* **Descripción de las variables y valores estadísticos (minimo, máximo, media, desviación, mediana, etc.). Estudia qué valores estadísticos son los convenientes según el tipo de variable y procede en consecuencia**.

Tenemos una base de datos de resultados de un análisis sanguíneo a diferentes pacientes.

Por un lado tenemos datos sobre los pacientes: la edad y el género, por otro lado la medida de ciertos componentes en su organismo:

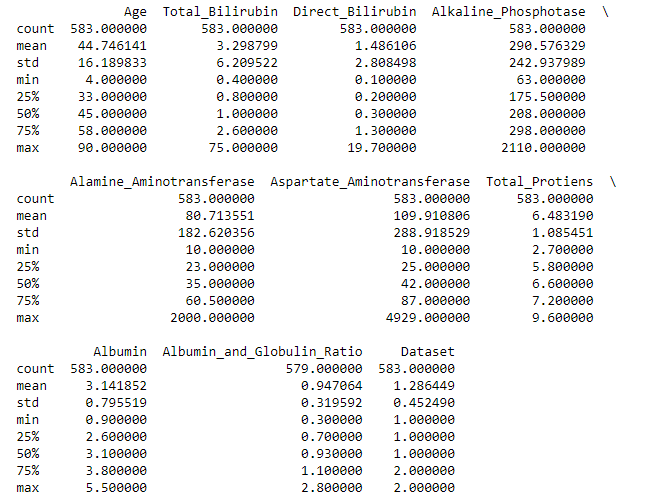
Primero el total de bilirrubina en el organismo y la bilirrubina directa. Y consultando con expertos nos informamos que **total de bilirrubina=bilirrubina directa + bilirrubina indirecta**. Y que la bilirrubina indirecta es aquella asociada a la albúmina (proteína portadora).

Seguimos con fosfatasa alcalina es una enzima que se encuentra en el torrente sanguíneo, los niveles normales varían según edad y sexo, los niveles en niños son más altos y disminuye con la edad.

Después aspartato aminotransferasa y alanina aminotransferasa que son encimas, el total de proteínas en el organismo. Y para finalizar con las variables la cantidad de albúmina, que como ya hemos mencionado es una proteína, y el ratio de albúmina y globulina.

También tenemos una clasificación en dos posibles categorías que desconocemos su significado.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Age | Total\_Bilirubin | Direct\_Bilirubin | Alkaline\_Phosphotase | Alamine\_Aminotransferase | Aspartate\_Aminotransferase | Total\_Protiens | Albumin | Albumin\_and\_Globulin\_Ratio | Dataset |
| count | 583 | 583 | 583 | 583 | 583 | 583 | 583 | 583 | 579 | 583 |
| mean | 44,746141 | 3,298799 | 1,486106 | 290,576329 | 80,713551 | 109,910806 | 6,48319 | 3,141852 | 0,947064 | 1,286449 |
| std | 16,189833 | 6,209522 | 2,808498 | 242,937989 | 182,620356 | 288,918529 | 1,085451 | 0,795519 | 0,319592 | 0,45249 |
| min | 4 | 0,4 | 0,1 | 63 | 10 | 10 | 2,7 | 0,9 | 0,3 | 1 |
| 25% | 33 | 0,8 | 0,2 | 175,5 | 23 | 25 | 5,8 | 2,6 | 0,7 | 1 |
| 50% | 45 | 1 | 0,3 | 208 | 35 | 42 | 6,6 | 3,1 | 0,93 | 1 |
| 75% | 58 | 2,6 | 1,3 | 298 | 60,5 | 87 | 7,2 | 3,8 | 1,1 | 2 |
| max | 90 | 75 | 19,7 | 2110 | 2000 | 4929 | 9,6 | 5,5 | 2,8 | 2 |



* Describe y realiza modificaciones en la base datos si lo consideras necesario. Por ejemplo, qué harías con valores nominales, si los hubiera.

El valor género que es un valor nominal, lo hemos cambiados por otros valores female=1 male=0.

* Estudia si es necesario normalizar los datos y cómo lo harías. Procede a modificar la base de datos (normalizar) si lo consideras necesario.
* Detección de valores extremos (outliers) y descripción de qué harías en cada caso.
* Detección de valores perdidos (missing values) y descripción de cómo actuarías para solventar el problema.
* Buscar correlaciones entre:
  + las variables predictoras, lo que permitirá ver si hay variables redundantes.
  + variables predictoras y la clase (target).
* Detecta, si hubiera, falsos predictores.
* Estudia si fuera conveniente segmentar alguna de las variables.
* Estudia si fuera conveniente crear nuevas variables sintéticas basada en las variables originales.