number`, `Pack`, `Bottle Volume (ml)`, `State Bottle Cost`, `Bottles Sold`, `Vendor Number` y `Sale (Dollars)`.

La presencia de valores negativos en estas variables sería indicativa de errores en la captura o registro, ya que, por ejemplo, no puede haber una cantidad negativa de botellas vendidas o un costo negativo.

También se verificó que ciertas columnas no tuvieran valores de cero, dado que estas variables reflejan cantidades o costos que, en el contexto del negocio, no pueden ser nulos. `Bottle Volume (ml)` no puede ser cero porque cada botella debe tener un volumen definido. `State Bottle Cost` no puede ser cero porque cada botella tiene un costo asociado. `Bottles Sold` no puede ser cero porque una venta implica al menos una unidad. `Sale (Dollars)` no puede ser cero porque representa el monto total de la venta. En total, se detectaron 491 filas con ceros en estas columnas, lo que motivó una revisión para mejorar la calidad de los datos.

Unicidad: Se identificaron 48,113 filas completamente idénticas en el dataset.

Detección de outliers: Para identificar valores extremos que pudieran distorsionar los análisis y modelos, se aplicaron los métodos de Z-score y Distancia de Mahalanobis.

Z-score: Este método mide cuántas desviaciones estándar se encuentra un dato respecto a la media de su variable. Los valores con z-score muy altos o bajos se consideran outliers. Se aplicó a columnas `Store Number`, `County Number`, `Pack`, `Bottle Volume (ml)`, `State Bottle Cost`, `Bottles Sold`, `Vendor Number` y `Sale (Dollars)`. Con este método se detectaron 146,495 outliers.

Distancia de Mahalanobis: Esta medida estadística toma en cuenta la correlación entre variables para identificar puntos que están alejados del centro multivariado de los datos. Es útil para detectar outliers considerando todas las variables simultáneamente. Se calculó usando todas las columnas del dataset. Con este método se encontraron 148,727 outliers.

Con base en este análisis inicial y el diagnóstico de calidad, se procedió a realizar las etapas de limpieza, corrección y preparación de los datos para garantizar que el conjunto final utilizado para modelar cumpliera con los estándares de calidad requeridos.

III. PREPARACIÓN DE DATOS

En esta fase, aplicamos todos los tratamientos necesarios detectados durante el entendimiento de datos para mejorar la calidad del dataset y prepararlo para el modelo.

Se identificaron valores nulos en las columnas `Zip Code` (21,663 nulos) y `County Number` (2,012,676 nulos). Estas filas fueron eliminadas para evitar sesgos o errores en el análisis geográfico.

```
Date                       0
Store Number               0
Zip Code               21663
County Number        2012676
Pack                       0
Bottle Volume (ml)         0
State Bottle Cost          0
Bottles Sold               0
Vendor Number              4
Sale (Dollars)             0
Category Name              0
dtype: int64
```

Figura 6: Visualización de los valores nulos detectados en las columnas `Zip Code` y `County Number`.

Se verificó que la columna `Date` tuviera el formato correcto (no se encontraron fechas en el formato incorrecto) y se convirtió a entero en formato YYYYMMDD para facilitar su uso.

```
Cantidad de fechas mal formateadas: 0
Fechas mal formateadas:
Series([], Name: Date, dtype: datetime64[ns])

Tipo final de 'Date': int64
       Date
0  20120327
1  20120306
2  20150409
3  20130130
4  20140318
```

Figura 7: Verificación y formato de la columna `Date`.

El `County Number` fue validado para que sólo contuviera valores entre 1 y 99, válidos para Iowa, sin encontrar valores inválidos. No se encontraron `county number` inválidos.

```
Cantidad de registros con 'County Number' inválido: 0
```

Figura 8: Validación de los valores en la columna `County Number`.

El `Zip Code` fue validado para contener cinco dígitos y estar dentro del rango 50001 a 52809, correspondiente a Iowa. Se eliminaron filas con códigos postales inválidos (3,435,123 casos).

```
Cantidad de ZIP Codes inválidos: 3435123
Ejemplos de ZIP Codes inválidos:
['712-2' '56201' '51653.0' '51534.0' '52205.0' '52601.0' '50314.0'
 '51241.0' '51501.0' '50662.0' '50833.0' '51443.0' '52087.0' '50317.0'
 '50676.0' '52637.0' '52761.0' '50058.0' '50010.0' '52806.0' '50266.0'
 '50651.0' '51050.0' '51334.0' '50613.0' '50647.0' '52245.0' '51106.0'
 '52732.0' '52302.0' '50472.0' '51054.0' '50616.0' '50313.0' '50315.0'
```

Figura 9: Inspección y limpieza de códigos postales en la columna `Zip Code`.

Se revisaron columnas donde los valores negativos no tienen sentido lógico, tales como `Store Number`, `County Number`, `Pack`, `Bottle Volume (ml)`, `State Bottle Cost`, `Bottles Sold`, `Vendor Number` y `Sale (Dollars)`. No se encontraron valores negativos, lo cual garantiza la coherencia de los datos.

```
'Store Number' no tiene valores negativos
'County Number' no tiene valores negativos
'Pack' no tiene valores negativos
'Bottle Volume (ml)' no tiene valores negativos
'State Bottle Cost' no tiene valores negativos
'Bottles Sold' no tiene valores negativos
'Vendor Number' no tiene valores negativos
'Sale (Dollars)' no tiene valores negativos
```

Figura 10: Revisión de la ausencia de valores negativos en columnas críticas.

