Universidad Pontificia Bolivariana

Fundada en 1936

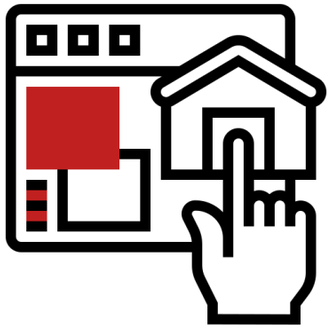
Modelo de predicción de precios de viviendas en el municipio de Rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz

Yuri Vanesa Grajales Alzate

Director: José Ricardo Zapata González

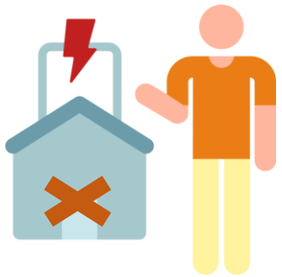
Problema

Incremento en el número de viviendas en el municipio de Rionegro



Información de precios de vivienda en varias páginas web.

¿Cuál es la mejor opción?



Decisiones Incorrectas

Contexto

La construcción en el Oriente Antioqueño.
Copyright 2017 por Oriente Comercial Digital

Oriente Antioqueño
Copyright 2016 por Cámara de Comercio Oriente Antioqueño

Alta valorización en el oriente antioqueño.

Municipios con mas proyección: Rionegro, Guarne, El Retiro.

En la próxima década muchas familias se asentarán allí.

Justificación

Invertir es una decisión sería que debe ser fruto de un proceso de meditación

Antes de realizar una inversión es importante hacer un estudio del mercado sobre donde se quiere destinar los fondos.

La tendencia de los precios de vivienda en el sector.

Ahorro en tiempo y esfuerzo, mejores decisiones

Rionegro no cuenta con un modelo o sistema que permita determinar precios de vivienda.

Marco conceptual

Universidad Pontificia Bolivariana
Fundada en 1936

Webscraping

- Acceder y extraer automáticamente grandes cantidades de información de un sitio web

CRISP-DM

- Metodología proyectos de minería de datos.
- 6 etapas

Algoritmos de machine learning

- Aprendizaje supervisado, no supervisado y por refuerzo.
- Clasificación y regresión.

Marco legal

Ley 527 de 1999

Uso de los mensajes de datos, del comercio electrónico y de las firmas digitales

La ley 1712 de 2014

Ley de transparencia y del derecho a la información pública nacional

Estado del arte

Universidad Pontificia Bolivariana
Fundada en 1936

- Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data (2015)
- Housing Price Prediction Using Neural Networks (2016)
- New Insights into rental housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listing (2017).
- Spatiotemporal Analysis of Housing Prices in China: A Big Data Perspective (2017)
- Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java (2018)

Estado del arte

Universidad Pontificia Bolivariana
Fundada en 1936

- Predicting House Price with a Memristor-Based Artificial Neural Network (2018)
- Exploiting filtering approach with web scrapping for smart online shopping: PPenny Wise: A wise tool for online shopping (2019)
- Web Scraping Crawling-based Automatic Data Augmentation for Deep Neural Networks-based Vehicle Classifications (2019)

Objetivo General

Universidad Pontificia Bolivariana
Fundada en 1936

Crear un modelo de datos para la predicción de precios de las viviendas en el municipio de Rionegro usando aprendizaje de maquina (machine learning), mediante la extracción de información disponible en páginas web para la toma de decisiones informada de compra y venta de propiedad raíz.

