Trabajo Grupal de Prácticas: Fallos Cardíacos

Componentes.

- Arturo Cortés Sánchez
- Abel José Sánchez Alba

1. Introducción

En este documento se tratará sobre un dataset relativo a los ataques al corazón producidos en 90 pacientes durante su seguimiento bajo un hospital obtenido aquí. Sobre ellos, se conocen variables binarias como el sexo o la condición de fumador del paciente, variables continuas y normales como la edad de los pacientes y variables como la condición de diabetes de los pacientes. Al ser un dataset pequeño, la generalización y la extracción del conocimiento pueden ser algo complicadas, sin embargo, es un dataset bueno para comprobar el efecto de algunas acciones que se pueden realizar sobre los datos, como la eliminación de outliers, el agrupar las instancias en clusters etc. Esto se debe a que la mejora o el empeoramiento en la efectividad de los algoritmos de clasificación será notable y es sencillo observar los efectos de cada tipo de preprocesamiento o clasificador empleado.

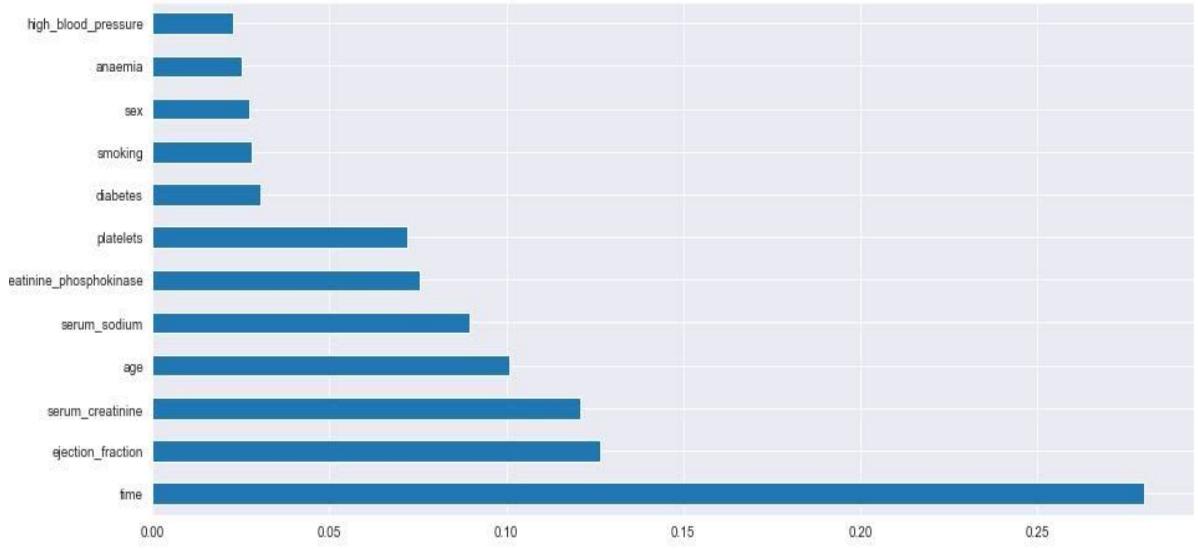
A lo largo de la práctica se ha realizado un análisis exploratorio de las variables, representando su distribución y correlación entre ellas, se han ejecutado algoritmos basados en árboles y algoritmos basados en la estadística por su adecuación al problema. En el caso de los árboles, su ejecución ayuda a la extracción del conocimiento pues es posible consultar sus criterios de información y la importancia de las variables del dataset en el algoritmo para la clasificación.

2. Exploración de variables

El dataset consta de una serie de variables relativas a los ataques al corazón. Comenzando por la variable principal, “DEATH_EVENT”, una variable indicadora de la mortalidad del ataque al corazón, y por tanto, la variable a clasificar. Se tienen variables relativas a los análisis de sangre de los pacientes, como son las plaquetas en sangre, anemia, diabetes etc.. Se tienen variables binarias, como la variable a clasificar, el hecho de que el paciente sea fumador, el sexo..etc. Y por otro lado se tienen variables normales como la edad, que se considera una variable continua y normal.

Es posible ver un análisis exploratorio dinámico en el notebook adjunto a este documento, el notebook de Python “ExploratorioDinamico.ipynb”

A continuación se exploran las variables más significativas, para ello, se ejecuta un algoritmo basado en árboles con el criterio de ganancia de información y se recurre a los parámetros de la clase del algoritmo “ExtraTrees”. Esta información se representa en la siguiente gráfica.



