Proceso de carga de datos en RapidMiner y los hallazgos en el Dataset Precios de la vivienda: técnicas avanzadas de regresión

Bryan Olivares, John Vásconez.

Escuela de Formación de Tecnólogos
Escuela Politécnica Nacional
Dirección Postal.
bryan.olivares@epn.edu.ec, john.vasconez@epn.edu.ec.

Resumen- En el presente documento se detalla la manera en que se realiza el proceso de carga de datos en la herramienta RapidMiner. El caso de estudio consta de un Dataset .csv de los precios de la vivienda: técnicas avanzadas de regresión en donde se requiere predecir los precios de venta. Se escogerán 3 algoritmos y los atributos necesarios para realizar el modelo predictivo. Posteriormente los Dataset de training y testing se volcarán a la base de datos local no SQL MongoDB. Finalmente se creará y subirán los archivos en un repositorio en la nube usando GitHub.

Palabras Clave- RapidMiner, MongoDB, modelo predictivo.

I. INTRODUCCIÓN

"El diseño de modelos predictivos es uno de los primeros frentes del trading cuantitativo. Para identificar tendencias, la regresión lineal es la más empleada, aunque también son frecuentes las parabólicas, polinómicas y logarítmicas. Sin embargo, estos algoritmos son bastante ineficientes en espacios multidimensionales en los que intervienen más de dos variables. En estos casos suele ser mejor utilizar algoritmos basados en modelos probabilísticos o estadísticos."[1].

El uso de la herramienta RapidMiner permitirá obtener una predicción del precio de venta de las propiedades en base a la elección de los atributos que tengan mas relación con esta variable. El Dataset que nos permitirá realizar este análisis es obtenido del enlace: https://www.kaggle.com/c/house-prices-advanced-regression.techniques/data_gue_contiene_un_archivo

regression-techniques/data que contiene un archivo train.csv que es el conjunto de entrenamiento, un archivo test.csv que es el conjunto de prueba, un data_description.txt que es una descripción completa de cada columna y un archivo sample_submission.csv que es una presentación de referencia de una regresión lineal en año y mes de venta, cantidad de lotes cuadrados y número de habitaciones.

II. DESARROLLO

Se busca construir un modelo predictivo empleado los datos del Dataset del caso de estudio de los precios de las viviendas. Es necesario elegir los atributos que esten mas relacionados con la variable que se desea predecir, en este caso la variable de estudio es "SalePrice". Primeramente, se realiza la carga de los datos; posteriormente, si es necesario, se realiza una preparación o limpieza de los datos utilizando

los operadores que RapidMiner nos ofrece. Luego se realiza la validación de los atributos que serán utilizados para predecir la variable objetivo "SalePrice" y finalmente generar el modelo utilizando 3 algoritmos diferentes y observar cual de estos resulta ser el de mayor precisión. La Fig. 1 muestra el esquema de desarrollo [1]

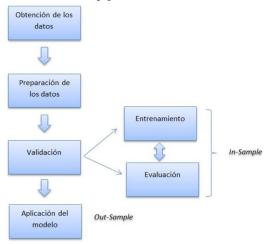


Fig. 1. Esquema general de desarrollo.

A. Análisis y carga de los datos

En la Fig. 2 se muestra el archivo de entrenamiento "train.csv" que consta de 81 columnas y 1460 registros. La Fig. 3 muestra el archivo de de prueba "test.csv" que tiene 80 columnas y 1459 registros, en donde este archivo no contiene el atributo "SalePrice" y que servirá para realizar el modelo.