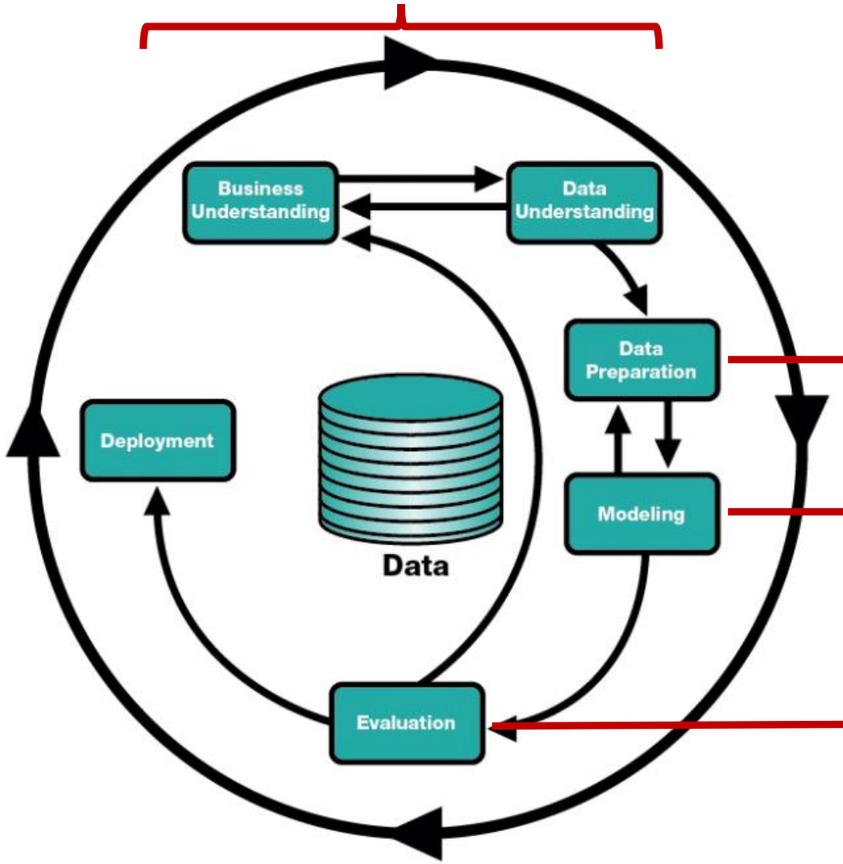
Objetivos Específicos

Universidad Pontificia Bolivariana
Fundada en 1936

- Consolidar la información relacionada con los precios de vivienda (casas y apartamentos) de páginas web previamente establecidas en el alcance del municipio de Rionegro.
- Realizar un análisis descriptivo de los datos de las viviendas que están en venta, de acuerdo con sus condiciones habitacionales y el precio, utilizando herramientas estadísticas y de análisis de datos.
- Generar modelos de aprendizaje de maquina (machine learning) que permitan predecir la relación entre el perfil de las viviendas y el precio de estas.
- Determinar el modelo que más se ajusta al comportamiento de los datos analizados.

Metodología

Consolidación de información



Análisis descriptivo

Modelamiento de los datos

Evaluación de los datos

Imagen 1. Metodología Crisp - DM para desarrollo de proyectos de minería de datos.
Fuente: DaimlerChrysler - SPSS - NCR (1996)

Análisis de Resultados
Consolidación de información

Universidad Pontificia Bolivariana
Fundada en 1936

fincaraíz.com.co

mercado Libre.com

2315 registros

38% duplicados exactos

Análisis de Resultados
Consolidación de información

0	https://www	4617440	Apartament	$ 355.000.00	73.71	3	2	1	apriv68,00 a	DescripciÃƒÂ	1	a Rionegro San Antonio d	68,00
1	https://www	4630318	Casa en Ven	$ 345.000.00	154	4	4	1	apriv154,00	DescripciÃƒÂ	2	gro El Porvenir	154,00
2	https://www	2076556	FOREST APA	Desde $ 333	65	2		2		DescripciÃƒÂ	1	Sector RÃƒÂ-o Golf - Rionegro Calle 40E	
3	https://www	4640678	Finca en Ver	$ 2.300.000.(	14720	3	2	NaN	area 14,72 H	DescripciÃƒÂ	3	egro Vereda	14,72
4	https://www	4624799	Apartament	$ 173.000.00	58.3	2	2	1	apriv58,30 a	DescripciÃƒÂ	1	a Rionegro Horizontes de	58,30
6	https://www	4918112	Casa en Ven	$ 780.000.00	220	3	4	NaN	apriv310,00	DescripciÃƒÂ	2	gro porvenir	310,00
7	https://www	3698025	Ventus Apar	Desde $ 330	66.29	2		1		DescripciÃƒÂ	1	Balcones II - Rionegro Calle 41 48AA-15	
8	https://www	4727145	Apartament	$ 190.000.00	57	3	2	1	aconst57,00	DescripciÃƒÂ	1	a Rionegro fontibon	

