INDICE DEL TRABAJO

**Introducción**

Elección del DataSet.

Contexto y preguntas por responder

Descripción de las variables

**Carga del dataset**

Explicación de las variables

**Parte 1. Preprocesado de datos**

**Conversión y estandarización.**

* Sustituir , por .
* Int a numérico
* String a Factor

**Estudio de Outliers.**

Sustitución por NaN

* Frecuencia cardíaca máxima.

**Estudio de Valores nulos.**

Estudio de las distintas técnicas comprobando cual es la más eficiente para el modelo de clasificación:

* KNN
* Sustitución por valores desconocidos
* Sustitución por estadísticos.

Variables con valores nulos:

* RestingBP
* Choresterol

**Normalización**

(minA , maxA)Valores continuos entre -1 a 1

**Discretización de variables.**

Discretizar las variables continuas:

* Edad
* …

***Reducción de la dimensionalidad***

En función del modelo

**Parte 2. Análisis de los datos**

1. **Análisis descriptivo de las variables.**

Estudio de media, mediana, varianza. Mínimos y máximos de cada variable, etc

1. **Estudio normalidad de las variables.**
   1. Contrastes de normalidad
   2. TLC
2. **Contraste de hipótesis. Responder a las siguientes preguntas**
   1. El colesterol es mayor en mujeres que en hombres.
   2. La probabilidad de sufrir una enfermedad cardíaca es mayor en hombres que en mujeres.
   3. Otras
3. **Estudio de correlación entre las variables.**

Matriz de correlación

1. **Regresión lineal** (Las variables independientes a elegir para el estudio pueden cambiar según los análisis anteriores)
   1. **Regresión lineal simple:**

Variable dependiente: colesterol, y variable a elegir.

* 1. **Regresión lineal múltiple:**

Variable dependiente, colesterol, variable independiente: Añadiremos el resto de las variables a estudiar dependiendo del resultado del modelo.

1. **Regresión logística:**
   1. **Regresión logística simple:**

Variable dependiente: heart disease , independente: chest pain type

* 1. **Regresión logística simple:**

Variable dependiente: heart disease , independente: a estudiar el resto del variables.

1. **Clasificación**
2. **Partición de los datos**
   1. Técnica de k-fold
3. **Modelos de clasificación**
   1. Árbol de decisión
   2. Random Forest
   3. RPart
4. **Predicción**
5. **Conclusiones e interpretación de la matriz de confusión**
6. **Clustering (Opcional)**

Realizar un modelo de clustering que nos permita realizar grupos de pacientes con aquellas variables asociadas a un tratamiento médico o tipo de diagnóstico que a priori no tienen relación aparente.

**Parte 3. Visualización de los datos y conclusiones**