Se puede observar que, a partir de la variable indicadora de las plaquetas de una persona, estas aportan más información al algoritmo que variables como la presión arterial o el sexo. Por lo que se seleccionan, dentro del conjunto de datos, estas variables. En el notebook existen varias configuraciones de variables seleccionadas de modo que se puede analizar el comportamiento del algoritmo y la distribución de información. Esto queda libre al lector si así lo desea.

En esta memoria se tendrá en cuenta el conjunto de variables que consta de:

1. Plaquetas
2. creatinine_phosphokinase
3. serum_sodium
4. edad
5. serum creatinine
6. ejection fraction
7. time

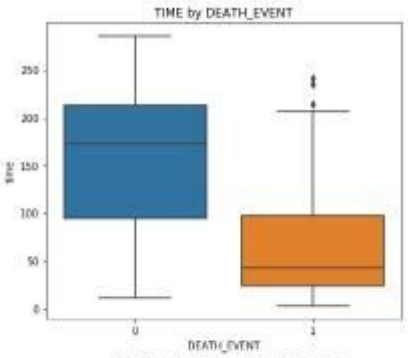
Se puede apreciar que la variable tiempo, indicadora del tiempo el cual el paciente se encuentra bajo seguimiento, es la variable más indicativa de la supervivencia al ataque al corazón. Por otro lado, la segunda variable más importante es “ejection_fraction”, esta es una variable indicadora del porcentaje de sangre que abandona el corazón tras cada contracción. El resto de variables corresponden a mediciones médicas autoexplicativas.

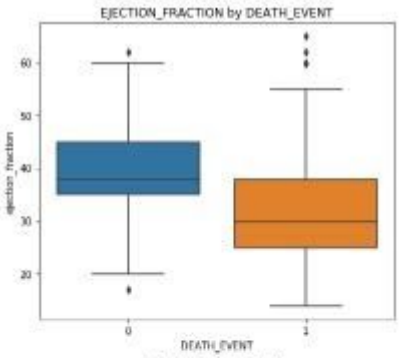
Con el fin de identificar casos que pueden perjudicar el comportamiento de los algoritmos de predicción, se representan diagramas boxplot de las variables seleccionadas que se pueden ver en la figura de la página posterior. En ella se puede ver como la variable “Ejection_fraction” presenta una instancia completamente fuera de los márgenes normales de la variable y por tanto, es una instancia que produce ruido y de cierta forma, contamina la

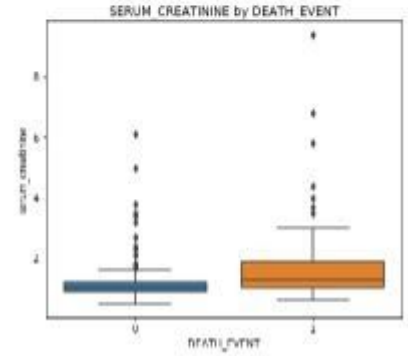
extracción de conocimiento. Por lo tanto, esta instancia es eliminada. Se puede ver otro caso donde la variable “creatinine_phosphicase” podría presentar instancias fuera de los márgenes comunes, sin embargo, tras investigar los valores de estas instancias y el conocimiento que representan, se ha decidido mantener las mismas.

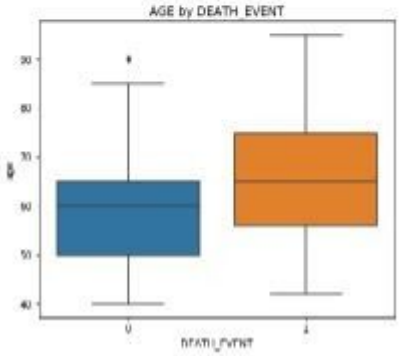
Dado que las variables seleccionadas son numéricas y normales, se han tenido en cuenta también algoritmos de clasificación propios de la estadística y la regresión multivariable, como son el algoritmo “Lineal discriminant analysis”(LDA), Naïve-Bayes y “Quadratic discriminant analysis”(QDA). Por otro lado, se tiene un conjunto de algoritmos de predicción basados en árboles de decisión, que son DecisionTree, RandomForest, GradientBoostTree,ExtraTree classifier. Finalmente se tiene un algoritmo del vecino más cercano.