Tabla 1. Datos obtenidos de vivienda

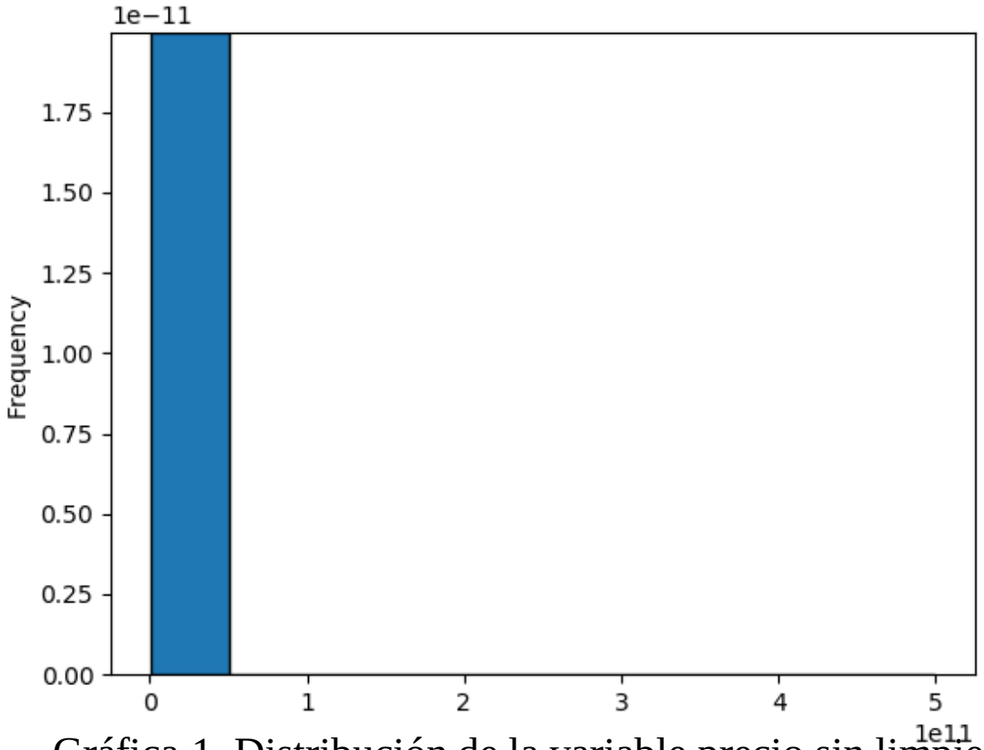
garage	boxcube
1	apriv68,00 aconst73,71 preciom2: 4.816.171/admon$180,000 Estrato: 5 Estado: Excelente anti 1 a 8 aÃ±os Piso No: 10Âº Sector: Ver Map
1	apriv154,00 aconst154,00 preciom2: 2.240.260/Estrato: 3 Estado: Bueno anti 16 a 30 aÃ±os Piso No: 3Âº Sector: Ver Mapa
2	

Tabla2. Datos contenidos en una misma columna.

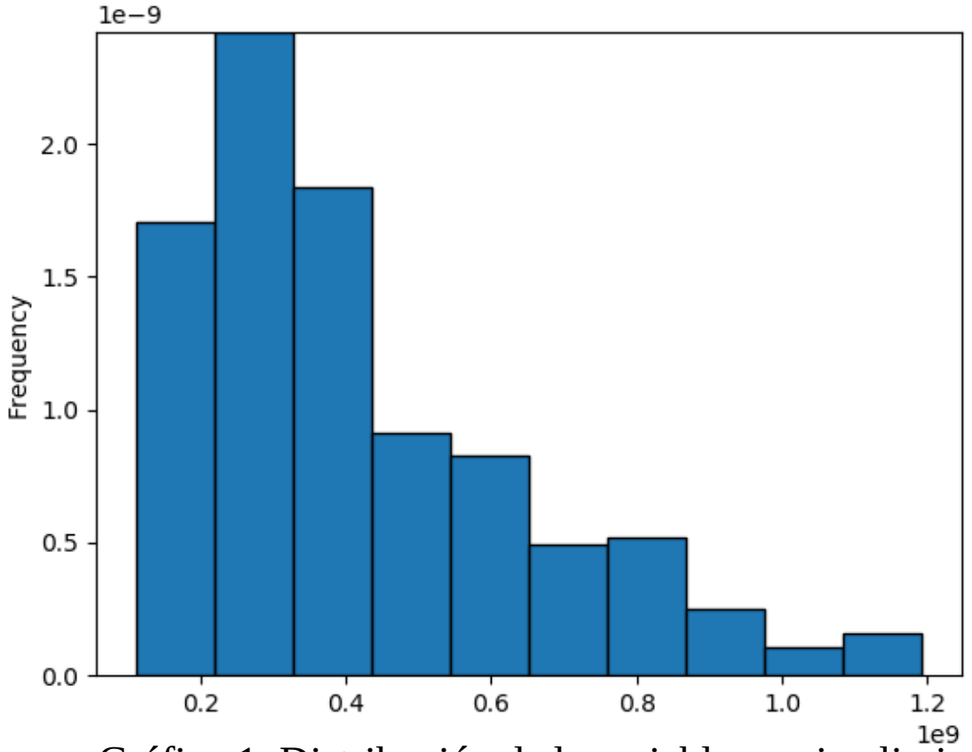
	apriva	aconst	preciom2	admon	estrato	esta
0	68.00000	73.71000	4816171.00000	180000.00000	5.00000	Excelent
1	154.00000	154.00000	2240260.00000	nan	3.00000	Bueno
3	nan	nan	15625.00000	nan	3.00000	Excelent

