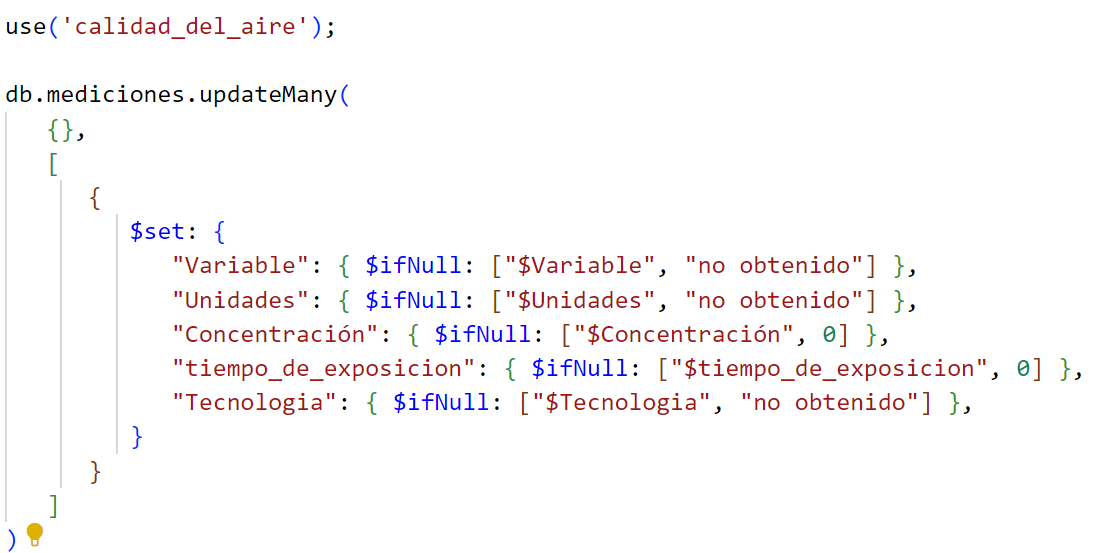
**TRATAMIENTO DE LA BASE DE DATOS CALIDAD DEL AIRE**

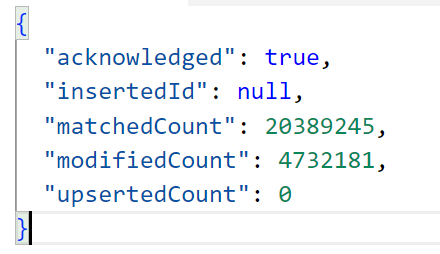
Para el tratamiento lo que hicimos fue la creación de 4 colecciones

|  |  |  |
| --- | --- | --- |
| Nombre de la colección | Descripción | Cantidad de Registros |
| autoridades\_ambientales | Se va registrar todos los atributos de las autoridades ambientales y las relaciones que presenten. | 28 registros |
| estaciones | Se va registrar todos los atributos de las estaciones y las relaciones que presenten. | 380 registros |
| mediciones | Se va registrar todos los atributos de la medición de calidad del aire y las relaciones que presenten con otras colecciones. | 20.4 millones de registros |
| departamentos | Se va registrar departamentos y municipios y las relaciones que presenten con otras colecciones. | 1124 registros |

Después se manipularon los datos que podían estar nulos y se sustituyeron con el siguiente aggregate:



Y se obtuvo el siguiente resultado



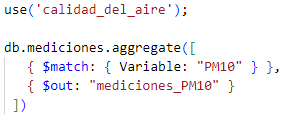
**Explicación:** Se contemplaron los archivos nulos que podían estar en la colección de mediciones y los tratamos de la siguiente manera.

|  |  |  |
| --- | --- | --- |
| **Nombre del Dato** | **Tipo de dato** | **Descripción** |
| Variable | String | Se cambiaron los datos nulos y se registro un valor de “no obtenido” |
| Unidades | String | Se cambiaron los datos nulos y se registro un valor de “no obtenido” |
| Concentración | Numérico | Se cambiaron los datos nulos y se registró un valor 0 |
| Tiempo de exposición | Numérico | Se cambiaron los datos nulos y se registró un valor de 0 |
| Tecnología | String | Se cambiaron los datos nulos y se registro un valor de “no obtenido” |

En la segunda imagen tenemos los archivos que cumplían con un dato nulo de los 5 que se contemplaron 4.732.181 de registros fueron los que cumplieron esta condición y sufrieron los cambios explicados en la tabla.

**Después se realizo un filtro donde se manejaron solo los datos que la variable era igual a la constante “PM10”**

1. Realizar un conteo en los 20.4 millones de registros para saber cuantos existen para la condición
2. Realizar una replicación en otra colección con los datos que cumplan la condición



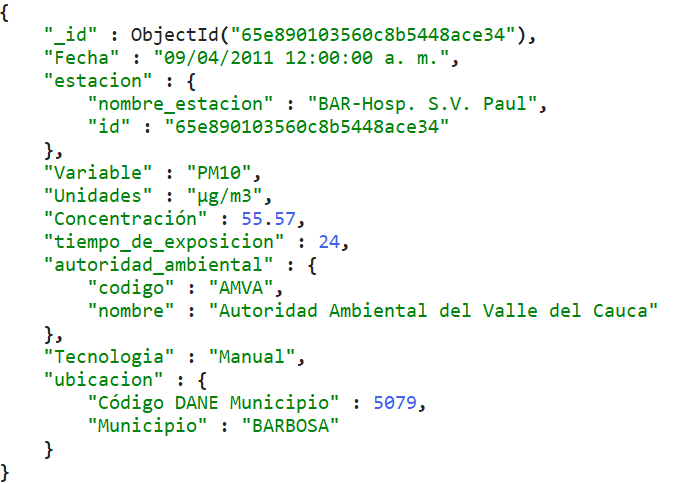
1. Salida esperada con los datos ya filtrados con 2,369.180 registros totales.