En las columnas `Bottle Volume (ml)`, `State Bottle Cost`, `Bottles Sold` y `Sale (Dollars)`, se detectaron 491 filas con valores cero, lo que no tiene sentido en el contexto (una botella no puede tener volumen cero, ni puede haber venta o costo nulo). Estas filas fueron eliminadas para asegurar la precisión de los datos.

```
Filas con 0 en columnas clave: 491
```

Figura 11: Identificación y eliminación de filas con valores cero en variables donde estos no tienen sentido lógico.

Se detectaron y eliminaron 48,113 filas completamente idénticas para evitar sesgos en el análisis.

```
Filas duplicadas exactamente: 48113
```

Figura 12: Detección y eliminación de filas duplicadas.

Se aplicó `One Hot Encoding` a la columna `Category Name` para convertir las categorías de whisky en variables booleanas. Posteriormente, se eliminaron las categorías sin presencia en el dataset limpio, quedando seis tipos de whisky para el análisis final:

- Canadian Whiskies
- Straight Bourbon Whiskies
- Tennessee Whiskies
- Scotch Whiskies
- Blended Whiskies
- Irish Whiskies

```
Category_BLENDED WHISKIES              bool
Category_CANADIAN WHISKIES             bool
Category_IRISH WHISKIES                bool
Category_SCOTCH WHISKIES               bool
Category_STRAIGHT BOURBON WHISKIES     bool
Category_TENNESSEE WHISKIES            bool
```

Figura 13: Proceso de transformación de la columna `Category Name` mediante One Hot Encoding.

Se aplicaron dos métodos complementarios para identificar valores atípicos:

`Z-score`: Evaluó la desviación estándar de valores en columnas numéricas clave, detectando 146,495 outliers, los cuales se eliminaron.

```
Filas con al menos un outlier: 146495
```

Figura 14: Identificación de valores atípicos utilizando el método Z-score en columnas numéricas clave.

`Distancia de Mahalanobis`: Identificó outliers multivariados considerando la relación entre variables, detectando 148,727 casos, los cuales se eliminaron.

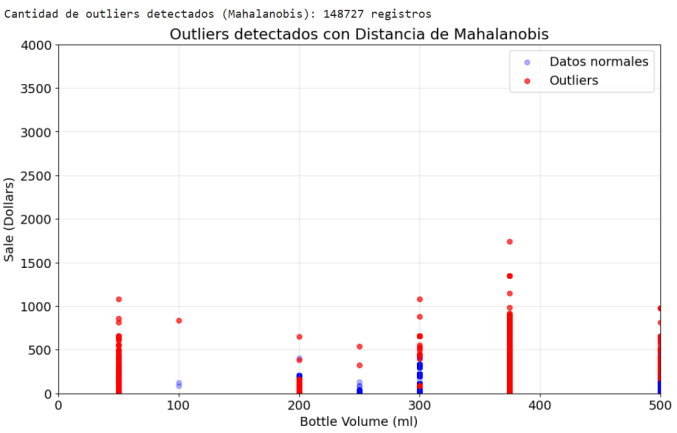

Figura 15: Detección de outliers multivariados mediante Distancia de Mahalanobis.

Ambos métodos ayudaron a asegurar que el dataset fuera consistente y sin valores extremos que pudieran afectar el modelado.

Luego de limpiar la base de datos, el dataframe `data_limpia` contiene un total de 2,597,977 registros dis-

tribuidos en 16 columnas con tipos de datos variados, incluyendo 6 variables booleanas, 6 variables enteras (`int64`) y 4 variables de punto flotante (`float64`).

```
data_limpia.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2597977 entries, 0 to 2597976
Data columns (total 16 columns):
 #   Column                          Dtype  
---  ------                          -----  
 0   Date                            int64  
 1   Store Number                    int64  
 2   Zip Code                        int64  
 3   County Number                   float64
 4   Pack                            int64  
 5   Bottle Volume (ml)              int64  
 6   State Bottle Cost               float64
 7   Bottles Sold                    int64  
 8   Vendor Number                   float64
 9   Sale (Dollars)                  float64
 10  Category_BLENDED WHISKIES       bool   
 11  Category_CANADIAN WHISKIES      bool   
 12  Category_IRISH WHISKIES         bool   
 13  Category_SCOTCH WHISKIES        bool   
 14  Category_STRAIGHT BOURBON WHISKIES  bool
 15  Category_TENNESSEE WHISKIES     bool   
dtypes: bool(6), float64(4), int64(6)
memory usage: 213.1 MB
```

Figura 16: Info del dataframe tras la limpieza del dataset.