Tabla3. Base de datos final

Análisis de Resultados
Consolidación de información



Gráfica 1. Distribución de la variable precio sin limpieza



Gráfica 1. Distribución de la variable precio eliminando atípicos

Análisis de Resultados

Análisis descriptivo

barrio	0	barrio	0.000000
price	0	price	0.000000
area	0	area	0.000000
room	21	room	1.397206
bath	35	bath	2.328676
garage	334	garage	22.222222
description	0	description	0.000000
tipo_vivienda	0	tipo_vivienda	0.000000
barrio_or	0	barrio_or	0.000000
apriva	482	apriva	32.069195
aconst	332	aconst	22.089155
preciom2	57	preciom2	3.792415
admon	1160	admon	77.178975
estrato	276	estrato	18.363273
esta	770	esta	51.230872
antg	298	antg	19.827013
otro1	863	otro1	57.418496

Tabla 4. Cantidad y porcentaje de datos nulos

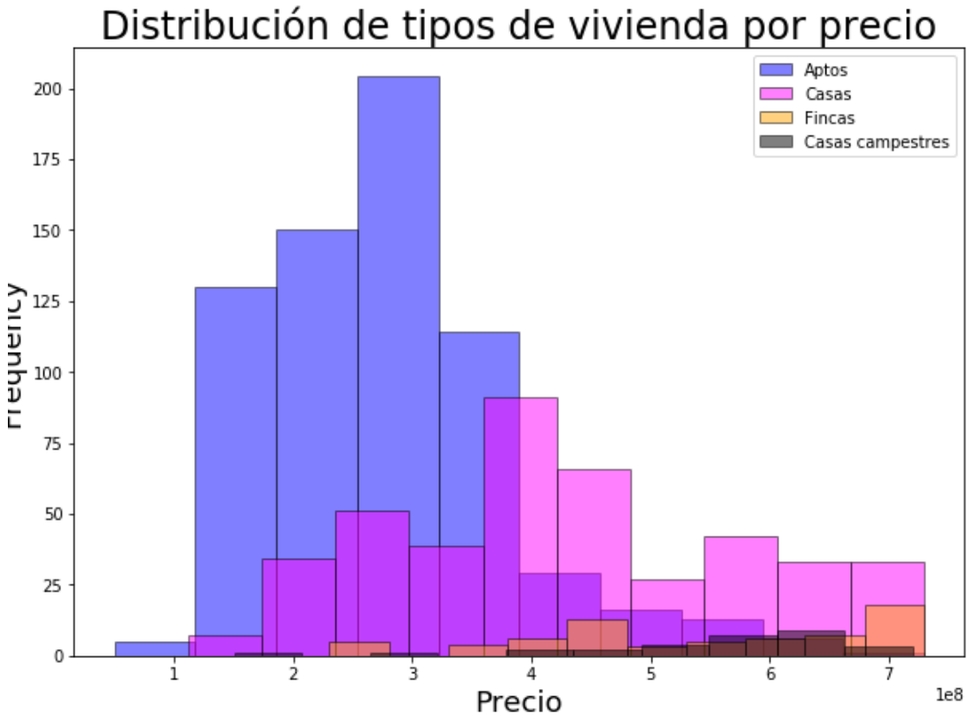
	price	area	room	bath	garage	tipo_vivienda	apriva	aconst	preciom2	admon
count	1116	1116	1103	1094	862	1116	764	976	1081	250
mean	420522629	4486.67922	2.99909338	2.54753199	1.35150812	1.67114695	423.183639	4795.7753	3337149.2	224973.864
std	230854418	134699.857	0.99499614	0.95124916	0.8864007	0.80598152	1492.745	144036.6	1497785.37	381942.424
min	110000000	41	1	1	1	1	13	41	133	120
25%	251500000	66	2	2	1	1	64	64	2571429	120000
50%	347500000	104.5	3	2	1	1	87	87.5	3307087	166420
75%	550000000	175	3	3	1	2	160	151	4135802	222250
max	1193465184	4500000	8	8	10	4	18000	4500000	17400000	5750000

