Informe sobre el proceso de carga de datos en RapidMiner y análisis de hallazgos encontrados en un dataset en Keaggle sobre "precios de la vivienda: técnicas avanzadas de regresión"

Xavier Calle, Sebastián Morales
Escuela de Formación de Tecnólogos
Escuela Politécnica Nacional
xavier.calle@epn.edu.ec, sebastian.morales@epn.edu.ec.

***Resumen*-**
En este documento se detalla los procesos a seguir para cargar, transformar y depurar datos en RapidMiner, así mismo se utilizará modelos predictivos para el análisis de un dataset. El dataset posee 3 archivos, el archivo sample_submission.csv posee información del precio de cada vivienda, test.csv posee la misma información que train.csv exceptuando el precio de venta de cada vivienda que será lo que nuestro modelo debe predecir, por último, el archivo train.csv posee toda la información referente a las características que contiene una vivienda como su tipo, clasificación, tamaño, barrio en el que se encuentra, año de construcción, etc. A partir de la obtención de estos datos se debe seleccionar que características son las que más influyen en el campo "SalePrice" el mismo que representa el precio de la vivienda y que se define como variable objetivo, la cual se va a predecir mediante un modelo. El objetivo de este informe es seleccionar los principales atributos que influyen en la variable objetivo y los modelos más adecuados para la predicción de dato.

***Palabras Clave*- RapidMiner, minería de datos, modelos predictivos, tuplas, atributos, dataset.**

I. INTRODUCCIÓN

"Las razones para comprar una casa son muy diversas, puedes desear adquirir una casa porque te has casado y deseas empezar una vida más familiar y requieres de un lugar para hacerlo, otra razón puede ser que tienes el dinero para invertirlo y no tienes una mejor opción de inversión que el inmueble, también puedes desear mudarte del lugar en donde estas o tienes planeado pasar mucho tiempo y buscas un sitio para vivir cómodamente, si no estás a gusto con tu vivienda actual, una buena opción es mudarte, si ya estás cansado del ambiente donde resides, puede ser un buen momento para adquirir una nueva propiedad y mudarte." [1]. En tiempos actuales las empresas de inmuebles buscan recopilar datos acerca de las características generales de las viviendas y así poder utilizar estos datos para predecir las tendencias de los usuarios interesados en adquirir viviendas propias. Además, es importante recalcar la estrecha relación que existe entre el precio de las viviendas y la facultad de los usuarios en adquirirlas.

Las empresas que utilizan modelos predictivos para el tratamiento de datos actuales de los usuarios y la correlación que tienen sus compras con el precio y características de las viviendas tienen una ventaja dentro del mercado de bienes inmuebles. El dataset que utilizaremos nos presenta variables explicativas que abarcan toda la información acerca de las viviendas y nos permite aplicar modelos predictivos que nos permitan predecir el precio final de la vivienda.

II. DESARROLLO

En este trabajo se usa el dataset que se encuentra situado en el siguiente link: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data [2] que corresponde a una base de datos bastantes atributos de las viviendas y su precio final. Kaggle es una página web que dispone de datasets de carácter público, se pueden encontrar cualquier tipo de datasets en esta plataforma, basta con buscar el tema de interés para que se muestren resultados exactos o relacionados al tema de interés, la mayoría de las veces se encuentran datasets limpios, es decir que no poseen datos erróneos dentro de los atributos pertinentes de la tabla, no poseen vacíos en cada tupla. Por lo tanto, los datasets que se encuentran en esa página web permiten obtener una información acertada del tema de interés, lo que permite un análisis exhaustivo del dataset, con el objetivo de obtener información, conocimiento y sabiduría lo cual es muy valioso en el momento de tomar una decisión. Esta data principalmente requiere pasar por un proceso ETL(Extract, Transform and Load) para que los datos lleguen totalmente limpios y posteriormente se aplica un proceso de minería de datos que se llevará a cabo con RapidMiner.