El dataframe `data_limpia` presenta una estructura sólida y coherente en todas sus columnas numéricas. La columna `Date` contiene valores enteros que reflejan un amplio rango temporal. `Store Number` y `Zip Code` muestran identificadores numéricos consistentes, propios de múltiples tiendas distribuidas en una región geográfica delimitada. La columna `County Number`, presenta valores enteros dentro de un rango lógico que corresponde a la codificación de condados. En cuanto a las variables relacionadas con el producto, `Pack` varía entre 1 y 24 botellas, lo que concuerda con presentaciones comerciales típicas, mientras que `Bottle Volume (ml)` abarca desde tamaños pequeños hasta botellas grandes, con valores dentro de límites esperados. El `State Bottle Cost` muestra un rango de precios plausible para los distintos tipos de whisky, y `Bottles Sold` tiene una distribución adecuada, con un rango de 1 a 54 botellas vendidas por orden, reflejando comportamientos de compra realistas. La variable `Vendor Number`, que identifica al proveedor, se mantiene dentro de valores esperados y estables. Finalmente, la variable `Sale (Dollars)` varía entre 2.8 y 935.52 dólares, con un promedio de 107.39 dólares. Por otro lado, las columnas categóricas, codificadas en formato booleano mediante one-hot encoding, reflejan correctamente la pertenencia a las diferentes categorías de whisky.

	Date	Store Number	Zip Code	County Number	Pack	Bottle Volume (ml)	State Bottle Cost	Bottles Sold	Vendor Number	Sale (Dollars)
count	2.597977e+06	2.597977e+06	2.597977e+06	2.597977e+06	2.597977e+06	2.597977e+06	2.597977e+06	2.597977e+06	2.597977e+06	2.597977e+06
unique	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	2.014514e+07	3.594571e+03	5.124982e+04	5.656453e+01	1.133207e+01	9.925169e+02	1.048779e+01	7.444345e+00	1.788084e+02	1.073903e+02
std	1.946596e+04	9.113011e+02	9.753994e+02	2.745561e+01	5.086375e+00	4.928584e+02	5.986495e+00	6.423915e+00	1.254818e+02	1.052346e+02
min	2.012010e+07	2.106000e+03	5.002200e+04	1.000000e+00	1.000000e+00	5.000000e+01	1.870000e+00	1.000000e+00	3.500000e+01	2.800000e+00
25%	2.013060e+07	2.613000e+03	5.031600e+04	3.100000e+01	6.000000e+00	7.500000e+02	6.550000e+00	2.000000e+00	6.500000e+01	3.312000e+01
50%	2.014102e+07	3.723000e+03	5.105000e+04	6.000000e+01	1.200000e+01	7.500000e+02	9.090000e+00	6.000000e+00	1.150000e+02	8.028000e+01
75%	2.016020e+07	4.312000e+03	5.224800e+04	7.700000e+01	1.200000e+01	1.750000e+03	1.375000e+01	1.200000e+01	2.600000e+02	1.261200e+02
max	2.022072e+07	6.313000e+03	5.280700e+04	9.900000e+01	2.400000e+01	1.750000e+03	3.998000e+01	5.400000e+01	5.780000e+02	9.355200e+02

Figura 17: Describe del del dataframe `Data Limpia` parte 1.

Category_BLENDED WHISKIES	Category_CANADIAN WHISKIES	Category_IRISH WHISKIES	Category_SCOTCH WHISKIES	Category_STRAIGHT BOURBON WHISKIES	Category_TENNESSEE WHISKIES
2597977	2597977	2597977	2597977	2597977	2597977
2	2	2	2	2	2
False	False	False	False	False	False
2129799	1610743	2517706	2387253	1991220	2353164
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN

Figura 18: Describe del del dataframe `Data Limpia` parte 2.

El análisis estadístico confirma que la base de datos fue limpiada y preparada adecuadamente, con variables numéricas en rangos consistentes y variables categóricas correctamente transformadas, asegurando así una estructura lista para el desarrollo de modelos de machine learning y análisis posteriores.

También se calculó la matriz de correlación del dataset (ver el codigo) para analizar la relación entre las variables y la variable objetivo, `Sale (Dollars)`. A continuación se puede observar las correlaciones respecto a `Sale (Dollars)`.

```
Sale (Dollars)                      1.000000
Bottles Sold                        0.712145
State Bottle Cost                   0.393429
Bottle Volume (ml)                  0.186819
Category_TENNESSEE WHISKIES         0.129432
Category_CANADIAN WHISKIES          0.051817
Date                                0.038983
Category_IRISH WHISKIES             0.038742
County Number                       0.012101
Zip Code                           -0.002441
Category_SCOTCH WHISKIES           -0.017010
Vendor Number                      -0.028515
Category_STRAIGHT BOURBON WHISKIES -0.028637
Store Number                       -0.065792
Category_BLENDED WHISKIES          -0.137655
Pack                               -0.156568
Name: Sale (Dollars), dtype: float64
```

Figura 19: Correlación respecto a `Sale (Dollars)`.

Posteriormente, se identificaron las nueve variables con mayor correlación positiva respecto a `Sale (Dollars)`. Tambien se creó un heatmap que visualiza las relaciones de manera clara y sintética.