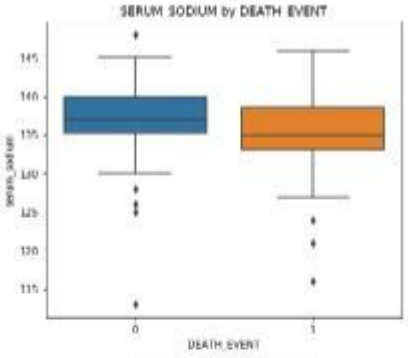
Para todos y cada uno de ellos, se creará una matriz de confusión de la cual se analizarán sus datos. Y para los algoritmos basados en árboles de decisión, se tiene la descripción de sus variables más representativas.

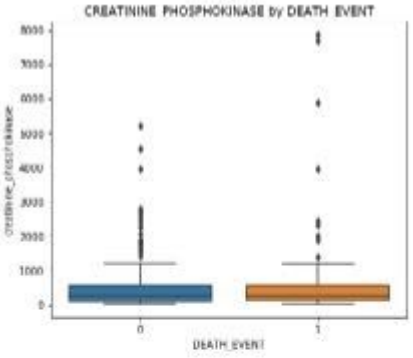

TIME by DEATH_EVENT
time
DEATH_EVENT

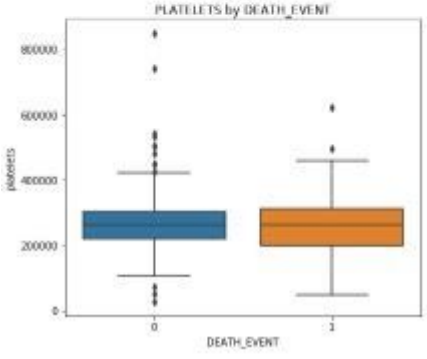

EJECTION_FRACTION by DEATH_EVENT
ejection_fraction
DEATH_EVENT


SERUM_CREATININE by DEATH_EVENT
serum_creatinine
DEATH_EVENT


AGE by DEATH_EVENT
age
DEATH_EVENT

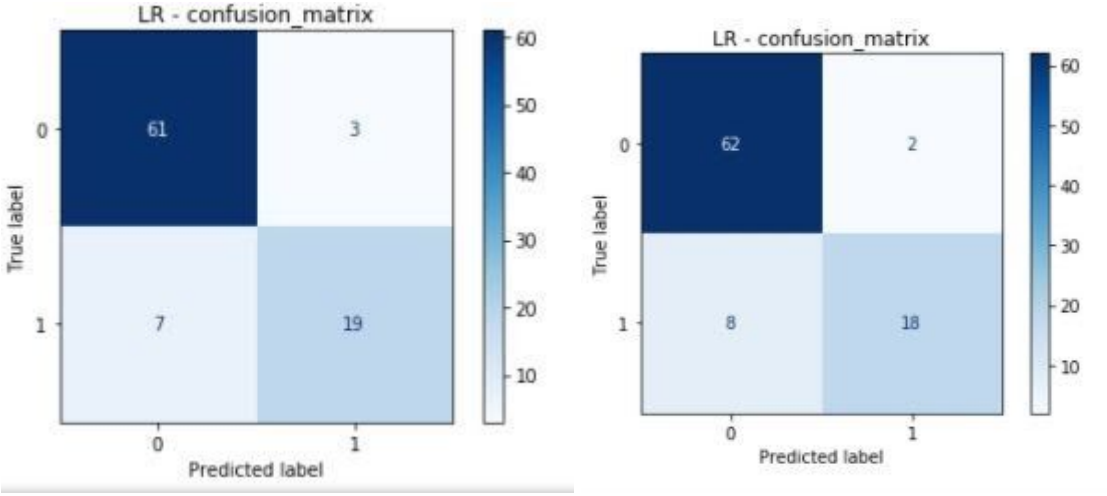

SERUM_SODIUM by DEATH_EVENT
serum_sodium
DEATH_EVENT


CREATININE_PHOSPHOKINASE by DEATH_EVENT
creatinine_phosphokinase
DEATH_EVENT


PLATELETS by DEATH_EVENT
platelets
DEATH_EVENT

3.Ejecución de los algoritmos

Se han ejecutado los algoritmos mencionados, en todos ellos se presenta una precisión bastante alta, por encima del 80%. Cabe destacar, de entre los algoritmos basados en árboles, el random forest, con una precisión de más del 88%. Y por parte de los algoritmos basados en la estadística, el LDA presenta una precisión similar. En la figura, se puede ver la matriz de confusión del random forest y del algoritmo LDA respectivamente.



