Avance 1. Análisis Exploratorio de Datos

PROYECTO INTEGRADOR: CO2 Mass Balance

PRESENTADO POR:

Miguel Ángel Aguilera Rodríguez - A00642541

Carlos Jesús Peñaloza Julio - A01793931

Alberto Patraca Sotomayor - A01793469

PROFESORA:

Dra. Grettel Barceló Alonso

Dra. Eduviges Ludivina Facundo Flores

MAESTRÍA EN INTELIGENCIA ARTIFICIAL APLICADA

TECNOLÓGICO DE MONTERREY

FACULTAD DE INGENIERÍA Y CIENCIAS

SEPTIEMBRE 29 DE 2024

Tabla de contenido

Tabla de figuras

1. Descripción General de los Datos

Durante la semana, tuvimos una llamada con Andrea, la Científica de Datos enfocada en la iniciativa del proyecto de CO2 Mass Balance, dentro de Heineken.

A pesar de que ya conocíamos la problemática existente, nos faltaban algunos detalles claves que ella nos pudo proporcionar, como los objetivos que se esperan lograr al final del desarrollo.

Nos explicó la expectativa de producción de CO2 teórica que hay sobre la cerveza, la cual según comprendimos es descrita por una curva normal, donde se cuentan los días a partir de que la cerveza se deposita en los tanques, y a partir de los 4 o 5 días (dependiendo de la marca) se empieza a producir la mayor cantidad de CO2 con la calidad requerida para su reutilización.

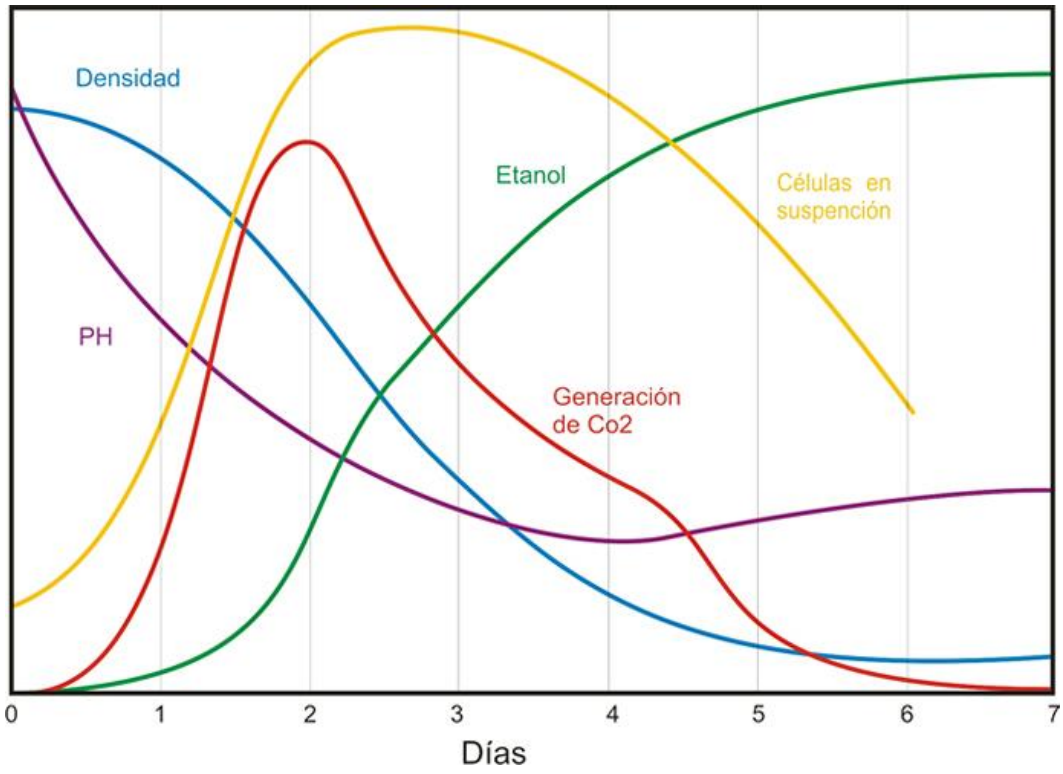

Figura 1 Diversas características de la cerveza a través del tiempo. Fuente: Tomado de https://www.cervezaartesanalsm.com.ar/fermentacion1/

El comportamiento en la Figura 1 siendo coherentes con lo ya mencionado indica que los tiempos de llenado son dependientes en gran medida a la marca, por lo que muy probablemente, los mostrados en la gráfica no son iguales a los manejados por Heineken.