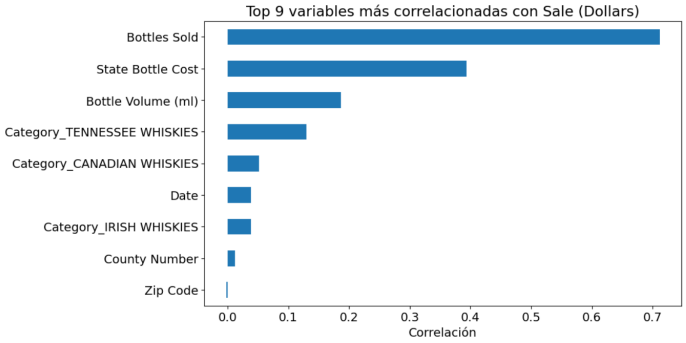

Figura 20: Gráfica de barras.

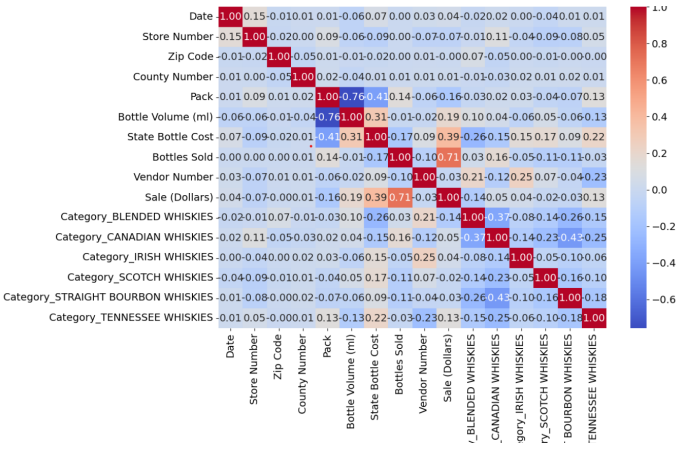

Figura 21: Heatmap.

Se observa que la variable que más se correlaciona con `Sale (Dollars)` es `Bottles Sold`, con una correlación positiva fuerte de aproximadamente 0.71. Esto indica que, a mayor número de botellas vendidas, mayor es el valor total de las ventas, lo cual es coherente con el conocimiento del negocio. Le siguen en importancia `State Bottle Cost` (precio al que el estado compra cada botella), con una correlación moderada cercana a 0.39, y `Bottle Volume (ml)`, con una correlación más baja pero aún positiva de aproximadamente 0.19, lo que sugiere que el tamaño de la botella también influye en las ventas, aunque en menor medida.

IV. MODELACIÓN

Una vez terminado el proceso de preparación de datos, nos encargamos de entrenar y escoger el mejor modelo para predecir la variable `Sale (Dollars)`, esto basándonos en el criterio de éxito propuesto anteriormente, en donde el MAE sea menor a un dólar de error.

En ese orden de ideas, `Sale (Dollars)` va a ser nuestra variable de salida, y las características con las que vamos a predecirla son `Date`, `Store Number`, `Zip Code`, `County Number`, `Pack`, `Bottle Volume (ml)`, `State Bottle Cost`, `Bottles Sold`, `Vendor Number`, `Category_BLENDED WHISKIES`, `Category_CANADIAN WHISKIES`, `Category_IRISH WHISKIES`, `Category_SCOTCH WHISKIES`, `Category_STRAIGHT BOURBON WHISKIES`, `Category_TENNESSEE WHISKIES`

	Date	Store Number	Zip Code	County Number	Pack	Bottle Volume (ml)	State Bottle Cost	Bottles Sold	Vendor Number	Category_BLENDED WHISKIES	Category_CANADIAN WHISKIES	Category_IRISH WHISKIES	Category_SCOTCH WHISKIES	Category_STRAIGHT BOURBON WHISKIES	Category_TENNESSEE WHISKIES
0	20120227	4222	50707	7.0	10	600	4.96	10	115.0	False	False	False	False	False	False
1	20120306	2591	50022	15.0	12	750	5.23	12	115.0	False	True	False	False	False	False
2	20150409	2806	52732	23.0	6	1750	13.71	6	259.0	False	False	False	False	True	False
3	20130130	4197	51054	97.0	12	750	13.54	12	85.0	False	False	False	False	False	True
4	20140318	4131	50322	77.0	12	750	12.25	12	260.0	False	False	False	True	False	False

Figura 22: Características de entrenamiento de datos

IV-A. DIVISIÓN Y ENTRENAMIENTO DE LOS DATOS

Luego de definir con que variables de nuestra data vamos a trabajar para realizar el modelado, definimos la estructura de aprendizaje que va a tener nuestro modelo de Machine Learnig así, definimos una semilla aleatoria fija asegurar la productividad del entrenamiento de los datos. Luego, dividimos el conjunto de los datos destinando así 1529636 entrenamiento (50 %) y 764818 para validación (25 %) y prueba (25 %) para la realización del modelado empleando la libreria `Scikit-Learn` para evaluar multiples modelos que nos permitan predecir de la mejor forma `Sale (Dollars)`.