Se puede ver como la clasificación de ambos algoritmos es similar. Entre las variables más representativas del algoritmo random forest se tiene la siguiente lista:

{'time': 0.3932150228678265, 'ejection_fraction': 0.16295587180998997, 'serum_creatinine': 0.18757721131664665, 'age': 0.1085802869865146, 'serum_sodium': 0.06858438342128573, 'creatinine_phosphokinase': 0.044669383573307596, 'platelets': 0.034417840024428936}

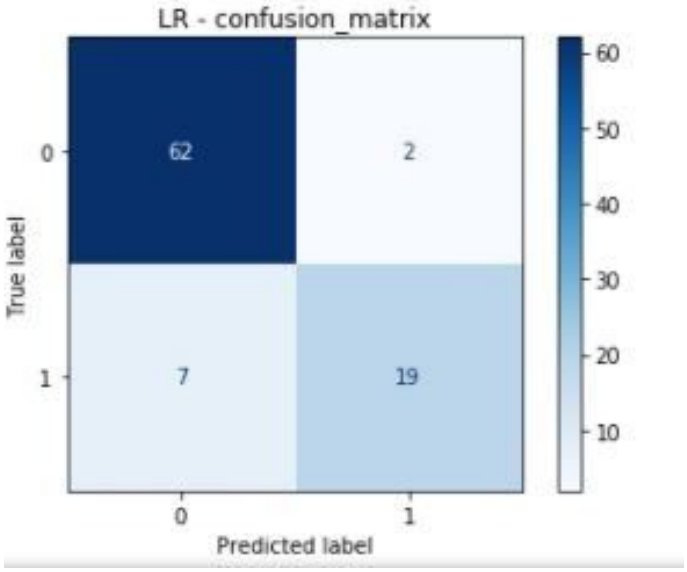
Se puede apreciar cómo, de nuevo, el tiempo en ser atendido un paciente representa la variable más importante del dataset.

4.Clustering como preprocesamiento.

Se ha ejecutado el algoritmo Kmeans como preprocesamiento del conjunto de datos, es decir, se toma la variable cluster dada por el algoritmo, y se añade al conjunto de variables del dataset. Esta operación se realiza con el objetivo de añadir información al dataset creando dichos clusters y así, ayudar a los algoritmos de decisión a mejorar su precisión. Pues se presenta información sobre la cercanía de las instancias, en definitiva, su cluster.

Una vez hecho esto, se ejecutan de nuevo dichos algoritmos. Se puede observar una mejora de los algoritmos basados en árboles, por lo que el preprocesamiento elegido es de

ayuda. Destaca la precisión del algoritmo random forest, que aumenta a un 90%. Donde su matriz de confusión es la siguiente



Se puede ver una ligera mejora en la clasificación. la información aportada por los clusters mejora levemente el comportamiento de los algoritmos. Al ser un dataset pequeño, cualquier pequeña mejora es sustancial, pero a su vez es difícil mejorar la abstracción de los algoritmos al poseer tan pocas instancias.

Las conclusiones se pueden extraer de las variables más empleadas por los algoritmos basados en árboles, es decir, las variables más importantes implicadas en la toma de decisiones de los mismos. De nuevo, es posible extraer la lista de variables del algoritmo, la cual es:

```
{'time': 0.4852948824630747, 'ejection_fraction': 0.13216178583589114,
'serum_creatinine': 0.0888216901567347, 'age': 0.07905896267487082,
'serum_sodium': 0.11133771588231858, 'creatinine_phosphokinase':
0.05241916588566074, 'platelets': 0.03641304347826086, 'cluster':
0.014492753623188404}
```

De nuevo, el tiempo se presenta como la variable más crucial para determinar si un ataque al corazón conlleva la muerte del paciente.

5. Reglas de asociación

Finalmente usando Knime se han ejecutado varios algoritmos de reglas de asociación. Para ello primero se han discretizado los datos siguiendo la discretización realizada en el enlace del dataset.