	49 NL	- 15	1450 Pers	MA	Beg	Let	AllFub	Incide	GH.	CelligCr	Monm.	Horm.	Fon	25tery	7		
2	20 Ft.	10	9600 Pass	NA	Req	Lel	ARPA	FR2	GN	Vessher	Foods	Nane.	Fin	15tery	- 6		
3	69.86	6.0	19250 Pere	MA	IB1	Let	Allha	Incide	GH .	CelligCr	Norm.	Horm.	Fon	2Stery	7		
- 4	79 Ft.	60	9550 Pana	NA	IR1	Lel	AIPA	Carner	Gil	Creafer	Norm.	Nane.	Fin	2Stary	7		
- 5	60 FG	14	14244 Pere	MA	IB1	Let	AllPub	FB2	64	NeRidon	Norm.	Herm.	Fen	2Stery			
- 6	50 FG	15	14115 Page	NA	IB1	Let	AIPA	Incide	Gil	Minchell	Norm.	Harm.	Fin	1.5Fin			
7	20 FG	15	10004 Pers	NA	Res	L-I	AllPub	bride	GH	Sement	Norm.	Nane.	Fen	Shery			
	49 No.	HA	90242 Page	MA	IR1	Let	Allha	Carner	QH .	MARKET	Part	Hann.	Fon	25tery	7	- 6	
4	So RM	91	6520 Pers	NA	Reg	Lel	AllPut	bride	GN	Old Town	detary	Nane.	Fen	1.5Fin.	7		
99	199.7%	50	7420 Pere	MA	Fire a	Let	ATPW	Carner	QH .	Bv9:516+	Actesy	Artery	2fmCon	1504		4	
- 11	20 Ft.	76	15200 Pave	NA	Req	Lel	ARPA	loide	GN	Sauger	Norm.	Name.	f en	Stary	s	9	
12	69.86	15	19924 Pere	MA	181	Let	Allha	Incide	GH .	Mridatte	Monm.	Horm.	Fon	2Stery			
12	20 Fit.	HA.	S2555 Pane	NA	IR2	Let	MIPA	Incide	Gil	Source	Norm.	Mann.	Fin	15tery	5	- 6	
14	20 %	91	10452 Pere	NA.	IB1	L+I	AllPub	Incide	64	CelleCr	Norm.	Herm.	Fen	15terr	7		
15	20 FG	HA	99920 Pane	NA	IB1	Let	AIPA	Carner	Gil	NAmer	Norm.	Harm.	Fin	15tery	- 6		
16	45 524	51	6120 Pers	NA	Res	L-I	AllPub	Carner	GH	BricSide	Norm.	Nane.	Fen	1500	7		
17	29 76	HA	11241 Page	MA	IR1	Let	Allha	Outposes	Qt/	MAnue	Morm.	Hann.	Fon	12tery	- 6	7	
10	99 FG.	72	16741 Pere	NA	Reg	Lel	AllPut	bride	GN	Source	Nerve.	Nane.	Daples	Shery	4		
99	29 FG	**	13645 Page	MA	Fire a	Let	ATPA	Incide	QH .	SourceY	PEA.	Horm.	Fon	12tery			
20	20 Ft.	76	7560 Pass	NA	Reg	Lel	ARPA	loide	GN	NAmer	Norm.	Name.	f en	Stary	s	- 6	
21	69 FG	191	14215 Pere	MA	161	Let	Allha	Carner	GH.	Mridatte	Monm.	Horm.	Fon	2Stery			
22	45 RM	57	7669 Pass	Grot	Req	Dale	AIPA	loide	Gil	IDOTER	Norm.	Nane.	Fin	1504	7	7	
23	20 %	15	9742 Pere	MA	Bea.	L+I	AllPub	Incide	64	CelleCr	Norm.	Herm.	Fon	15herr			
24	120 Fe4	- 44	4224 Page	NA	Bee	Let	AIPA	Incide	Gil	Meedes/7	Norm.	Harm.	Tuntet	15tery		7	
29	20 FG	NA.	1246 Pers	NA	IB1	L-I	AllPub	Incide	GH	Source	Norm.	Nane.	Fen	Shery			
26	20 76	110	94229 Page	MA	Fire a	Let	Allha	Carner	Qd	Middellt	Morm.	Hann.	Fon	12tery			
27	20 FG.	60	7200 Pere	255	Reg	Lel	AllPub	Career	GN	Namer	Norm.	Norm.	Fen	Shery	9	7	
24	29 76	94	19470 Page	MA	Fire a	Let	ATPA	Incide	QH .	Middelft	Monm.	Horm.	Fon	12tery			
29	20 Ft.	47	16321 Pass	NA	IR1	Lel	ARPA	Cultisee	GN	NAmer	Norm.	Nane.	F en	Sitery	s	- 6	
30	39.595	6.0	6324 Pere	MA	161	Let	Allha	Incide	QH .	BrkS16+	Foods	FIRMs	Fon	15terr	- 4	4	
24	79 O (40)	Se	6500 Pana	Pere	Req	Lel	MIPA	loide	Gil	IDOTER	Feeds	Nane.	Fin	2Stary	4	- 4	
32	29 FG	HA	1544 Pere	MA	IB1	L+I	AllPub	CulDSec	64	Source	Norm.	Herm.	Fon	15herr		6	
22	20 Fit.	15	19049 Pass	NA	Reg	Lel	AIPA	Carner	Gil	CalligOr	Norm.	Hann.	Fin	15tery			
34	20 FG	76	10552 Pere	MA	IB1	Lel	AllPub	Incide	GH	Namer	Norm.	Herm.	Fen	Shery	9	- 5	
25	120 FG	60	7212 Page	MA	Fire a	Let	AIDA	Instide	Qd	Middellt	Morm.	Hann.	Tuntet	12tery			
24	60 RL	160	12410 Pers	NA	Reg	L+I	AllPut	bride	GN	Neidalle	Nerve.	Nane.	Fen	2Stery			
37	20 ML	112	90059 Pany	MA	Beg	Let	Alltra	Carner	GH .	CelligOr	Monm.	Hore.	Fon	15tery			
29	20 Ft.	74	1922 Pass	NA	Reg	Lel	ARPA	beide	GN	NAmer	Norm.	Nane.	F-m	Shery	9	- 6	
39	20 FL	- 64	T922 Pere	MA	Bea	Let	AllPub	Incide	GH.	MAmor	Horm.	Hore.	Fon	1Dery		7	
40	99 FG	65	6040 Pane	NA	Bee	Let	MIPA	Incide	GH	Edwards	Norm.	Name.	Deeles	15tery	- 4		