IV-B. Entrenamiento de modelos para la predicción

Para buscar el mejor modelo para la predicción nos enfocamos en evaluar 330 árboles de decisión, 180 random forest y 125 gradientes; todos los modelos evaluados fueron realizados con base en la regresión. Además, se aplicó la estrategia computacional de computación en paralelo de la librería `joblit` donde se importó `Parallel, delayed` dado el

costo operativo que con lleva la volumetría de nuestros datos para agilizar tiempos y procesos de lectura y entrenamiento de datos.

Se importó de la librería `Scikit-Learn` el tipo de modelo de aprendizaje de los cuales, se dividió en ciclos para buscar el mejor modelo donde los hiperparámetros están definidos como rangos o listas de valores que puede tomar el modelo usado. Además se llamaron también `sklearn.metrics`, `sklearn.base` para calcular las métricas de los modelos escogidos (`MAE`, `MSE`, `MAPE`) y para clonar el modelo.

IV-B1. Arboles de decisión

Importamos `DecisionTreeRegressor` para buscar el mejor modelo. Para realizar este proceso, se dividió la búsqueda en dos ciclos.

Para el primer ciclo, se usaron los hiperparámetros `max depth`, `min samples split`, `min samples leaf` con los rangos (5, 10), (2, 14, 2), (5, 10) respectivamente donde se evaluaron 150 modelos. Luego, el mejor modelo encontrado por el ciclo fue

```
DecisionTreeRegressor
DecisionTreeRegressor(max_depth=9, min_samples_leaf=9, random_state=42)
```

Figura 23: hiperparámetros del `DecisionTreeRegressor1` con mejor MAE.

```
Metricas mejor árbol 1:
MSE: 19.007987907504017
MAE: 1.3515385351519469
MAPE: 0.024215615315658107
```

Figura 24: Métricas del `DecisionTreeRegressor1`.

Para el segundo ciclo se usaron los hiperparámetros `max depth`, `max features`, `ccp alpha` con los valores (6, 11), ['sqrt', 'log2', None, 0.5, 0.8], [0.0, 0.01, 0.02, 0.03, 0.05, 0.1] respectivamente donde se evaluaron 180 modelos. Despues, el mejor modelo del ciclo es

```
DecisionTreeRegressor
DecisionTreeRegressor(max_depth=10, random_state=42)
```

Figura 25: hiperparámetros del `DecisionTreeRegressor2` con mejor MAE

```
Metricas mejor árbol 2:
MSE: 15.533907673872596
MAE: 0.8402940681153099
MAPE: 0.014720105151997181
```

Figura 26: Métricas del `DecisionTreeRegressor2`

IV-B2. Random forest

Importamos `RandomForestRegressor` para la busqueda del mejor modelo. Además, se estableció para la programación en paralelo en el ciclo donde se escoje el modelo con menor MAE evaluado `Parallel(n jobs=-1)` y como hiperparámetro fijo de los bosques `n jobs=2` esto, para agilizar los tiempos de computación y de entrenamiento del modelo. Similarmente al caso anterior se dividió la busqueda en dos ciclos.

Para el primer ciclo se usaron los hiperparámetros `max depth`, `n stimators`, `min samples split` en rangos (40, 43), (60, 80,4) y (3, 5) respectivamente donde se evaluaron 30 modelos. El modelo escogido con menor MAE fue

```
RandomForestRegressor
RandomForestRegressor(max_depth=41, min_samples_split=3, n_estimators=76,
n_jobs=2, random_state=42)
```

Figura 27: hiperparámetros del `RandomForestRegressor1` con mejor MAE.

```
Metricas mejor random forest 1:
MSE: 9.69856848540296
MAE: 0.13088736928232167
MAPE: 0.003377570657279555
```

Figura 28: Métricas del `RandomForestRegressor1`.

Luego, para el segundo ciclo usaron los hiperparámetros `max depth`, `n stimators`, `min samples leaft` en rangos (7, 12), (10, 20) y (3, 6) respectivamente donde se evaluaron 150 modelos. El modelo escogido con menor MAE fue

```
RandomForestRegressor
RandomForestRegressor(max_depth=11, min_samples_leaf=3, n_estimators=19,
n_jobs=2, random_state=42)
```

Figura 29: hiperparámetros del `RandomForestRegressor2` con mejor MAE.

```
Metricas random forest 2:
MSE: 12.178389340215068
MAE: 0.5217685843083973
MAPE: 0.00944705552946948
```

Figura 30: Métricas del `RandomForestRegressor2`.

IV-C. Gradient Boostig

Importamos `HistGradientBoostingRegressor` ya que aprovecha el "binning" de características en histogramas para acelerar el cálculo de gradientes y la búsqueda de cortes, lo que reduce memoria y tiempo de entrenamiento sin sacrificar precisión. Por lo que, descartamos el uso `GradientBoostingRegressor` ya que significa un costo computacional enorme por lo que nos limitaba y no nos permitía encontrar de forma medianamente rapida y efectiva una solución así, nos vimos forzados a usar una variante del mismo. Además, como en el caso anterior para decidir cual

```
Metricas modelo elegido en test:
MSE: 9.448727286614789
MAE: 0.1155178905513807
MAPE: 0.0029281052110591193
```

Figura 34: Métricas del `modelo_final`.

fue el mejor modelo se usó `Parallel(n jobs=-1)` para óptimizar la elección.

