**Referencia: Problema de gestión global de datos**

**Fecha: 08/09/2021**

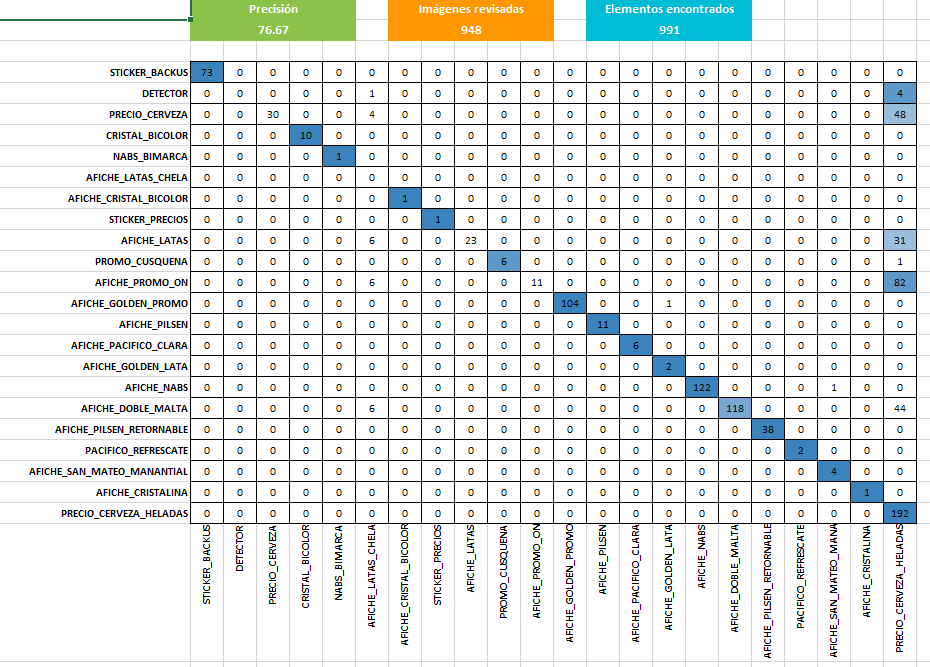
**Responsable: Carlos Jorge Pallares Urieles**

Se solicitó una revisión de datos el día 03/09/2021 para un entrenamiento de Perú/Tradicional, los datos relacionados son los siguientes:

**Link entrenamiento**:

<https://saturno.aituring.co/training/details/dataset/4a31b96b-1c5f-4cf4-803e-04c1e278fbf8/>

**Matriz de confusión**:

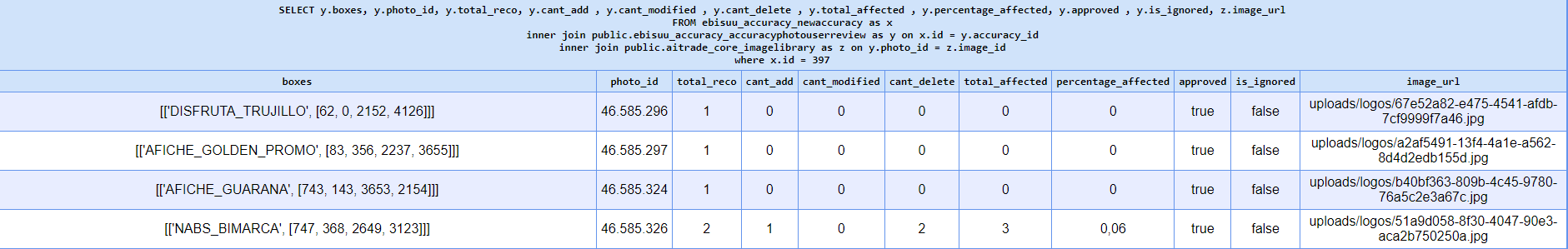


Se empezó la exploración de datos desde el día **06/09/2021** hasta **07/09/2021** siendo totalmente **fallida** ya que no tenemos acceso a los datos usados para el accuracy de forma explícita, dicha información debe estar incluida luego de cada entrenamiento al igual que la siguientes:

* **Etiqueta especifica de la imagen**
* **Tarea especifica de la imagen. No mezclar tareas. Si el cliente envía imágenes de una tarea diferente entonces asignar a la tarea que le corresponde**
* **Listado de imágenes usadas para validar el modelo**
* **Listado de imágenes usadas para entrenar el modelo**
* **Número de datos por etiqueta**
* **Numero de datos por tarea**
* **Listado completo de las imágenes de referencia en la mejor resolución posible**

A la falta de toda esa información se procedió a una consulta manual en SATURNO para extraer imágenes de las clases PRECIO\_CERVEZA Y PRECIO\_CERVEZA\_HELADAS, cuando realizamos esto nos dimos cuenta de que imágenes de PRECIO\_CERVEZA aparecen en la búsqueda de PRECIO\_CERVEZA\_HELADA y viceversa por el simple de hecho de **compartir la cadena de caracteres “PRECIO\_CERVEZA”** por lo cual debíamos hacer una categorización manual. No se realizó por temas de tiempo y pertinencia del área.

