PREDICCION DE MASA MOLAR USANDO REDES CONVOLUCIONALES

Jesus Emmanuel Ramos Davila
Facultad de Ciencias Fisico Matemáticas

1. Introducción

Una fórmula química es una expresión gráfica de los elementos que componen un compuesto químico cualquiera. Las fórmulas expresan los números y las proporciones de sus átomos respectivos y, en muchos casos, también el tipo de enlaces químicos que los unen. A cada molécula y/o compuesto conocido le corresponde una fórmula química, así como un nombre a partir de ella de acuerdo a las reglas de la nomenclatura química.

Otro de los componentes por el cual podemos conocer su formula química es por si estructura, La **estructura química** se define como el arreglo espacial de átomos dentro de una molécula. La estructura química determina la geometría molecular de una molécula especifica. Esto fue descubierto por el químico ruso Alexander Butlerov el cual reconoció que las moléculas no son arreglos aleatorios de átomos sino que estan formados dentro de un patrón concreto [Figura 1].
El peso molar (**molar weight**) de una sustancia es la masa en gramos de un mol de la sustancia.

Este estudio esta dirigido a el procesamiento de imágenes las cuales contienen
estructuras químicas para cada formula se tiene su respectiva masa molar.

2. Objetivo

El objetivo se centrara en realizar una red convolucional para una regresión en cual trataremos de estimar la masa molar dada su estructura química, esto servirá como toma inicial de decisión para revisión de etiquetas para su pictograma de seguridad. ya que en algunos casos no se reporta sus pictogramas de seguridad relacionados lo cual podría ser peligroso al manejo y distribución de la sustancia química.

3. Metodologia

Preprocesamiento: El preprocesamiento es una de las tareas las cuales llevo una serie de subprocesos los cuales se enumeraran:

1. **Parseo y extracción** de columnas dentro de una serie de archivos **XML** esto nos dara las columnas iniciales en un formato CSV para su análisis.
2. **Web Scrapping fase 1 :** Este se realizo tomando en cuenta el paso 1, el webscrapping se realizo para la extracción de especificaciones de la formula tales como su formula , masa molar.
3. **Web Scrapping fase 2** : Este se realizo con el objetivo de obtener la imagen de su estructura química correspondiente e.g **Figura 1 y Figura 2**.
4. **Almacenamiento en base de datos:** relacionar para su posterior union con su respectiva formula y masa molar.

Imputación: Reemplazo de algunos valores nulos por valores ofuscados a fin de no mostrar datos sensitivos

Análisis: Búsqueda de correlaciones , histogramas , boxplots y pruebas no-parametricas a fin encontrar hallazgos interesantes

Modelo: El modelo propuesto es una Red Neuronal Convolucional CNN **Figura 4** usando las imagenes de la fase 2 de Web Scrapping, nuestra variable de respuesta es en este caso la masa molar.

Validación: La validación actual de este modelo ya que su naturaleza es un modelo de regression se optara por un error de promedios al cuadrado o MSE.

5. Importancia del Proyecto

La importancia del proyecto reside en la seguridad y una toma de decisión previa ante la falta en algunos casos de sus correspondientes pictogramas de seguridad, ya que este modelo tiene como fin determinar si determinada formula tiene un alto nivel de masa molar lo cual podría indicar que esta tiene un alta probabilidad de ser peligrosa tanto en su manejo como en su exposición ante esta formula.

4. Resultados

Los resultados para este modelo de Red Convolucional fueron de moderados a buenos **Grafico 1**, una de las principales causas las cuales afecto el modelo de datos desbalanceados ya que solo algunas sustancias reportaban masas molares muy altas, este tipo de masa molar no se trato de eliminar ya que no se considera un dato atípico o un error ya que tanto su masa molar como otras propiedades fueron reportadas correctamente.
En la parte del modelo se observo que a partir de las primeras 13 épocas la medida de loss tuvo una ajuste considerable pero en siguiente épocas se mantuvo sin ninguna mejora **Figura 3**.