RapidMinder posee un sinnúmero de herramientas que se pueden utilizar para la minería de datos para poder sacar el máximo provecho de la información que se obtenga de un dataset, se utilizará el RapidMiner Studio el mismo que asegura "la experiencia integral de ciencia de datos desde la preparación de datos hasta la implementación del modelo" [3]. Respecto al almacenamiento de la data se utilizará MongoDB, MySQL y archivos de extensión .csv

En primera instancia se debe aplicar una matriz de correlación para definir los atributos más influyentes al costo de la vivienda. Con este conjunto definido de atributos y con la ayuda de la herramienta (RapidMiner), se aplica un proceso de Auto modelado que nos permitirá seleccionar los modelos más eficientes que serán recreados posteriormente de manera manual para su análisis de resultados. La metodología de cómo se realizará se muestra a continuación:

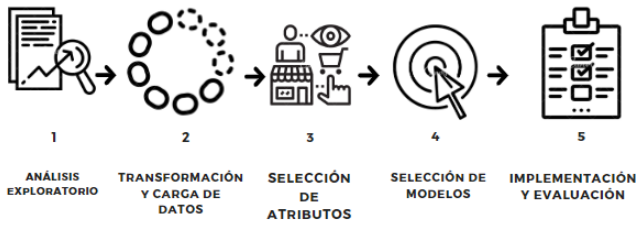


Fig. 1. Metodología de trabajo.

A. *Análisis exploratorio e ingesta de datos.*

La información que se presenta en los documentos de Kaggle se puede resumir de la siguiente manera:

Tabla I

Fuente de datos	
Archivo	Detalles
train.csv	Set de entrenamiento
test.csv	Set de pruebas
sample_submission.csv	Archivo de comparación

En el archivo "train" tenemos 80 atributos que describen las características de las viviendas que nos servirán para entrenar los diferentes modelos. El archivo "test" tiene 79 atributos excluyendo el "SalePrice" y servirá para comprobar la eficiencia de los modelos. Por último, el archivo "simple_submission" tiene solo dos atributos, el de identificación para las tuplas del archivo "test" y el "SalePrice" que corresponde a las mismas para posteriormente verificar la eficiencia del modelo a usar.

El primer paso que debemos realizar es cargar los datasets a nuestro sistema.

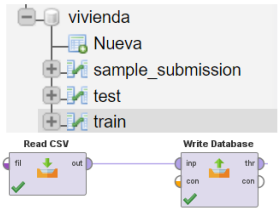


Fig. 2. Ingesta de datos.

B. *Transformación y carga de datos.*

Debemos realizar una limpieza de valores faltantes en los datasets y luego cargamos los datos.

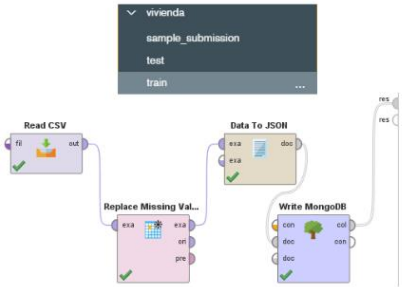


Fig. 3. Carga de datos.

C. *Selección de atributos influyentes.*

Para la selección de los principales atributos que influyen en el costo del bien inmueble se aplica una matriz de correlación sobre la data de entrenamiento del DWH, obteniendo el siguiente resultado:

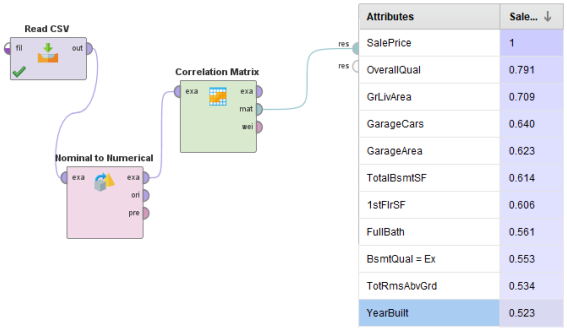


Fig. 4. Selección de atributos.

Seleccionamos los 10 atributos con más correlación con el atributo "SalePrice"

D. *Auto modelado y selección de los modelos más eficientes.*

Para poder seleccionar los modelos de predicción más eficientes podemos utilizar la herramienta de automodelado de RapidMiner. Podemos cargar datos temporales del archivo "train" y seleccionamos el atributo "SalePrice" como variable objetivo para predecir con los modelos.