Para este caso, sólo se uso un ciclo con los hiperparámetros `max depth, learning rate, max iter` con valores [5, 7, 10, 12, 15], [0.02, 0.03, 0.05, 0.1, 0.15] y [100, 300, 500, 700, 1000] respectivamente; además, se usaron algunos hiperparámetros fijos `min samples leaf=60, max iter no change=300, early stopping=True, max features=0.8` donde se evalúan 125 modelos. El mejor modelo encontrado en el ciclo es

```
HistGradientBoostingRegressor
HistGradientBoostingRegressor(early_stopping=True, learning_rate=0.03,
                              max_depth=10, max_features=0.8, max_iter=10000,
                              min_samples_leaf=60, n_iter_no_change=300,
                              random_state=42)
```

Figura 31: hiperparámetros del `HistGradientBoostingRegressor` con mejor MAE.

```
Metricas gradient boosting 1:
MSE: 11.16470083115123
MAE: 0.40031211457092536
MAPE: 0.008924377403739286
```

Figura 32: Métricas de `HistGradientBoostingRegressor`.

IV-C1. Modelo elegido

El modelo que predecia mejor `Sale (Dollars)` fue `RandomForestRegressor1` ya que es el que tenia el MAE más pequeño de todos ($\approx$ 13 centavos) por lo que, se decidió clonar este modelo y se reentreno con todos los datos de validación y entrenamiento y el entrenamiento de este modelo se definio para la aceleración de los tiempos de entrenamiento `modelo_final.set_params(n_jobs=-1)` así, el modelo reentrenado tiene los hiperpárametros y métricas.

```
RandomForestRegressor
RandomForestRegressor(max_depth=41, min_samples_split=3, n_estimators=76,
                      n_jobs=-1, random_state=42)
```

Figura 33: Hiperparámetros del `modelo_final`.

Vemos que las métricas son relativamente parecidas a las de validación, lo cual indica que el modelo está generalizando correctamente con nuevos datos que no ha visto.

Estas métricas de test (prueba) son las que sí representan realmente cómo se comportará nuestro modelo a futuro, y son muy importantes para el Negocio puesto que nos ayudan a saber qué esperamos que pase en el futuro (ejemplo: el MAE nos dirá en promedio 11 centavos al predecir el precio de los Whiskys.

V. EVALUACIÓN

En esta sección vamos a calcular y visualizar errores como el MAE, MAPE y MSE que obtenemos al comparar la predicción con nuestros datos de validación, para así determinar si nuestro modelo está teniendo un buen desempeño.

V-A. Visualización del error absoluto

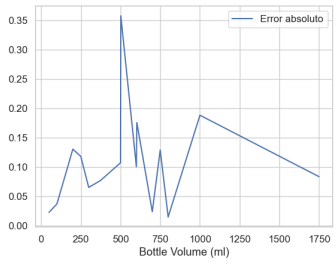

Figura 35: Volumen

Podemos ver que el modelo presenta un error bajo en volúmenes muy pequeños (¡200 ml) y muy grandes (¿1 500 ml), pero presenta picos en torno a 550 ml (0,35 USD) y 1 000 ml (0,19 USD), lo que indica que el modelo falla especialmente con esos formatos intermedios.

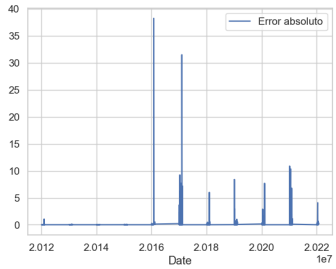

Figura 36: Fecha

El error absoluto es casi nulo en la mayor parte del periodo, pero aparecen picos muy pronunciados alrededor de 2016 (38 USD) y 2017 (31 USD), y otros más moderados en 2020–2021 (5–11 USD).

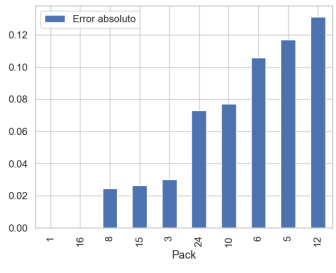

Figura 37: Pack

El modelo acierta perfectamente en los packs 1 y 16 (error 0), mantiene un MAE muy bajo en packs intermedios

pequeños (3, 8, 15: ¡0,03 USD) y exhibe errores crecientes en packs más grandes (24 → 0,07; 10 → 0,08; 6 → 0,11; 5 → 0,12; 12 → 0,13 USD), lo que indica que el tamaño de pack influye en la precisión. Sin embargo, el error presentado por el modelo es muy pequeño dado que esta en centavos.

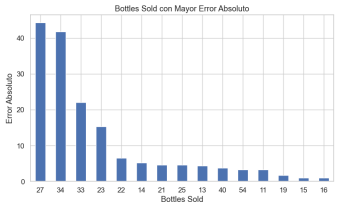

Figura 38: Botellas

La gran mayoría de los recuentos (11 de 15) presentan errores muy bajos (¡5 USD), mientras que solo 27 y 34 botellas generan picos de 45 USD y 42 USD, seguidos de 33 y 23 botellas errores moderados (22 USD y 15 USD).