**TRATAMIENTO DEL ARCHIVO CSV**

Se selecciono una muestra del 10% que cuenta aproximadamente con 2 millones de registros de los datos de la colección mediciones estos escogidos de manera aleatoria.

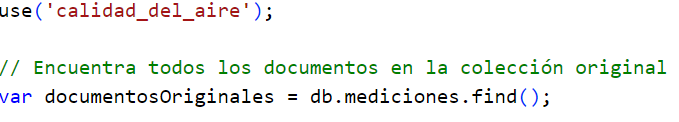
**Proceso de adecuación del registro**



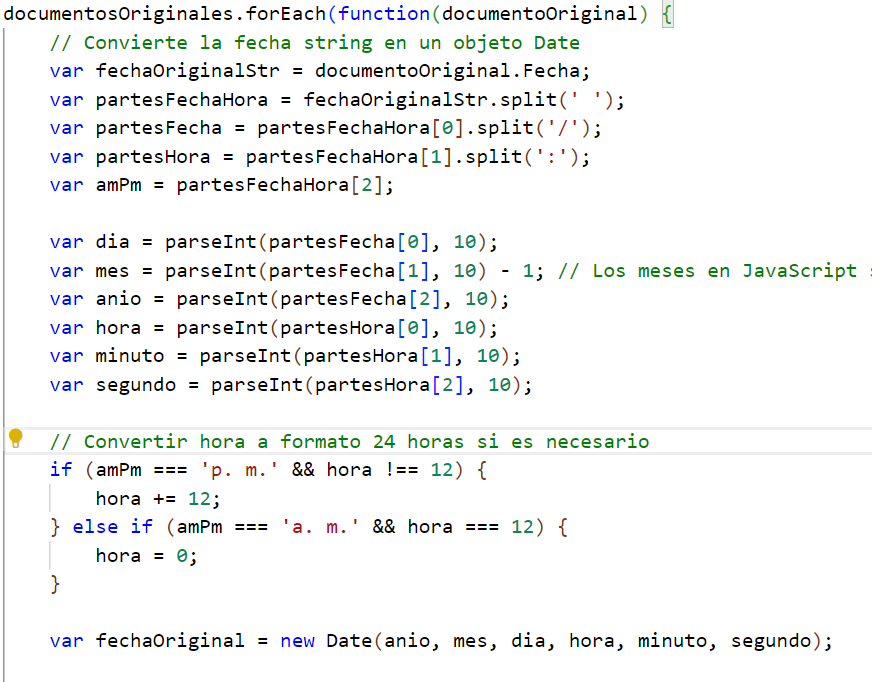
JSON DE ENTRADA. Este es la estructura que se manejo en mediciones y se va a adecuar para el modelo ML

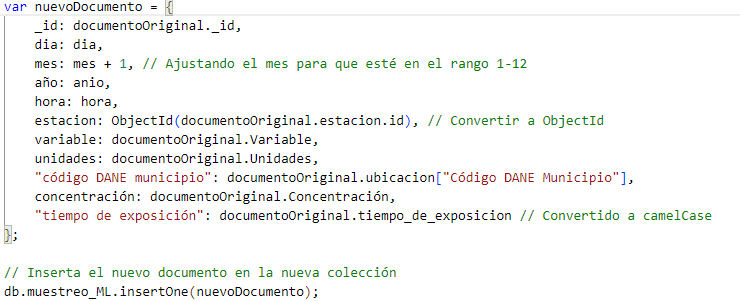
**Creación del script**

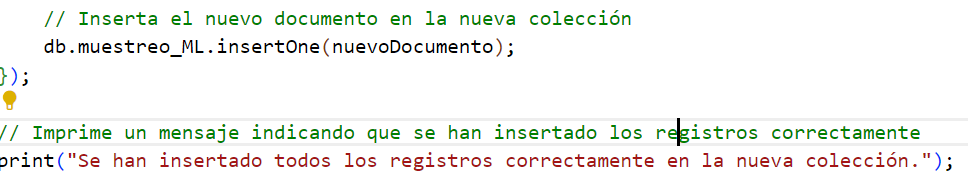
1. Se realiza la conexión con la base de datos y el llamado de la colección a tratar



1. Se recorre la colección para que cada iteración se haga la conversión de la fecha en campos individuales y se adecua la hora en formato de 24hrs para su posible entrenamiento.



1. Creación del JSON final lo que hacemos es llamar los datos que se necesitan en para el entrenamiento de red neuronal.
2. Inserción de los datos de manera aleatoria en una colección de prueba



1. Salida esperada y Conversión de la colección prueba a el archivo csv



Esta es la estructura del archivo final que va ser enviado al modelo ML ese para su posible entrenamiento