También nos explicó qué información tendremos para trabajar, la cual en parte se basa, como mencionamos en la entrega anterior, con implementación de IoT. Ya que ya se tienen sensores conectados a la sección de producción de CO2, y al consumo en las líneas requeridas. Así como registros donde se conserva la información de los tanques que contienen las diferentes cervezas, sus capacidades, y varias fechas.

2. Análisis de Datos

2.1 Estructura de los datos

Describiendo más a fondo la información que nos proporcionó Andrea, esta se distribuye de la siguiente manera:

1. Un diccionario con extensión *.json* que contiene el número representativo de cada tanque como llave, y la capacidad, en litros, de cada uno, como valor.
2. Un archivo Excel, con registros manuales de los tanques, con qué cerveza (marca y tipo) que se llenaba cada uno, así como fecha y hora del inicio y fin de llenado, ya que este lleva un tiempo, también como fecha y hora de inicio y fin conexión a la red de producción y almacenamiento de CO2.
3. Registros de las mediciones (por segundo) de los sensores IoT ubicados al inicio de la sección de producción de CO2, así como al inicio de las líneas de envasado.

Con lo antes mencionado, podemos notar que hay datos relevantes, en primera instancia claro que nos interesa la capacidad de cada tanque, al igual que las diferentes cervezas que se manejan.

De entre las fechas, la conexión de los tanques al sistema de producción de CO2 siempre ocurren después del fin del llenado. Como a nosotros solo nos interesa a partir del momento en que los tanques se unen al sistema de producción, podemos descartar la información de las fechas de llenado, conservando únicamente las de inicio y fin de conexión.

En primera instancia, nos interesaban ambos registros del flujo de CO2, la producción y el consumo de este. Sin embargo, debido a la dificultad del proyecto, y división de los conjuntos de datos que pudimos notar en la junta con Andrea, acordamos que tenemos que acotar nuestros objetivos, de modo que nos enfocaremos en únicamente la producción, obviando los datos del consumo.

2.2 Análisis univariante

Para analizar el conjunto de datos relacionado con la capacidad de los tanques, se decidió hacer uso del diccionario teniendo en cuenta que la información es sencilla de observar por contar con sólo dos columnas y 50 líneas de valores, la relación fue 1:1 y se pudo identificar que no se tienen valores atípicos, nulos o con formato errado. En este sentido no se aplicaría preprocesamiento ni se expondrá lo encontrado en la sección 2.4.

En el conjunto de datos que contiene el registro de los tanques, no se emplearon técnicas de visualización especializadas porque la sola validación en forma de tabla imprimiendo las cinco (5) primeras líneas permitieron conocer el formato de los campos de fecha de las variables FillingStart, FillingEnd, ConnStart y ConnEnd. Lo anterior se puede apreciar en la Figura 2.

	UT	Brand	FillingStart	FillingEnd	ConnStart	ConnEnd
0	13	Amstel Utra	2022-04-12 11:43:00	2022-04-12 22:05:00	2022-04-13 13:00:00	2022-04-17 11:00:00
1	13	Carta Blanca	2022-01-17 15:49:00	2022-01-18 00:59:00	2022-01-18 19:00:00	2022-01-22 09:00:00
2	13	Carta Blanca	2022-03-21 06:21:00	2022-03-21 15:47:00	2022-03-22 17:00:00	2022-03-26 02:34:00
3	13	Carta Blanca	2022-04-22 04:32:00	NaN	2022-04-23 05:12:00	2022-04-28 06:00:00
4	13	Carta Blanca	2022-05-27 07:50:00	2022-05-27 16:05:00	2022-05-28 07:00:00	2022-06-01 05:36:00

Figura 2 Revisión del contenido del conjunto de datos CO2_Connection_202409271420.csv. Fuente: Elaboración propia.

Aparte de las variables mencionadas, se tiene la de UT relacionada con la identificación del tanque, la numeración inicia en 13 y finaliza en 50.

La Figura 3 muestra la cantidad de registro que se obtuvo de cada tanque, por ejemplo, el tanque #50 tuvo 25 registros mientras que el tanque #35 mostró un valor cercano a 45 registros.