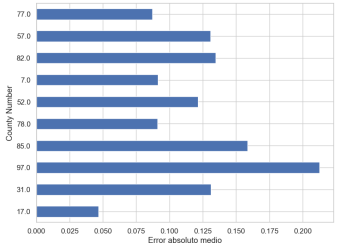

Figura 39: Condado

- Cada barra muestra el error absoluto medio de predicción para un condado (según frecuencia de registros).
- Los condados **30** y **56** registran el mayor error (0.24–0.25), indicando predicciones menos fiables.
- El condado **50** tiene el error más bajo (0.02), con predicciones muy ajustadas.
- La mayoría de los condados se agrupa en un rango de **0.10 - 0.18** de error medio, sugiriendo un desempeño estable.

V-B. Residuos

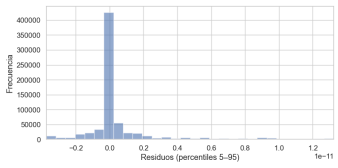

Figura 40: Histograma 1

- La **gran mayoría** de los residuos se agrupa muy cerca de cero (columna central alta), lo que indica que el modelo predice el precio con **muy poca desviación media**.
- La distribución es prácticamente simétrica y concentra su "cúpula" alrededor de cero, sugiriendo ausencia de sesgo sistemático.
- Al truncar extremos (5 % inferiores y superiores), vemos que los valores remanentes quedan muy apretados, demostrando una alta precisión en las predicciones.
- Sólo hay pocos residuos más alejados, lo que confirma que los outliers son excepcionales y no afectan la calidad global del ajuste.

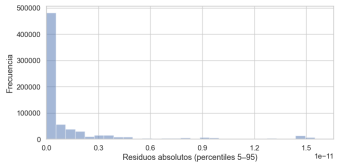

Figura 41: Histograma 2

- La mayoría de los residuos absolutos se concentra muy cerca de cero, mostrando que el error medio es prácticamente insignificante en la mayoría de las predicciones.
- La distribución está muy sesgada a la derecha, con una "cola" de pocos valores más grandes, lo que indica algunos casos donde el modelo se desvía más.
- La alta altura de la primera barra refleja que decenas de miles de predicciones caen en el rango mínimo de error.
- Al eliminar los extremos, se confirma que los errores grandes son raros y no afectan el ajuste global.

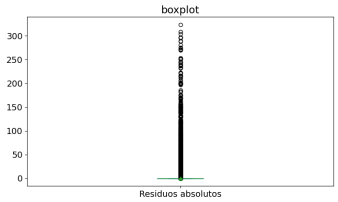

Figura 42: Box-plot

- La mediana de los residuos absolutos está muy cerca de cero, indicando que la mitad de los errores son prácticamente insignificantes.
- La caja es muy estrecha, mostrando que la mayoría de los errores se concentran en un rango muy pequeño. - Aparecen multitud de outliers hacia arriba, lo que revela casos puntuales con errores absolutos elevados.
- La gran cantidad de puntos externos sugiere que, aunque el modelo es muy preciso en la mayoría de los casos, existen observaciones aisladas con desviaciones considerables.

En conjunto, este comportamiento de los residuos respalda que el modelo de Random Forest (rf2) es adecuado para la predicción de precios, con alta exactitud y sin sesgos significativos, dejando espacio para mejorar solo en casos aislados.

V-C. Intervalos de confianza

Teniendo en cuenta que n ¿30, por el teorema del límite central, un intervalo de confianza al 99 % para el MAE es:

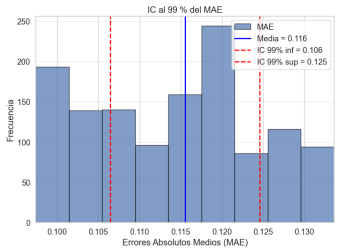

Figura 43: MAE

Con un 99 % de confianza, el MAE real de nuestro modelo se sitúa entre 0,106 y 0,125; ese intervalo estrecho alrededor de la media (0,116) denota baja variabilidad y alta consistencia en las predicciones, confirmando así un rendimiento muy sólido y preciso.

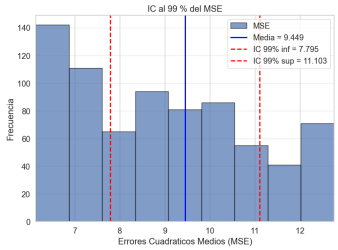

Figura 44: MSE

Con un 99 % de confianza, el MSE verdadero de nuestro modelo se sitúa entre 7,795 y 11,103; aunque este intervalo (3,308) es más amplio que el del MAE, sigue reflejando una variabilidad moderada y confirma la solidez y consistencia del desempeño en términos de error cuadrático medio.

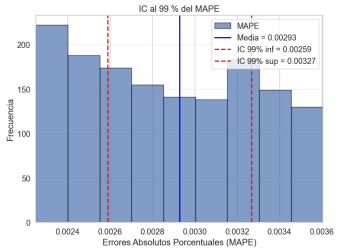

Figura 45: MAPE

Con un 99 % de confianza, el MAPE verdadero de nuestro modelo se sitúa entre 0,00259 y 0,00327; este intervalo estrecho alrededor de la media (0,00293) indica muy poca variabilidad porcentual y respalda la consistencia y precisión de las predicciones relativas.

