# Protocolo seguido para generar el modelo de VDss

## Endpoint

Volume of Distribution at steady-state (VDss)

## Origen de los datos

Los datos vienen del Supplementary Information de los artículos:

* “***In Silico* Models of Human PK Parameters. Prediction of Volume of Distribution Using an Extensive Data Set and a Reduced Number of Parameters**” de Berellini (2020). DOI: <https://doi.org/10.1016/j.xphs.2020.08.023>
* **“An Accurate *In Vitro* Prediction of Human VDss Based on the Øie–Tozer Equation and Primary Physicochemical Descriptors. 3. Analysis and Assessment of Predictivity on a Large Dataset”** <https://doi.org/10.1124/dmd.119.088914>
* **“A Benchmarking Dataset with 2440 Organic Molecules for Volume Distribution at Steady State”** (<https://github.com/da-wen-er/VDss/tree/main/VDss_dataset>)

## Tratamiento de los datos

Del Excel original nos quedamos con las columnas “SMILES” y “human VDss (L/kg)”.

Después, y antes de pasar por Hygieia, se usó el script de Ágata (Elbow\_Undersampling.py) para quitar los outliers según peso molecular.

## Transformación de la “y”

Se ha llevado a cabo la transformación log10 (manualmente con numpy, no usando el 1/log que tiene Neo) del valor respuesta para realizar el modelo.

## Train/test ratio

Se ha realizado una partición del 70% para el train y del 30% para el test. Durante el proceso de partición se han mantenido los compuestos señalados como *quite dissimilar*.

## Scaler

Se ha usado el *Standard Scaler*