Hay un tercer conjunto de datos que tiene que ver con el flujo de CO2, este no pudo ser obtenido antes de la entrega. Sin embargo, conociendo la estructura general de los datos, realizamos un conjunto de datos de prueba con datos aleatorios, por lo que no tendremos datos nulos o erróneos, aunque por lo que entendimos, estos tampoco deberían existir en los conjunto de datos originales, debido a que los registros manuales son completados y revisados rigurosamente, y los sensores IoT poseen los estándares más altos de calidad, por lo que la comunicación nunca es interrumpida.

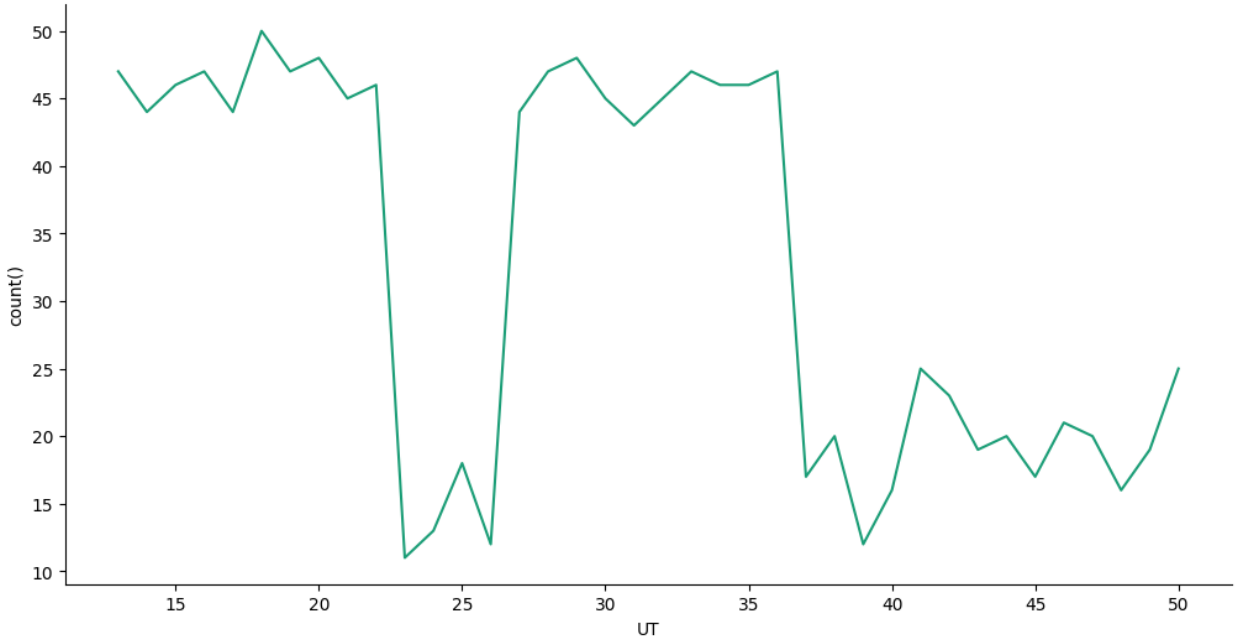

Figura 3 Relación entre la variable UT y el conteo de datos por periodo. Fuente: Elaboración propia.

2.3 Análisis bi/multivariante

Hasta este punto, aunque se conoce la relación entre las variables que participan en la generación del CO2 en las plantas de producción de cerveza, no se tiene unificación de estas, especialmente porque el flujo medido por la solución de IoT aún no se tiene. Sin embargo, nos aventuramos en revisar la correlación entre UT y Volumen sabiendo que en varias unidades de tanque no se tiene el registro completo. En la siguiente entrega, cuando tengamos todos los datos consolidados, exploraremos esa relación entre la cantidad de cerveza en los tanques y la producción de CO2. La intuición sugiere que debería haber una correlación directa, ya que la carbonatación de la cerveza depende del CO2. Sin embargo, será interesante ver cómo se refleja esto en los números.

En la Figura 4 se tiene la matriz de correlación entre UT y Volumen, el coeficiente de 0.13 resulta bajo lo cual indica que no se tiene correlación fuerte entre la identificación de los tanques y el volumen, aquí se debe considerar que en el conjunto de datos según lo visto en la Figura 3, no siempre se tienen registros completos en el periodo de tiempo objetivo.

El resultado no es concluyente, la correlación no siempre es directamente proporcional a la causalidad por lo que es requerido explorar el conjunto de datos del flujo de CO2 y cómo este afecta la generación por tipo de cerveza y los tiempos de llenado que serían calculados por la resta del FillingEnd y FillingStart. Aquí se tendría UT, FillingDuration in minutes, CO2flow y la conversión de Brand en variable numérica.