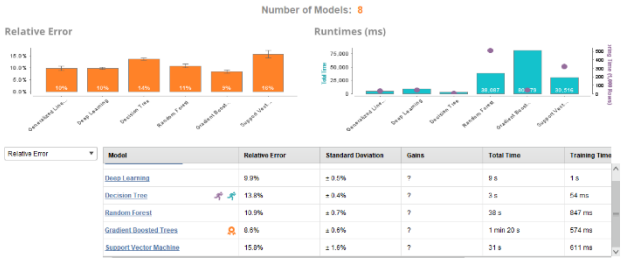


Fig. 5. Modelos arrojados con Auto Model.

Tabla 2

MODELOS OBTENIDOS CON AUTO MODELADO

Model	Total Time	Relative Error	Standard Deviation
Generalized Linear Model	5 s	9,9%	0,9%
Deep Learning	9 s	9,9%	0,5%
Decision Tree	3s	13,80%	0,004
Random Forest	38 s	10,9%	0,7%
Gradient Boosted Trees	1min 20 s	8,6%	0,6%
Support Vector Machine	31 s	15,8%	1,6%

Seleccionamos los modelos en base al menor tiempo de ejecución, por lo que escogimos los modelos: "Decision Tree", "Deep Learning" y "Generalized Linear Model".

E. *Implementación, evaluación de resultado y análisis comparativo de los modelos.*

Implementamos el modelo "Decision Tree" y tenemos el siguiente resultado:

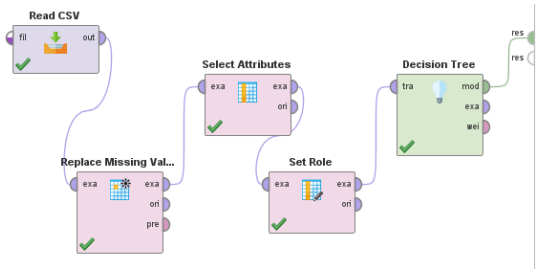


Fig. 6. Implementación del modelo Decision Tree (proceso).

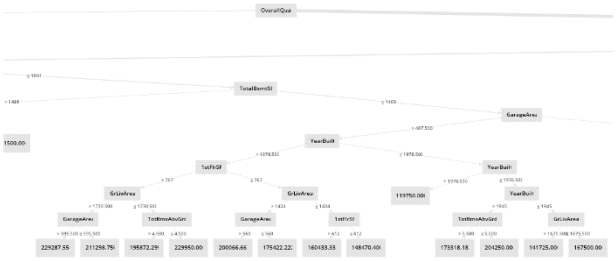
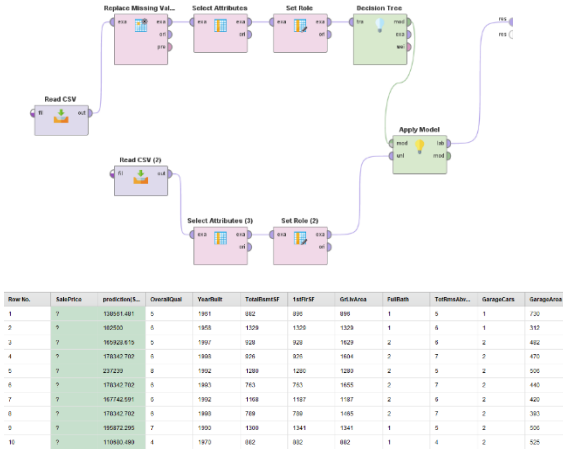


Fig. 7. Implementación del modelo Decision Tree.



Implementamos el modelo “Deep Learning” y se genera los siguientes resultados:

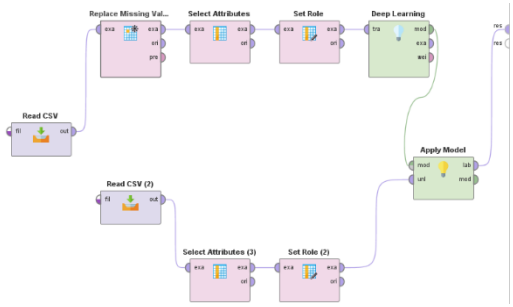


Fig. 8. Implementación del modelo Deep Learning.

