**DATASET – JUSTIFICACIÓN DE LAS VARIABLES**

Nuestro datagenerator cuenta con las siguientes variables: edad, patología previa, tensión arterial, kilómetros, % de grasa, nivel de colesterol, fumador, si bebe alcohol y minusvalía.

Para justificar las variables, podemos hacer uso de internet con estudios ya realizados que muestren la correlación de dichas variables con la edad, así como la esperanza de vida. Sin embargo, tiene más valor que, habiendo aprendido estadística con Python, intentemos justificarnos mediante esa herramienta.

Hemos realizado una búsqueda exhaustiva en kangle para dar con algún dataset que pudiera ser una fuente de información en relación con nuestras variables en la medida de lo posible. Era evidente que no era posible encontrar algo que encajase con todas nuestras variables.

De entre todos los dataset que hemos encontrado, hemos elegido uno específicamente dado que a pesar de no tener un source claro, es el que más confianza nos ha generado por la identidad del creador. Dicho creador no solo tiene públicos sus datos en la plataforma, sino que también proporciona su LinkedIn. Dicha persona es Svetlana Ulianova y estudia Data Science Student at Ryerson University (<https://www.linkedin.com/in/svetlanaulianova/>). A parte, el dataset cuenta con un tamaño muestral de 70.000 personas.

El primordial objetivo con el dataset era darle uso a la variable más determinante para el cálculo de un seguro de vida, como es la edad. Antes que nada, ha sido necesario realizar una conversión de la columna edad porque venía dada en días y preferimos trabajar con la edad en años, por eso hemos creado una columna adicional con el nombre ‘age\_years’ (la cual es nuestra variable predicha). A fin de justificar que a parte de tener en cuenta la esperanza de vida de cada individuo (o usuario mejor dicho en este caso), también hay que considerar que a medida que envejecemos tenemos peores hábitos y estilos de vida. Para ellos utilizamos como predictoras las variables colesterol, actividad física y reportes cardíacos (que simularía la variable ‘patología previa’ en nuestro generator). Desafortunadamente, ninguna de las variables estudiadas (a parte de la edad), son cuantitativas, por lo que no hemos podido realizar ningún scatterplot.

Dentro del dataset contamos con otras variables interesantes como la de fumar y beber, dichas variables hemos preferido relacionarlas con el sexo, dado que no consideramos que según envejecemos bebemos y fumamos más. Sin embargo, que las mujeres fumen y beban menos si y demostramos que efectivamente así es. El objetivo de esta segunda parte de estadística con Python era demostrar que las mujeres tienen mejores hábitos que los hombres y que junto a que también tienen una mejor esperanza de vida (85,06 años frente a 79,59 en España, véase https://datosmacro.expansion.com/demografia/esperanza-vida/espana), se justifica que la variable sexo en nuestro generator, también tenga cierta presencia.

También hay que comentar que ha sido necesario trabajar renombrando los valores de las variables nominales con las que hemos trabajado. Por ejemplo: el valor 1 en sexo es mujer y el valor 2 es hombre, de modo que hemos creado ‘gender\_cat’ para mostrar ‘Female’ y ‘Male’ y así no generar dudas. Dicho procedimiento ha sido utilizado también para las variables (a modo de variable\_cat): colesterol (donde el valor 1 figura como ‘Saludable’, el 2 como ‘En riesgo’ y 3 como ‘No saludable’), fumador (donde 0 pasa a ser ‘No fumador y 1 es ‘Fumador’), bebedor (donde 0 es ‘No bebe’ y 1 es ‘Bebe’ , patología cardiovascular (‘Sin reporte cardiovascular’ para el valor 0 y ‘Con reporte cardiovascular’) y actividad física (‘No hace ejercicio’ para el valor 0 y ‘Hace ejercicio’ para valor 1).

Por último añadir que las variables restantes del dataset como id, altura, peso, glucosa y la presión arterial (sistólica y diastólica), no han sido interesantes para hacer el estudio. Esta última variable mencionada hubiera sido de gran utilidad de no ser porque al ser un dataset estadounidense utiliza otras mediciones para el cálculo de ambos tipos de presiones arteriales. Al no poseer ninguno del grupo los conocimientos necesarios para una posible investigación y su posterior conversión, hemos preferido no meternos en un terreno que desconocemos y no trabajar con la variable. No sería la primera vez que de un dataset tenemos que dejar de lado ciertas variables por emplear mediciones distintas en ámbitos que no manejamos como la medicina.

El resto de datasets que han sido valorables y revisados han sido los siguientes:

<https://www.kaggle.com/raminhashimzade/cardio-disease>

<https://www.kaggle.com/yudhaislamisulistya/cardio-train-new>

<https://www.kaggle.com/kirillformado/cardio>

<https://www.kaggle.com/soufianezemrani/cardio>