Tabla 5. Análisis Descriptivo general de los datos

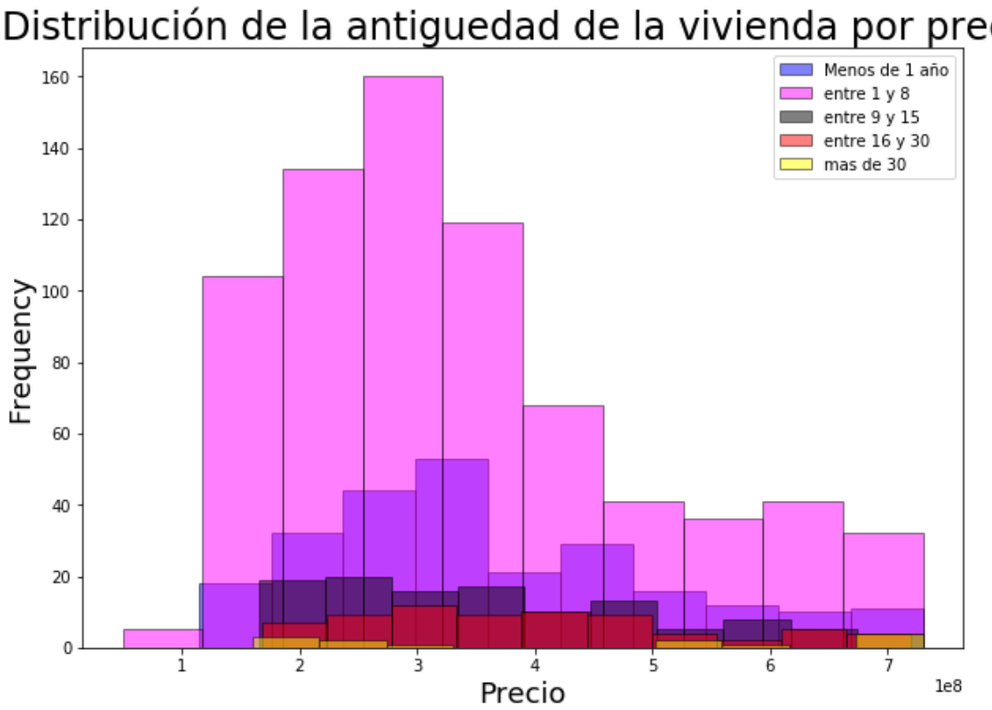
	estrato	esta	antg	piso
count	1220.0	727	1199	637.0
unique	4.0	3	5	16.0
top	4.0	Excelente	1a8	1.0
freq	548.0	537	587	289.0