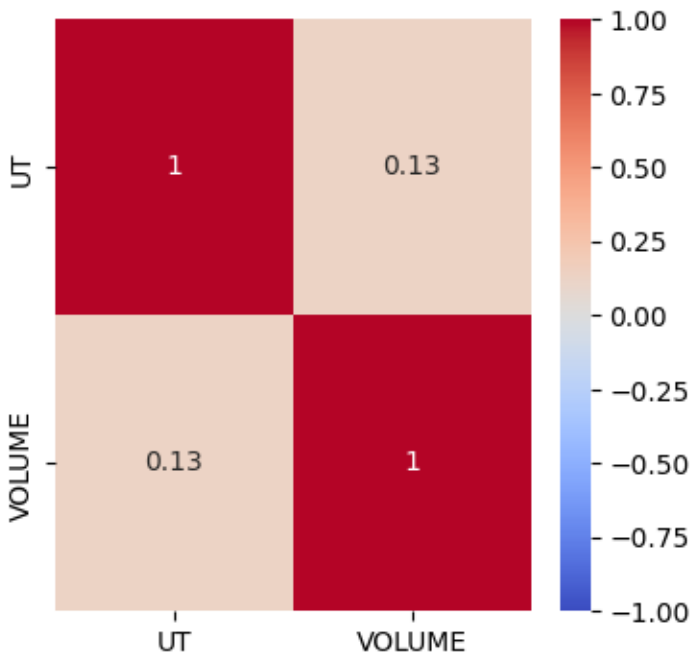

Figura 4 Matriz de correlación entre UT y Volumen para el conjunto de datos CO2_UT_VOL_202409271421.csv. Fuente: Elaboración propia.

2.4 Preprocesamiento.

De los tres (3) conjuntos de datos, sólo el correspondiente al registro por tipo de cerveza *"CO2_Connection_202409271420.csv"* mostró valores nulos, en la sección 2.2 del notebook en github se aprecia lo descrito en las Figuras 5 y 6.

```
In [14]: registers_df.isna().sum()
```

Out[14]:

	0
UT	0
Brand	0
FillingStart	0
FillingEnd	15
ConnStart	26
ConnEnd	60

Figura 5 Revisión de los valores erróneos en el conjunto de datos CO2_Connection_202409271420.csv. Fuente: Elaboración propia.

```
In [15]: registers_df.isnull().sum()
```

Out[15]:

	0
UT	0
Brand	0
FillingStart	0
FillingEnd	15
ConnStart	26
ConnEnd	60

dtype: int64

Figura 6 Revisión de los valores nulos en el conjunto de datos CO2_Connection_202409271420.csv. Fuente: Elaboración propia.

Se decidió hacer la eliminación de los datos nulos y erróneos bajo la premisa de que seguramente son los mismos, únicamente que duplicados. Nuestra decisión de cómo manejarlos se basó simplemente en quitar los registros con este tipo de datos y observar cuántos registros totales realmente se eliminan (ya que varios pueden encontrarse en los mismos renglones). Luego de aplicar la técnica evidenciamos que perdimos 84 registros (alrededor de un 6% del conjunto de datos original).

Conclusiones

Luego de realizar el análisis de datos basado en la validación de la estructura y posteriormente el análisis univariante, bivariante y hacer preprocesamiento, se pudo determinar cuáles variables se deben considerar en la siguiente fase del proyecto. En la exploración de los conjuntos de datos que tienen que ver con la capacidad de los tanques de almacenamiento, el de registro por tipo de cerveza y el flujo de CO2 simulado se notó que el primero no contiene datos nulos ni erróneos al igual que el último y que, además, se trata de datos más limpios en el sentido de que son pocos con relación a lo que suponemos será el de flujo de CO2. El conjunto de datos que contiene la relación entre UT, tipo de cerveza y tiempo de llenado permitió validar que la solución de IoT puede estar sujeta a mejoras en la conexión ya que en términos de disponibilidad estaría en 98% (26 de los 1246 registros no mostró información del inicio de la conexión).

La siguiente fase supone la consolidación de las variables UT, Tiempo de llenado por tipo de cerveza en distintos periodos complementado por el flujo de CO2 medido por la solución de IoT y ya con esto podemos avanzar en el planteamiento del modelo de ML para lograr la predicción de la generación que es finalmente el objetivo del proyecto.