MACHINE LEARNING

Trabajo de Clasificación binaria

Este trabajo está orientado a predecir una variable **binaria** a través de diferentes algoritmos de clasificación.

Normas para la realización del trabajo

1) El trabajo es individual.
2) Se entregará el trabajo en pdf en una copia digital enviada al Campus Virtual.
3) El trabajo deberá estar explicado (no basta con responder a las cuestiones), indicando, si es necesario, el código utilizado. Se valora la claridad de exposición en el informe y la estructura. Puede contener Anexos de datos y gráficos o no, todo según vuestro criterio. El trabajo es libre, con lo cual se agradece sentido común.
4) El trabajo se debería recuperar en septiembre si se da al menos una de las siguientes circunstancias:

- La presentación y explicaciones son escasas.
- Los modelos comprobados son demasiado limitados.
- Se observan signos de copia de otros trabajos de otros alumnos.

5) Se debe responder obligatoriamente a cada una de las cuestiones a) hasta la f) expuestas más abajo. Si falta una sola de las cuestiones o algoritmos el trabajo está para septiembre.

Se construirán los modelos sobre **una** matriz de datos a elegir voluntariamente. La calificación también dependerá de la complejidad de los datos a tratar. Se piden al menos cerca de 500 observaciones y 5 variables input posibles, de las cuales al menos una debe ser categórica.

En el ANEXO hay información sobre datos a utilizar. Podéis utilizar datos vuestros, de otras webs, etc. siempre que se cumplan las condiciones.

Cuestiones generales a responder

Se trata de conseguir obtener el mejor método/algoritmo para predecir, estable en cuanto a su performance, y en comparación con un modelo de regresión logística con selección de variables stepwise.

a) Se deben realizar pruebas suficientes para obtener una buena selección de variables, obteniendo uno o varios conjuntos de variables tentativos
b) Se requiere la comparación entre los mejores algoritmos y regresión logística ;
c) Se comprobará el efecto de la variación de los parámetros básicos de cada algoritmo (tuneado) (número de nodos en redes, shrink en gradient boosting, etc.).
d) Los algoritmos a utilizar son obligatoriamente y como mínimo:
    - Redes Neuronales
    - Regresión Logística
    - Random Forest
    - Bagging
    - Gradient Boosting
    - Support Vector Machines

- También si se quiere y para comprender los datos se puede probar con un simple árbol.

e) Es necesario utilizar validación cruzada, validación cruzada repetida o como mínimo training/test repetido. Se comparará con diferentes particiones y semillas.

f) Es necesario hacer alguna prueba de ensamblado.

Se puede utilizar el programa R aportado por el profesor y cualquier modificación o uso de código. También se puede utilizar R y/o cualquier otro paquete, siempre que se cumplan los apartados (a)-(f).

La calificación tendrá en cuenta (sin orden): a) La complejidad de los datos a tratar b) Que sea completo en cuanto a la exploración de posibilidades c) Que se tenga correctamente en cuenta el remuestreo para elegir el modelo d) el manejo de código.

ANEXO

WEBS DE DATASETS PARA APLICAR TÉCNICAS DE MACHINE LEARNING

RECOMENDADOS PARA EMPEZAR, MUY BIEN ESTRUCTURADOS
****
https://archive.ics.uci.edu/ml/datasets.html
https://sci2s.ugr.es/keel/datasets.php
https://vincentarelbundock.github.io/Rdatasets/datasets.html
****
1) En uci y keel están los archivos ordenados por clasificación o regresión.
****
2) En la web Rdatasets_
https://vincentarelbundock.github.io/Rdatasets/datasets.html
Archivos de más de 1200 observaciones, lista elaborada por mí:

Item	Rows	Cols	modelo
BEPS	1525	10	clasificación
Caravan	5822	86	clasificación
Gunnels	1592	10	clasificación
Hdma	2381	13	clasificación
Hmda	2381	13	clasificación
VerbAgg	7584	9	clasificación
WVS	5381	6	clasificación
Wells	3020	5	clasificación
YouthRisk2007	13387	6	clasificación
azcabgptca	1959	6	clasificación
dengue	2000	13	clasificación
flchain	7874	11	clasificación
mexico	1359	33	clasificación
mifem	1295	10	clasificación
monica	6367	12	clasificación
ohio	2148	4	clasificación
spam7	4601	7	clasificación
student	9679	13	clasificación
turnout	2000	5	clasificación
voteincome	1500	7	clasificación
Car	4654	70	clasificación multiclase
Chile	2700	8	clasificación multiclase
Kakadu	1827	22	clasificación multiclase
msqR	6411	79	correspondencias
colon	1858	16	cox
cricketer	5960	8	cox
mgus2	1384	10	cox
nwtco	4028	9	cox
BudgetFood	23972	6	regresión
BudgetItaly	1729	11	regresión
BudgetUK	1519	10	regresión
Computers	6259	10	regresión
DoctorContacts	20186	15	regresión
HI	22272	13	regresión
InstInnovation	6208	25	regresión
Males	4360	12	regresión
Males	4360	12	regresión
MathPlacement	2696	16	regresión
MedExp	5574	15	regresión
PatentsRD	1629	7	regresión
SLID	7425	5	regresión
SaratogaHouses	1728	16	regresión
Schooling	3010	28	regresión
Snmesp	5904	8	regresión
Star	5748	8	regresión
VietNamH	5999	11	regresión
Vocab	30351	4	regresión
Wage	3000	11	regresión
Wages	4165	12	regresión
Wages	4165	12	regresión
Workinghours	3382	12	regresión
azpro	3589	6	regresión
baseball	21699	22	regresión
diamonds	53940	10	regresión
mdvis	2227	13	regresión
medpar	1495	10	regresión
nlschools	2287	6	regresión
rwm5yr	19609	17	regresión
science	1385	7	regresión
DoctorAUS	5190	15	regresión muchos dep posibles ver web
OFP	4406	19	regresión varias dependientes

VietNaml	27765	12	regresión y clasi
Gestation	1236	23	regresión y clasificación
NCbirths	1450	15	regresión y clasificación

ZIPEADOS POCO EXPLICADOS

http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets?CGISESSID=10713f6d891653ddcbb7ddbdd9cffb79
https://www.cs.waikato.ac.nz/ml/weka/datasets.html

DATOS COMPLICADOS PERO INTERESANTES

https://www.nature.com/sdata/
http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html

CONCURSOS

https://www.kaggle.com/
https://www.drivendata.org/competitions/
http://www.chalearn.org/challenges.html
https://www.kdd.org/kdd-cup