VI. DESPLIEGUE

Después de haber obtenido nuestro mejor modelo y haberlo guardado en un archivo pickle, creamos un nuevo notebook de Python en donde realizamos una aplicación con interfaz interactiva para que los usuarios puedan ingresar los datos característicos de la venta de Whisky que van a realizar con el fin de que el modelo implementado les prediga el precio esperado de la compra para que así puedan ajustar sus ventas y sacar las mejores ganancias posibles.

Para realizar esta aplicación utilizamos la librería **Tkinter** que nos permite una interfaz gráfica.

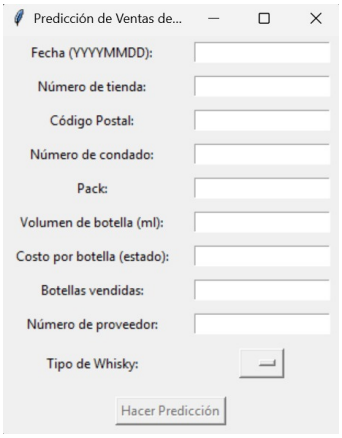

Figura 46: Interfaz del programa

Para realizar este programa, primero importamos nuestro modelo obtenido en la fase anterior. Se implementan funciones para recibir los datos de los usuarios, ejecutar la predicción del precio del Whisky y el despliegue y aspecto de la interfaz.

Definimos también función de validación de datos que nos ayude a garantizar un buen manejo del modelo y los datos que se usarán para la predicción, esto en relación a los formatos que deben cumplir varias variables como la no negatividad en las variables `Store Number`, `Pack`, `Bottle Volume` entre otras, y los rangos adecuados para `Date`, `Zip Code` y `County Number` como se vio en la sección de validación de los datos. En las siguientes figuras podemos ver ejemplos de cómo el programa no nos deja ejecutar la predicción si tenemos datos inválidos.

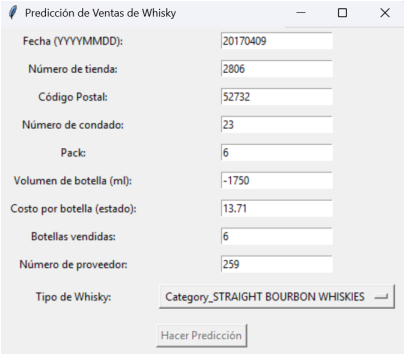

Figura 47: Volumne de botella negativo

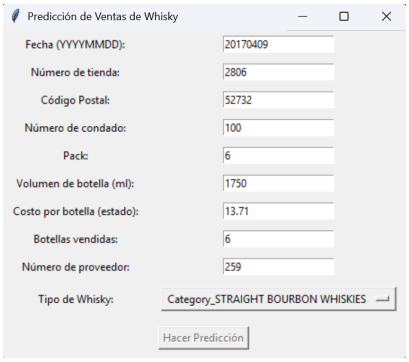

Figura 48: Condado fuera del rango permitido (1-99)

Por último, podemos ver un ejemplo de un buen ingreso de los datos al programa y su predicción de precio:

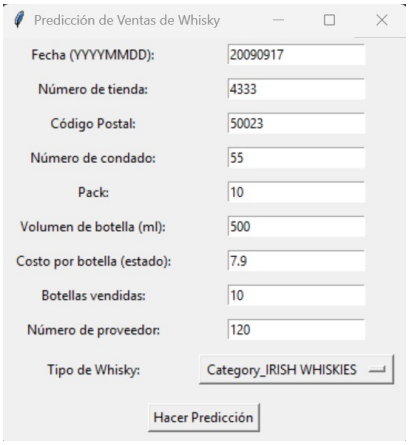

Figura 49: Datos de venta para la predición

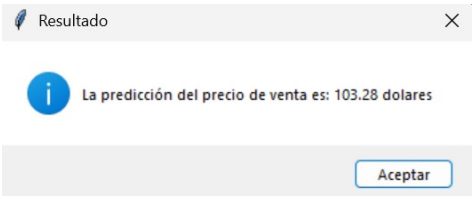

Figura 50: Predicción precio de venta del Whisky

REFERENCIAS

[1] T. D. Business, *US alcohol industry set for slow recovery after reset year in 2023*, Accessed: 2024-05-15, 2024. dirección: https://www.thedrinksbusiness.com/2024/12/us-alcohol-industry-set-for-slow-recovery-after-reset-year-in-2023/?utm_source=chatgpt.com.

[2] A. R. Treatment, *History of alcohol addiction in America*, Accessed: 2024-05-15, n.d. dirección: https://axisresidentialtreatment.com/alcohol-addiction/history-in-america/?utm_source=chatgpt.com.

[3] A. B. D. (Commerce), *Iowa Liquor Sales*, Última actualización de metadatos: 5 de abril de 2025, 2025. dirección: https://catalog.data.gov/dataset/iowa-liquor-sales.