Fig. 2. Archivo de entrenamiento "train.csv".

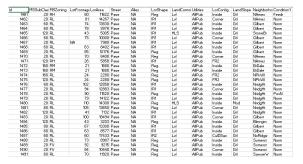


Fig. 3. Archivo de prueba "test.csv".

Ahora, utilizando la herramienta RapidMiner se procede a realizar la carga de los Dataset "train.csv" y "test.csv" utilizando el operador "Read CSV" como se muestra en la Fig. 4.

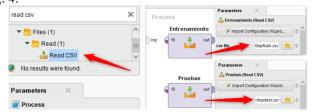


Fig. 4. Carga de Dataset "train.csv" y "test.csv".

B. Selección de los atributos

La selección de los atributos debe ser elegidos de acuerdo con la influencia que estos tienen en el atributo "SalePrice". La Fig. 5 muestra los atributos para el Dataset entrenamiento y la Fig. 6 el Dataset prueba que se eligieron. En ambos casos se han elegido 7 atributos.

LotArea	YearBuilt	YearRemodA	GarageCars	GarageArea	FullBath	GrLivArea	SalePrice
8450	2003	2003	2	548	2	1710	208500
9600	1976	1976	2	460	2	1262	181500
11250	2001	2002	2	608	2	1786	223500
9550	1915	1970	3	642	1	1717	140000
14260	2000	2000	3	836	2	2198	250000
14115	1993	1995	2	480	1	1362	143000
10084	2004	2005	2	636	2	1694	307000
10382	1973	1973	2	484	2	2090	200000
6120	1931	1950	2	468	2	1774	129900
7420	1939	1950	1	205	1	1077	118000
11200	1965	1965	1	384	1	1040	129500
11924	2005	2006	3	736	3	2324	345000
12968	1962	1962	1	352	1	912	144000
10652	2006	2007	3	840	2	1494	279500
10920	1960	1960	1	352	1	1253	157000
6120	1929	2001	2	576	1	854	132000
11241	1970	1970	2	480	1	1004	149000
10791	1967	1967	2	516	2	1296	90000
13695	2004	2004	2	576	1	1114	159000
7560	1958	1965	1	294	1	1339	139000
14215	2005	2006	3	853	3	2376	325300
7449	1930	1950	1	280	1	1108	139400
9742	2002	2002	2	534	2	1795	230000

Fig. 5. Selección de datos de entrenamiento.

LotArea	YearBuilt	YearRemodA	GarageCars	GarageArea	FullBath	GrLivArea
11622	1961	1961	1	730	1	896
14267	1958	1958	1	312	1	132
13830	1997	1998	2	482	2	162
9978	1998	1998	2	470	2	160
5005	1992	1992	2	506	2	128
10000	1993	1994	2	440	2	165
7980	1992	2007	2	420	2	118
8402	1998	1998	2	393	2	146
10176	1990	1990	2	506	1	134
8400	1970	1970	2	525	1	88
5858	1999	1999	2	511	2	133
1680	1971	1971	1	264	1	98
1680	1971	1971	1	320	1	109
2280	1975	1975	2	440	2	145
2280	1975	1975	1	308	1	83
12858	2009	2010	3	751	2	233
12883	2009	2010	3	868	2	154
11520	2005	2005	3	730	2	169
14122	2005	2006	3	678	2	182
14300	2003	2004	3	958	2	269
13650	2002	2002	3	756	2	225
7132	2006	2006	2	484	2	137
18494	2005	2005	2	430	2	132

Fig. 6. Selección de datos de prueba.