Tabla 6. Análisis variables categóricos

Análisis de Resultados
Análisis descriptivo

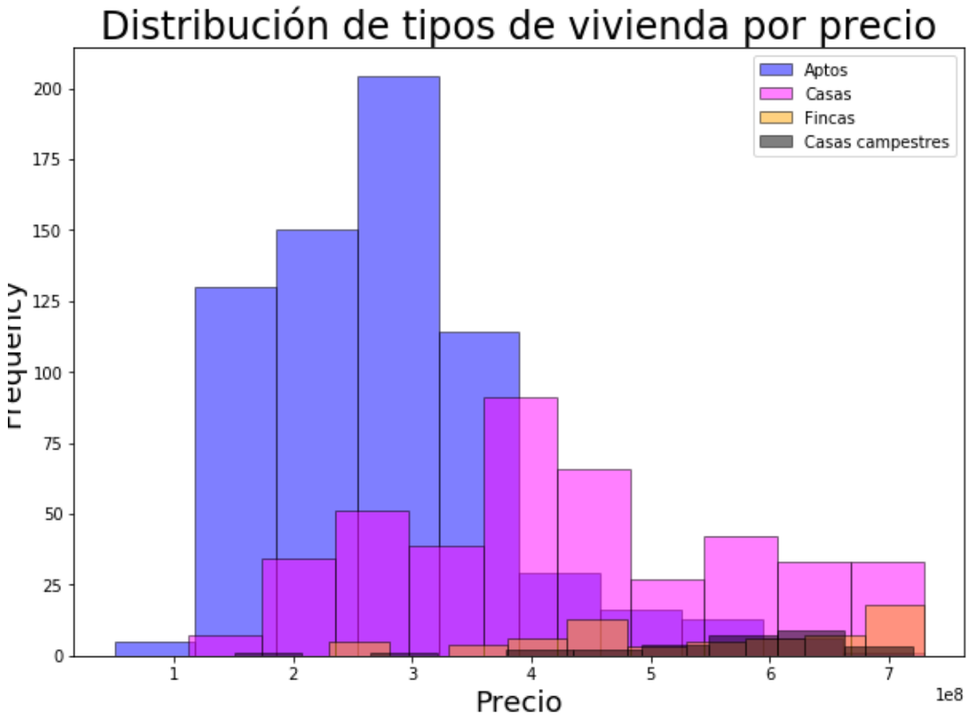


Gráfica 2. Distribución de tipos de vivienda por precio

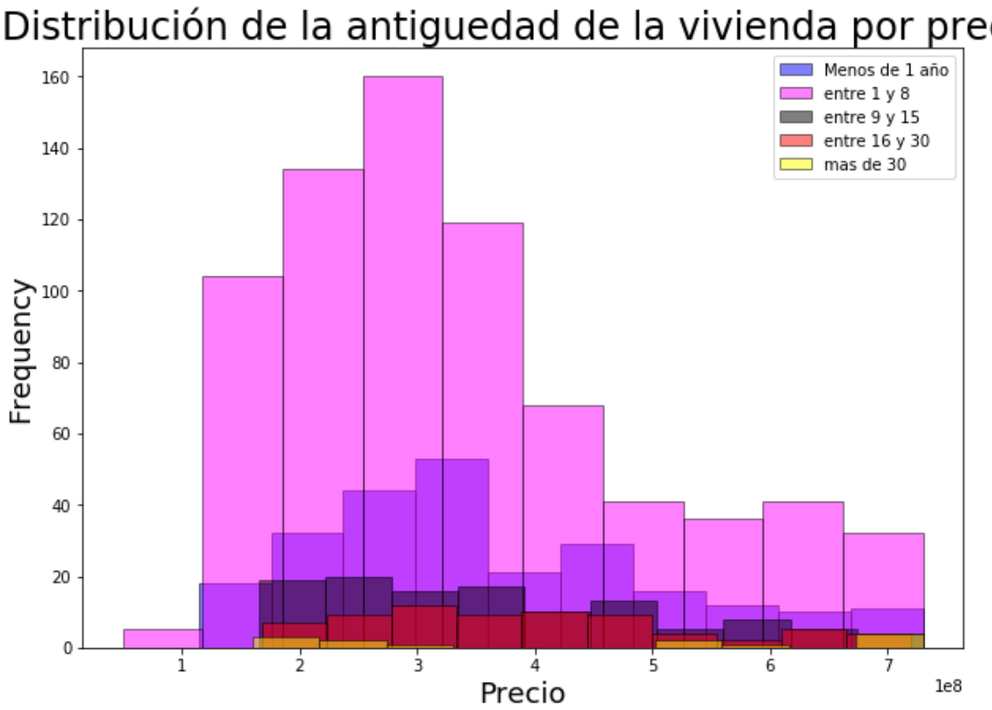


Gráfica 3. Distribución de antigüedad por precio

Análisis de Resultados
Análisis descriptivo



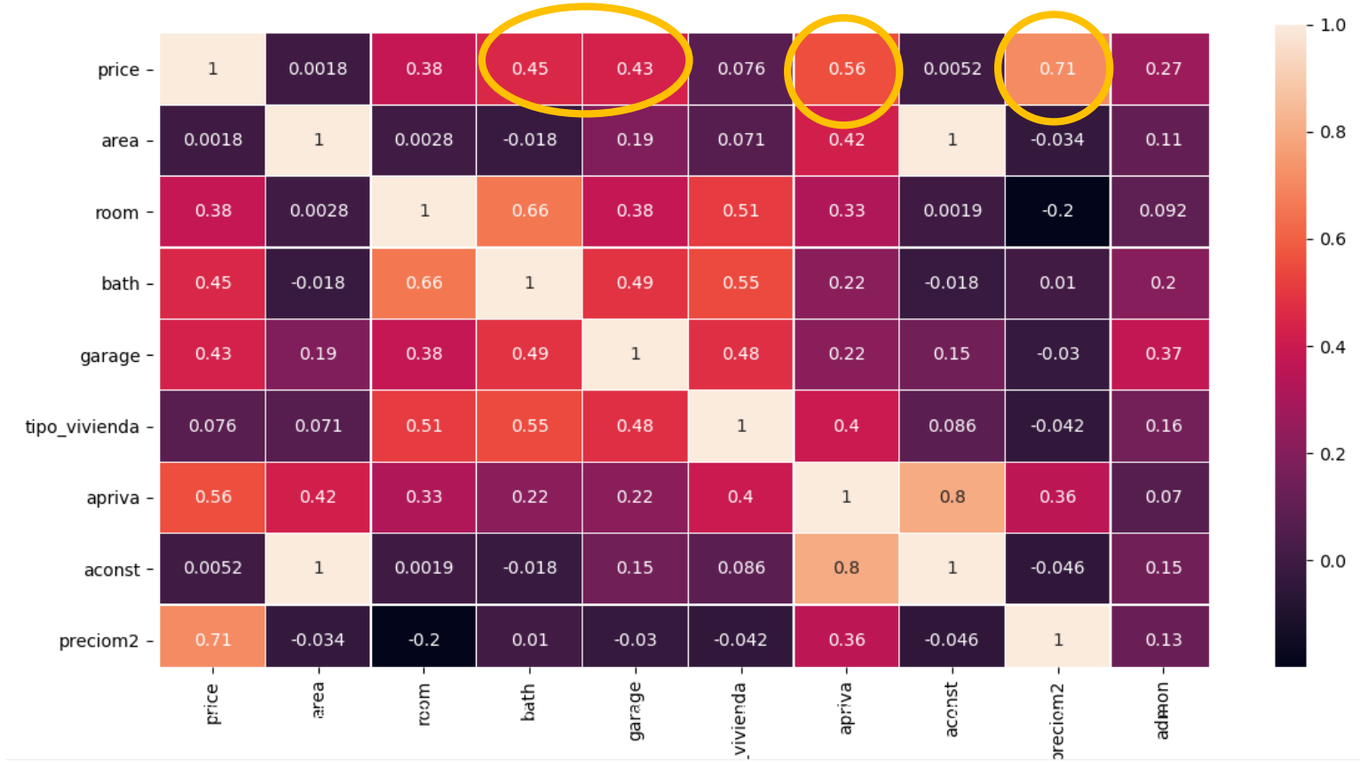
Gráfica 4. Distribución de tipos de vivienda por precio



Gráfica 5. Distribución de antigüedad por precio

Análisis de Resultados

Análisis descriptivo



Gráfica 6. Correlaciones

Análisis de Resultados
Modelos de aprendizaje automático

Universidad Pontificia Bolivariana
Fundada en 1936

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)
print(X_train.shape)
print(y_test.shape)
```

	R2	RMSE	MAE	MAPE	SMAPE
Regresión Lineal	0.530519	9.658456e+07	7.552920e+07	0.249191	0.109879
Decision Tree	0.551575	9.439382e+07	6.075658e+07	0.176933	0.088676
Random Forest	0.702875	7.683654e+07	5.148748e+07	0.152312	0.074897
Gradient Bosting Machine	0.710440	7.585210e+07	5.510707e+07	0.167122	0.080208
SVM	0.034413	1.385142e+08	1.064484e+08	0.339682	0.154847

Tabla 6. Resultados modelos de machine learning

```
Cross-validation scores: [0.82007193 0.78296132 0.70891265 0.59836476 0.80252586 0.67771331
 0.80998721 0.65028433 0.75746135 0.79569438]
Average cross-validation score: 0.74
```

Gráfica 3. Resultados Cross Validation modelo random forest

```
Cross-validation scores: [0.80227522 0.77194485 0.73888581 0.63177276 0.80076383 0.72844551
 0.82625837 0.67745147 0.72407652 0.75014975]
Average cross-validation score: 0.75
```

Gráfica 4. Resultados Cross Validation modelo gradient boosting

Análisis de Resultados

Evaluación modelos de aprendizaje automático

Universidad Pontificia Bolivariana
Fundada en 1936