Los algoritmos ejecutados han sido:

- Apriori
- Hotspot
- PredictiveApriori
- FilteredAsociator
- Tetrius
- Asociation rule learner

De todas las reglas obtenidas, se han eliminado las que no contenían la característica DEATH_EVENT, pues lo que se busca es ver como esta característica es influida por las demás. Cómo seguían siendo demasiadas reglas, se volvió a filtrar el conjunto para mantener únicamente aquellas que tuviesen DEATH_EVENT como consecuencia.

serum_creatinine=05-139 time=228-259 28 ==> DEATH_EVENT=no 28 acc:(0.99354)
ejection_fraction=33-40 time=228-259 25 ==> DEATH_EVENT=no 25 acc:(0.9932)
serum_creatinine=05-139 time=172-200 23 ==> DEATH_EVENT=no 23 acc:(0.9929)
creatinine_phosphokinase=23-806 serum_creatinine=05-139 time=200-228 23 ==> DEATH_EVENT=no 23 acc:(0.9929)
age=50-55 creatinine_phosphokinase=23-806 high_blood_pressure=no 22 ==> DEATH_EVENT=no 22 acc:(0.99272)
age=50-55 creatinine_phosphokinase=23-806 serum_creatinine=05-139 22 ==> DEATH_EVENT=no 22 acc:(0.99272)
anaemia=no creatinine_phosphokinase=23-806 time=228-259 21 ==> DEATH_EVENT=no 21 acc:(0.99252)
age=50-55 creatinine_phosphokinase=23-806 sex=hombre 20 ==> DEATH_EVENT=no 20 acc:(0.99228)
age=60-65 ejection_fraction=33-40 19 ==> DEATH_EVENT=no 19 acc:(0.99202)
age=50-55 anaemia=no creatinine_phosphokinase=23-806 18 ==> DEATH_EVENT=no 18 acc:(0.99171)
ejection_fraction=33-40 time=200-228 16 ==> DEATH_EVENT=no 16 acc:(0.99094)
anaemia=no high_blood_pressure=no smoking=no time=228-259 16 ==> DEATH_EVENT=no 16 acc:(0.99094)
platelets=190080-272570 time=200-228 15 ==> DEATH_EVENT=no 15 acc:(0.99044)
platelets=190080-272570 time=4-32 15 ==> DEATH_EVENT=si 15 acc:(0.99044)
age=50-55 creatinine_phosphokinase=23-806 platelets=190080-272570 14 ==> DEATH_EVENT=no 14 acc:(0.98985)
anaemia=si time=200-228 13 ==> DEATH_EVENT=no 13 acc:(0.98913)
creatinine_phosphokinase=23-806 ejection_fraction=33-40 time=4-32 13 ==> DEATH_EVENT=si 13 acc:(0.98913)
ejection_fraction=33-40 time=172-200 13 ==> DEATH_EVENT=no 13 acc:(0.98913)
age=40-45 anaemia=no platelets=190080-272570 11 ==> DEATH_EVENT=no 11 acc:(0.98714)
age=50-55 creatinine_phosphokinase=23-806 smoking=si 11 ==> DEATH_EVENT=no 11 acc:(0.98714)
diabetes=no time=116-144 10 ==> DEATH_EVENT=no 10 acc:(0.98573)

De la mayoría de reglas finales se deduce que el paciente sobrevive, pero hay dos reglas en las que el paciente muere, y una de ellas es bastante interesante, ya que es muy similar a otra en la que sobrevive.

platelets=190080-272570 time=200-228 15 ==> DEATH_EVENT=no 15 acc:(0.99044)
platelets=190080-272570 time=4-32 15 ==> DEATH_EVENT=si 15 acc:(0.99044)

Para cierto rango de plaquetas se ha observado que la variable tiempo está ampliamente relacionada con el hecho de que un fallo cardíaco en un paciente conduzca a la muerte. No obstante, no se recomienda tomar esta variable para realizar predicciones sobre nuevos pacientes en el momento inicial del seguimiento, pues el valor de la variable tiempo es

tomado en el momento que el paciente sufre el fallo, por lo que es imposible estimarlo en el inicio de su seguimiento.

Como curiosidad se han seleccionado estas dos reglas ya que remarcan claramente la relacion entre sexo de una persona, si es fumadora y si ha muerto en el incidente.

```
sex=mujer DEATH_EVENT=no 71 ==> smoking=no 70        <conf:(0.99)> lift:(1.45) lev:(0.07) [21] conv:(11.4)
smoking=si DEATH_EVENT=no 66 ==> sex=hombre 65       <conf:(0.98)> lift:(1.52) lev:(0.07) [22] conv:(11.59)
```