C. Algoritmo Decision Tree

El algoritmo Decision Tree "es un árbol como una colección de nodos destinados a crear una decisión sobre la afiliación de valores a una clase o una estimación de un valor objetivo numérico. Cada nodo representa una regla de división para un atributo específico. Para la clasificación, esta regla separa los valores que pertenecen a diferentes clases, para la regresión los separa para reducir el error de manera óptima para el criterio del parámetro seleccionado."[2].

La implementación del algoritmo y el árbol de decisión se muestra en la Fig. 7 y 8.

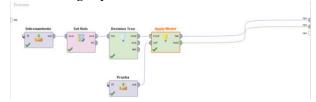


Fig. 7. Algoritmo "Decision Tree".

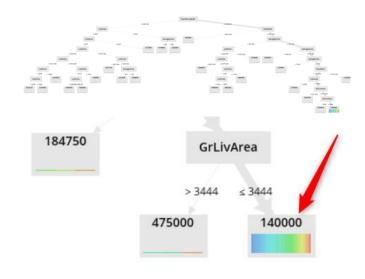
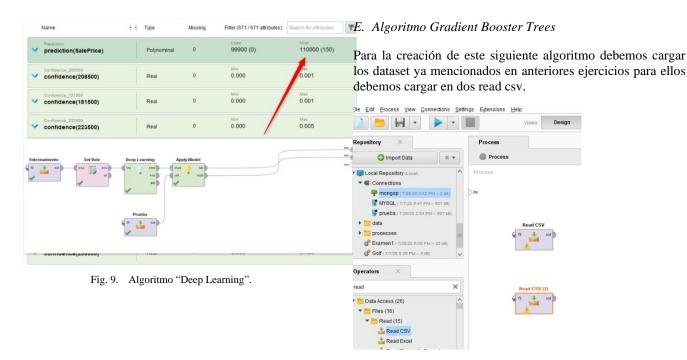


Fig. 8. Árbol de decisión.

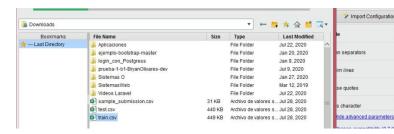
D. Algoritmo Deep Learning

"Deep Learning (DL) que tiene una desviación estándar de +-0.5% (el segundo mejor por este indicador) con un Error Relativo de 9.9% (El segundo mejor por este indicador)."[2].

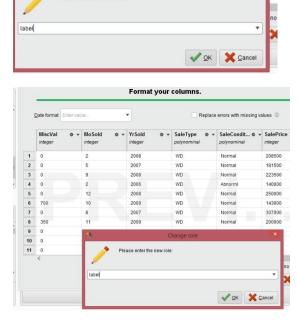
La Fig. 9 muestra el resultado de este algoritmo.



Para cargar los data set nos vamos al apartado de Import Configuration Wizard e importamos primero a train

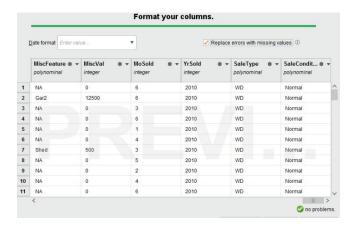


Al momento de cargar el train debemos cambiar el rol a la columna price a label.

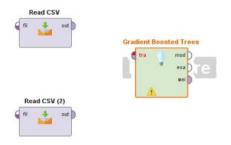


Please enter the new role

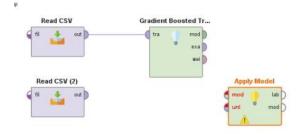
Para cargar el data set de test no va ser necesario cambiar el rol ya que es la columna que vamos a predecir.



En este momento vamos a seleccionar el algoritmo en este caso gradient el que tiene menos porcentaje de error.



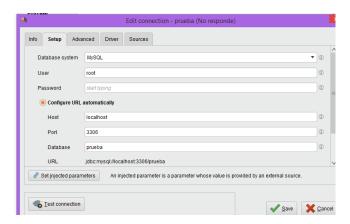
Para aplicar este modelo gradient vamos a unirlo con Apply Model.



Ahora si todo el procedimiento se completó correctamente debemos correrle y nos aparecerá el siguiente cuadro con los resultados.



Ahora vamos a la creacion de las conecciones para poder tener los datos en mongo db y en mysql.



III. CONCLUSIONES

El algoritmo de "Decision Tree" nos da una predicción del precio de venta es 140000 aproximadamente y en el algortimo "Deep Learning" es de 110000 el mayor precio de venta.

REFERENCIAS

- [1] "Minería de datos: Modelos predictivos con RapidMiner." http://www.tradingsys.org/mineria-de-datos-modelos-predictivos-con-rapidminer (accessed Jul. 28, 2020).
- [2] M. René and P. Eddy, "Minería de Datos Caso de Estudio House Prices."