Se optó entonces realizar una petición SQL en la base de datos de AITrade, los resultados son los siguientes:



Siendo cada fila una foto usada para el accuracy y la primera columna una lista con la etiqueta y su posición en la foto. Pero las imágenes de accuracy son obtenidas por fuera del dataset original usado para entrenamiento, por lo cual es posible que al hacer una búsqueda por etiqueta (**nuestro único indicador**) para el cálculo de accuracy ocurra el problema mencionado en el párrafo anterior.

**NOTA: No se opto por los datos de entrenamiento ya que se haría una revisión del accuracy, y los datos del accuracy no están en el dataset con que se entrena**

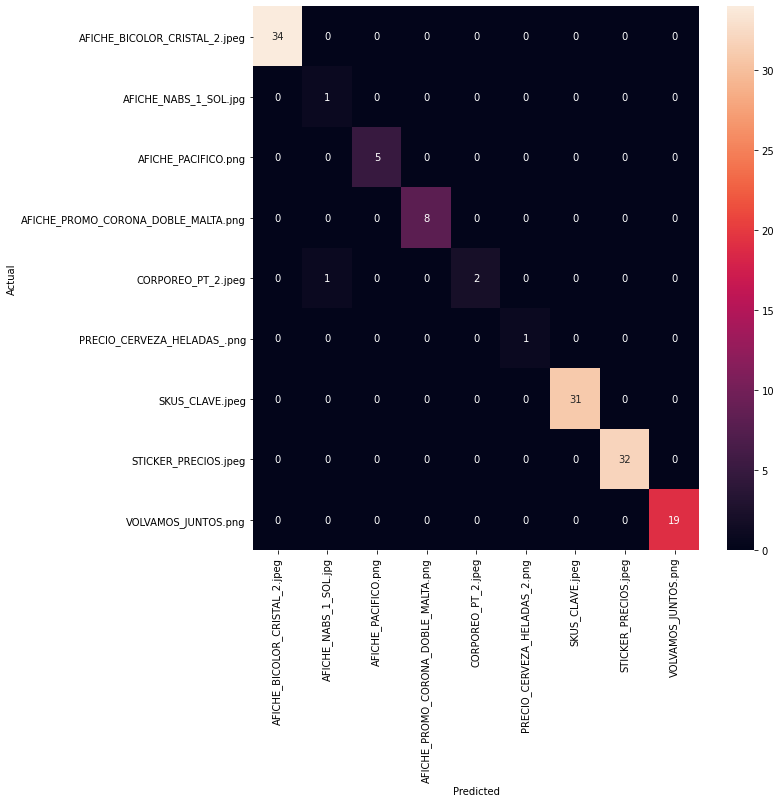
Se debe tomar acción de forma inmediata para la organización de la información y estructura un gobierno de datos para evitar contratiempos en operaciones y el grupo de inteligencia artificial, tratar los datos de nuestra empresa como un activo económico.

**Recomendación para este caso**:

1. Organizar la información de forma explícita como se especificó en los 7 incisos mencionados en este documento
2. Permitir el acceso intuitivo a los datos y a los 7 incisos mencionados en este documento
3. Implementar no training a los casos de POP, independiente del país. Se adjunta matriz de confusión de la última evaluación del modelo sobre los datos en, los links de imágenes de referencia y objetivo usados para esto:

**Imágenes referencia**:   
<https://drive.google.com/drive/folders/1JqAtHZ3Xoa1G8XLlKVsxF2otMwWDOsU7?usp=sharing>  
  
**Imágenes objetivo**: https://drive.google.com/drive/folders/17h8-r74qnu9NrBSsfZTrTN\_kVI1R6i4H?usp=sharing

**Matriz de confusión**:



El dato (CORPOREO\_PT\_2.jpeg , AFICHE\_NABS\_1\_SOL.jpg) = 1 es correcto ya que es una imagen colada.