Mejoras a considerar:

1. Algunos de los puntos a considerar como mejora para este modelo es el cambio de estrategia para su predicción al pasar de un modelo de regresión y variable de respuesta masa molar a una predicción multiclase la cual se considera en esta las 9 posibles pictogramas de seguridad que podría tener una formula química.
2. Realizar un webscrapping mas exhaustivo a fin de obtener más formulas químicas y su estructura molecular.
3. Realizar procesamiento con las imagenes que se obtuvo otro tipo de extension SVG, JPEG, JPG , y de requerir su debida transformacion y escala.
4. En caso de requerir una división sistemática a fin de obtener 2 grupos los cuales se podrian dividir por algun tipo de mediana, o moda la cual nos permita seccionar (Formulas químicas divididas en grupos mayor y menos masa molar)

Grafico 1: Valor actual vs Predecido

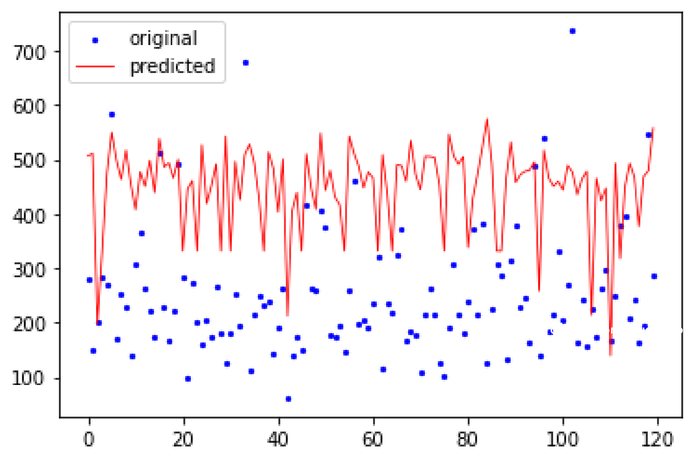

Figura 3: Epocas con mayor desempeño

```
Epoch 9/100
7/7 [==============================] - 13s 2s/step - loss: 215389.3906
Epoch 10/100
7/7 [==============================] - 13s 2s/step - loss: 215389.3750
Epoch 11/100
7/7 [==============================] - 14s 2s/step - loss: 215389.2969
Epoch 12/100
7/7 [==============================] - 14s 2s/step - loss: 215389.2656
Epoch 13/100
7/7 [==============================] - 13s 2s/step - loss: 215389.2188
```

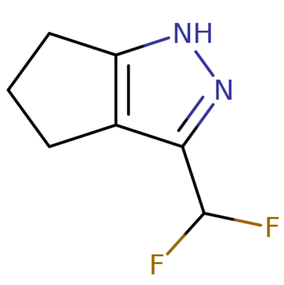

Figura 1

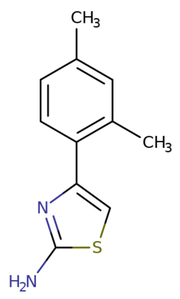

Figura 2

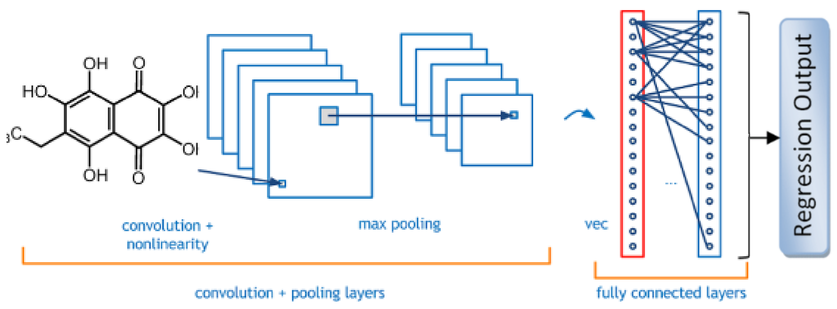

Figura 4

6. Conclusión

Este modelo tuvo una predicción moderada a baja , pero a pesar de su desempeño el modelo tiene ciertas mejoras que se podrían adquirir al ingresar mas datos y al posiblemente realizar al menos 2 grupos de formulas las que tienen masa molar reportada como alta y otras con masa molar reportada como promedio o baja. Al realizar esto nuestro modelo no presentaría ningún tipo de sesgo por formulas con masa molar alta y podríamos cambiar nuestro enfoque al propuesto en la sección **Mejoras a considerar.**
Al ser una primera fase de modelado y un rendimiento moderado a bajo nos da señales de que nuestros imágenes pueden ser consideradas para este tipo de modelos de convolucion **CNN**.

References
Figura 4: https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529
"Fórmula química". Autor: Dianelys Ondarse Álvarez. De: Argentina. Para: Concepto.de. Disponible en: https://concepto.de/formula-quimica/. Última edición: 15 de julio de 2021. Consultado: 21 de octubre de 2023
Nishinari K, Fang Y. Molar mass effect in food and health. Food Hydrocoll. 2021 Mar;112:106110. doi: 10.1016/j.foodhyd.2020.106110. Epub 2020 Sep 3. PMID: 32895590; PMCID: PMC7467918.
Fuente: https://concepto.de/formula-quimica/#ixzz8GqTdfdLl