# **Variable a modelar**

**ClinvarHC**

La variable target se extrajo de la base de datos clinvar. Sólo se consideran variantes con múltiples submisions, sin conflicto de interpretación y pertencecientes a las categorías

0: 1(benigna) 2(probablemente benignas)

1: 5(Patogénica)

# notar que todas las variantes en la clase positiva tienen al menos una anotación como patogénica. Aquellas sólo con categoría 4(probablemente patogénica) fueron descartadas.

Este tipo de estrategia ayuda a limpiar un poco la base de datos de clinvar que puede en caso contrario, estar bastante contaminada.

# Features del dataset

## **Conservación**

Gerp\_gerp\_nr

Gerp\_gerp\_rs

Phastcons\_phastcons100\_vert

Phastcons\_phastcons30\_mamm

phylop\_phylop100\_vert

Phylop\_phylop100\_vert\_r

Phylop\_phylop30\_mamm

## **Expresión diferencial**

gtex\_counts

## **Bases de datos poblacionales**

*#### indels / CNVs:*

exac\_gene\_exac\_del\_score

exac\_gene\_exac\_dup\_score

exac\_gene\_exac\_cnv\_score

exac\_gene\_exac\_cnv\_flag

*##### Loss of Function:*

gnomad\_gene\_oe\_lof\_min

gnomad\_gene\_lof\_z\_max

gnomad\_gene\_pLI\_max

Loftool\_loftool\_score #Gene intolerance score based on lof (exac)

*### missense variants*

gnomad\_gene\_oe\_mis\_min

gnomad\_gene\_mis\_z\_max

*### Synonymous variants*

gnomad\_gene\_oe\_syn\_min

gnomad\_gene\_syn\_z\_max

gnomad\_gene\_pRec\_max #prob. Recessive condition

gnomad\_gene\_pNull\_max #prob. Ser tolerante en heterocigosis y

hoozigosis

## **Estructural**

Interpro\_domain\_simple

## **Essential Essential or non-essential genes (dbNSFP)**

ess\_gene\_ess\_gene

ess\_gene\_ess\_gene\_crispr

ess\_gene\_ess\_gene\_crispr2

ess\_gene\_ess\_gene\_gene\_trap

ess\_gene\_indispensability\_scor

## **Redes PPI**

biogrid\_counts

intact\_counts

## **Otros**

Ghis\_ghis # Genome Wide Haploinsuf score (dbNSFP)

Prec\_prec # prob. Recessive condition

Repeat\_flag # UCSC hg38