Row No.	SalePrice	prediction(S...	OverallQual	YearBuilt	TotalBsmtSF	1stFlrSF	GrLivArea	FullBath	TotRmsAbv...	GarageCars	GarageArea
1	?	145740.822	5	1961	882	896	896	1	5	1	730
2	?	160590.192	6	1958	1329	1329	1329	1	6	1	312
3	?	183131.233	5	1997	928	928	1629	2	6	2	482
4	?	183486.043	6	1998	926	926	1604	2	7	2	470
5	?	209807.426	6	1992	1280	1280	1280	2	5	2	506
6	?	175702.423	6	1993	793	793	1655	2	7	2	440
7	?	169314.159	6	1992	1168	1187	1187	2	6	2	420
8	?	169883.132	6	1998	789	789	1465	2	7	2	393
9	?	209050.610	7	1990	1300	1341	1341	1	5	2	506
10	?	130323.727	4	1970	882	882	882	1	4	2	525

Fig. 9. Implementación del modelo Deep Learning.

Implementamos el modelo “Generalized Linear Model” y se genera los siguientes resultados:

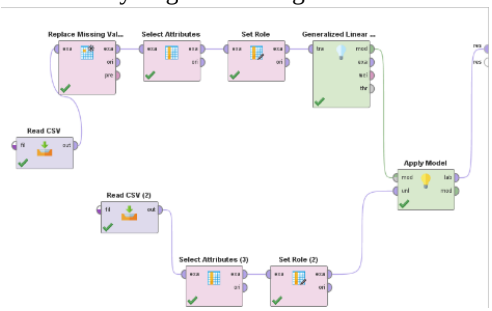


Fig. 10. Implementación del modelo Generalized Linear Model.

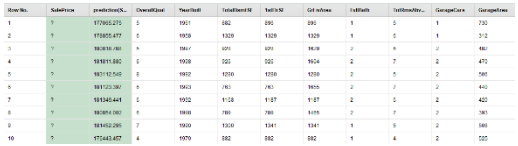


Fig. 11. Implementación del modelo Generalized Linear Model.

Se observa que en los 3 modelos realizados poseen diferentes resultados, ya que cada uno de ellos posee algoritmos diferentes de trabajo, en cada uno se ve un ligero aumento en el campo “SalesPrice”. Los modelos se utilizan en base a los resultados obtenidos en el auto modelado del dataset train.csv, los modelos se escogen en base enfocándose en el tiempo de respuesta ya que el error es relativo, así esperando obtener la mayor eficiencia y la certeza de los datos arrojados por cada modelo.

III. CONCLUSIONES

Todos los datasets que utilizamos en los modelos de predicción tienen que ser limpiados para así contar con datos fiables que aporten de manera correcta al proceso y análisis de resultados.

Dentro de la matriz de correlación podemos observar que hay pocas variables que influyen directamente al atributo “SalePrice” por lo que seleccionamos solo las que tienen mayor influencia en este atributo para posteriormente aplicarlos el modelado.

No se pueden utilizar otros modelos que no salgan dentro del auto modelado realizado previamente, ya que los modelos se arrojan dependiendo del dataset en el que se trabaja.

Cada modelo posee su propia característica única por lo que ya es decisión personal en que modelo basarse, pero para esto se debe tomar en cuenta los datos de un dataset.

IV. REFERENCIAS

[1] "¿Por qué es importante tener casa propia? - Urbanova - Desarrollos Urbanos", Urbanova - Desarrollos Urbanos, 2020. [En línea]. Disponible: https://www.urbanova.bo/por-que-es-importante-tener-casa-propia/#%C2%BFPor_que_comprar_una_casa.

[2] "House Prices: Advanced Regression Techniques | Kaggle", *Kaggle.com*, 2020. [En línea]. Disponible: https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

[3] "Get Started with RapidMiner | RapidMiner", *RapidMiner*, 2020. [En línea]. Disponible: https://rapidminer.com/get-started/.