**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

[1. Descripción General de los Datos 2](#_Toc178285206)

[2. Análisis de Datos 3](#_Toc178285207)

[2.1 Elegir las características más relevantes para reducir la dimensionalidad y aumentar la capacidad de generalización del modelo. 3](#_Toc178285208)

[2.2 Abordar y corregir los problemas identificados en los datos. 4](#_Toc178285209)

[3. Convenios 4](#_Toc178285210)

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.

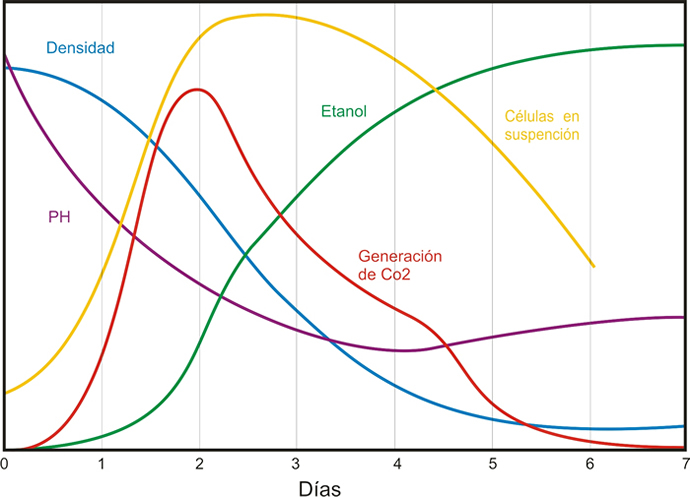


Imagen 1. (https://www.cervezaartesanalsm.com.ar/fermentacion1/). Diversas características de la cerveza a través del tiempo. Cabe recalcar, como ya mencionamos, que los tiempos son dependientes en gran medida a la marca, por lo que muy probablemente, los mostrados en la gráfica, no son iguales a los 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.

1. Análisis de Datos

### 2.1 Elegir las características más relevantes para reducir la dimensionalidad y aumentar la capacidad de generalización del modelo.

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

1. Un diccionario (archivo .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 del mismo. Sin embargo, debido a la dificultad del proyecto, y división de los datasets 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 Abordar y corregir los problemas identificados en los datos.

Actualmente enfrentamos un reto. Precisamente por la falta de organización de los datos, y la manera en como alguno de estos son obtenidos, de manera manual, mediante registros que se van llenando personalmente, por lo que toman tiempo. Debido a esto, no hemos podido obtener los datos reales, por lo que conociendo la estructura general de los datos, realizamos datasets 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 datasets 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.