**Datos clínicos**

**///////////////////////////////**

Pacientes: 41

Atributos descriptores: 45

Clases: N,I,II,III

Valores perdidos

|  |  |
| --- | --- |
|  | Missing(%) |
| Diabetes | 2 |
| Hb-Al | 2 |
| HOMA-IR | 34 |
| Urea | 7 |
| Creatininio | 12 |
| Bilirrubina | 15 |
| BD | 17 |
| BI | 17 |
| Filtracion-glomerular | 12 |
| LDL | 2 |
| Haptoglobina | 5 |
| Proteina-C-Reactiva | 12 |
| Ceruloplasmina | 5 |
| Insulina-Basal | 34 |
| Tiroxina | 7 |
| Ferritina | 2 |
| IST | 2 |
| Cobre | 5 |
| Protombina | 2 |
| Transferrina | 2 |

- El atributo sexo no se tiene en cuenta, todos los pacientes son mujeres.

-Se incluye el atributo categórico Diabetes con dominio {No, Si-DM-2, Si-G, GBA}

- El atributo suero/plasma se elimina por tener un único valor en su dominio

- El atributo Síndrome metabólico se eliminó por tener demasiados valores perdidos.

- Se eliminó el atributo “Triiodotironina (no unida a proteina)” por tener solo dos valores de 41.

- Se eliminó el atributo “Triodotironina (80-200) ng/dl” por tener solo seis valores de 41.

-Se eliminó el atributo Autoanticuerpos por tener un único valor en su dominio

- El atributo IMC se calcula a partir de los atributos talla y peso.

- El atributo ICC se calcula a partir de los atributos cintura y cadera.

- HOMA-IR se calcula a partir de glucosa e insulina, la glucosa está completa pero la insulina tiene muchos valores perdidos.

- Fueron detectados algunos outliers en atributos como Trigliceridos, Creatininio y Protombina. Los outliers se borraron y se estimaron como si fueran valores perdidos.

- Los atributos con valores perdidos fueron remplazados por la moda (atributos discretos) o la mediana (atributos numéricos). El completado de datos se realizó teniendo en cuenta solamente a los pacientes que pertencen a la misma clase donde se encuentra el valor perdido.

- Se ejecutaron modelos de selección de atributos, todos calculan un ranking de los atributos, y al final se promedian todos los ranking. De esta manera se estima mejor los atributos. Se empleó una validación cruzada en la estimación. Todos los métodos de selección de atributos son supervisados. Esto métodos pueden extraer mejor información al tener en cuenta los grupos de pacientes, mientras que los métodos no supervisados no lo hacen.

- Se normalizaron todos los atributos, para que todos estuvieran en la misma escala, de esta manera los algoritmos que utilizan funciones de distancia no se ven influenciados por determinados atributos.

Genes

///////////////////////cala E-290 y muchos otros en escala de los

Pacientes: 41

Atributos descriptores: 65

Clases: N,I,II,III

Valores perdidos

|  |  |
| --- | --- |
|  | Missing(%) |
| Mogat 1 | 29.3 |
| Mogat 2 | 31.7 |

Atributos con outliers: IGFBP3, SOCS2, Akt1,CA150, CELF4, ESRP1, FB11, MAGOH, NOVA1, nSR100, PSF, RBM22, RBM45, SF3B TV1, SF3B TV2, SKIP, SND1, SRSF1, SRSF4, SRSF5, SRSF6, TIA1, U1, U11, U12, U2, U4, U4ATAC, U5, U6, Ghelin, sst5md4

- El atributo nsr100 se tuvo que eliminar pues tenía valores que muchos valores que estaban en la escala E-290 y muchos otros en escala de los millares, por otra parte tenía valores perdidos.