Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Taller 8: Ensamble

Análisis de datos – Prof. David Porta

Maestría en Estadística Aplicada

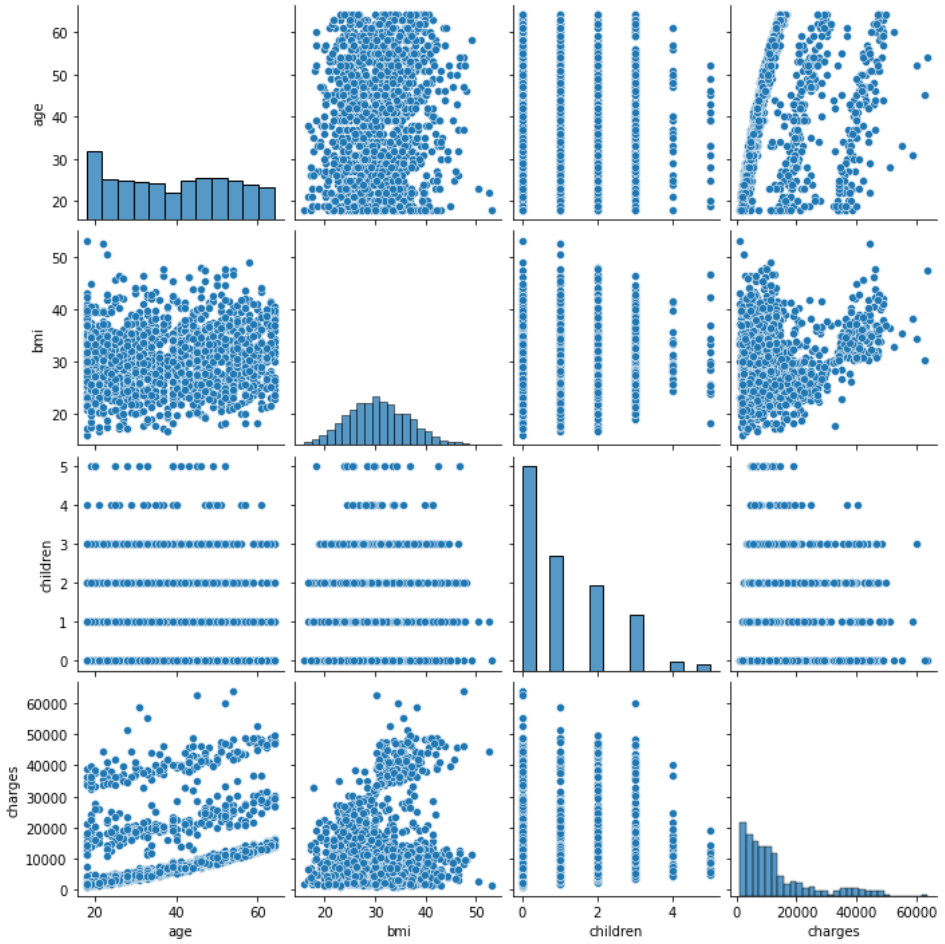
TRANSFORMAMOS VIDAS
CON EDUCACIÓN DE EXCELENCIA

Presentado por:

DIEGO HERRERA MALAMBO - malambod@utb.edu.co

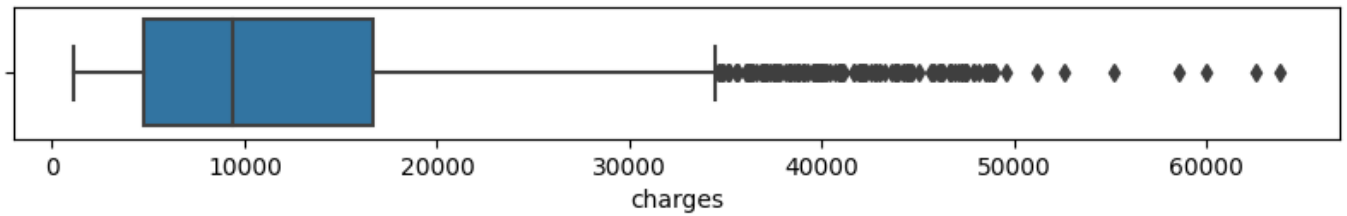
Dataset

El dataset corresponde a los datos simulados de costo de servicios de salud de una aseguradora de pacientes de Estados Unidos creado desde los datos demgraficos del U.S. Census Bureau y refleja aproximadamante las condiciones reales en su momento (2013). Este dataset costa de 7 variables de las cuales una es la dependiente continua (charges). Para mayor referencia se puede consultar el libro ***Machine Learning with R*** en la pagina 173.



	count	mean	std	min	25%	50%	75%	max
age	1,338.00000	39.20703	14.04996	18.00000	27.00000	39.00000	51.00000	64.00000
sex	1,338.00000	0.49477	0.50016	0.00000	0.00000	0.00000	1.00000	1.00000
bmi	1,338.00000	30.66340	6.09819	15.96000	26.29625	30.40000	34.69375	53.13000
children	1,338.00000	1.09492	1.20549	0.00000	0.00000	1.00000	2.00000	5.00000
smoker	1,338.00000	0.20478	0.40369	0.00000	0.00000	0.00000	0.00000	1.00000
charges	1,338.00000	13,270.42227	12,110.01124	1,121.87390	4,740.28715	9,382.03300	16,639.91251	63,770.42801
region_northeast	1,338.00000	0.24215	0.42855	0.00000	0.00000	0.00000	0.00000	1.00000
region_northwest	1,338.00000	0.24290	0.42900	0.00000	0.00000	0.00000	0.00000	1.00000
region_southeast	1,338.00000	0.27205	0.44518	0.00000	0.00000	0.00000	1.00000	1.00000
region_southwest	1,338.00000	0.24290	0.42900	0.00000	0.00000	0.00000	0.00000	1.00000

Distribución de costo



Notebook: https://github.com/dherrerambo/analisis-de-datos-avanzados/blob/master/Clase%208%20-%20Ensamble%20Models/Tarea8.ipynb

Implementación de métodos de Ensamble

Se implementaron 10 modelos de regresión: ***RandomForestRegressor, AdaBoostRegressor, GradientBoostingRegressor, ExtraTreesRegressor, BaggingRegressor, HistGradientBoostingRegressor, KNeighborsRegressor, DecisionTreeRegressor, MLPRegressor, LinearRegression***.
Para validación de los modelos con Cross-Validation de 5 folds se uso el scoring por Mean Square Error (MSE) y para comparación de los modelos se usaron dos métricas R2 y RMSE.
Como dato relevante se tiene que el modelo GradientBoostingRegressor tuvo un desempeño que es casi igual al de los métodos de Stacking o Voting, pero debido a la complejidad de estos modelos, seria ideal usar este modelo o RandomForest.

Algorithms	R2	RMSE
Stacking(All)	0.88277	0.06631
GradientBoostingRegressor	0.88237	0.06642
Stacking(Best)	0.88077	0.06687
Voting(Best)	0.87406	0.06873
Voting(All)	0.86749	0.07050
RandomForestRegressor	0.85900	0.07272
HistGradientBoostingRegressor	0.85780	0.07303
BaggingRegressor	0.85355	0.07411
AdaBoostRegressor	0.84667	0.07583
ExtraTreesRegressor	0.82072	0.08200
KNeighborsRegressor	0.78416	0.08997
MLPRegressor	0.77150	0.09257
LinearRegression	0.76498	0.09389
DecisionTreeRegressor	0.66933	0.11136

Notebook: https://github.com/dherrerambo/analisis-de-datos-avanzados/blob/master/Clase%208%20-%20Ensamble%20Models/Tarea8.ipynb