```
area = 0.28351808959154357
room = 0.075425042016236 3
bath = 0.053274324785183784
garage = 0.05759291486796008
apriva = 0.14443244916229012
aconst = 0.09995553351300661
barrio = 0.04744573620167542
tipo_vivienda = 0.05011764758119632
estrato = 0.08796116295238988
esta = 0.01963514287996115
antg = 0.045821752809432154
piso = 0.034820203639124525
```

Tabla 7. Importancia de variables para gradient boosting

```
area = 0.31507286932359135
room = 0.024382082377121084
bath = 0.09345710633235858
garage = 0.05203410820979993
apriva = 0.07871204181558535
aconst = 0.205174445677024
barrio = 0.016361500687163695
tipo_vivienda = 0.10188214040403817
estrato = 0.08404468923920619
esta = 0.002771008932015859
antg = 0.015703808566301918
piso = 0.010404198435794484
```

Tabla 17. Importancia de variables para random forest

CONCLUSIONES

- Mediante la extracción de información de páginas web, se logró evaluar un modelo de gradient boosting que pudiese predecir el precio de las viviendas en el municipio de Rionegro.

- Se logró consolidar una base de datos estructurada para el posterior análisis y construcción de los modelos de machine learnig a través de la información extraída de las páginas web de mercado libre y finca raíz.

- Los modelos que mejor desempeño presentaron fueron gradient bosting machine y random forest, ya que arrojaron R2 de 0.75 y 0.77 respectivamente. Las variables que más influyen son el área de la vivienda, el área construida, el área privada, el tipo de vivienda, el estrato y el número de baños.

TRABAJOS FUTUROS

Universidad Pontificia Bolivariana
Fundada en 1936

- Abarcar más páginas web de manera que se pueda tener un cubrimiento más amplio de las viviendas que están a la venta en el municipio de Rionegro.

- Tener un periodo más prolongado de recolección de los datos, de manera que se pueda evaluar cómo van cambiando los precios a través del tiempo y si es que verdaderamente el municipio se está valorizando o es solo una *"burbuja"* o comportamiento atípico

- Tener en cuenta otros municipios aledaños para observar los cambios en el comportamiento entre municipios y así tener más información para tomar decisiones de inversión.

Referencias

- Alisenda Inmobiliaria. (01 de Enero de 2016). *El proceso de compra de una vivienda explicado paso a paso.* Obtenido de https://www.alisedainmobiliaria.com/blog/el-proceso-de-compra-de-una-vivienda-explicado-paso-a-paso/
- Argos. (20 de 04 de 2017). *El Oriente Antioqueo es el futuro Medellín.* Obtenido de http://grandesrealidades.argos.co/oriente-antioqueno-futuro-medellin/
- Boeing, G., & Waddell, P. (2017). New Insights into rental housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listing. *Journal of planning education and research*.
- Camara de Comercio del Oriente Antioqueño (CCOA). (2018). *El presente y futuro de la construcción se edifica en el oriente.* Recuperado el 05 de 05 de 2019, de http://www.orientecomercialdigital.com/sitio/noticias_detalle.php?id=588
- Camara de Comercio del Oriente Antioqueño. (05 de 04 de 2018). *Concepto económico del oriente antioqueño.* Obtenido de https://www.ccoa.org.co/Portals/0/Biblioteca%20virtual/Publicaciones%20regionales/2018/Concepto%20econ%C3%B3mico%202018.pdf?ver=2019-02-01-105326-537
- Cardona, D. F., González, J. L., Rivera, M. L., & Cárdenas, E. H. (2013). Aplicación de la regresión lineal en un problema de pobreza. *Revista interaccion, 12*, 73-84. Obtenido de http://www.unilibre.edu.co/revistainteraccion/volumen12/art4.pdf
- DaimlerChrysler - SPSS - NCR. (1996). *CRISP-DM.* Recuperado el 10 de 05 de 2019, de http://crisp-dm.eu/home/about-crisp-dm/
- Deloitte. (2016). *Bienes Raíces y Transacciones de Inversión Inmobiliaria.* Recuperado el 05 de 05 de 2019, de https://www2.deloitte.com/content/dam/Deloitte/mx/Documents/bienes-raices/Bienes-Raices-Folleto-2016.pdf
- Febrita, R., Alfiyatin, E., Taufiq, A., & Mahmudy, W. (2018). Data-driven fuzzy rule extraction for housing price prediction in Malang, East Java. *International Conference on Advanced Computer Science and Information Systems*, 351-358.

Referencias

- Galvis, R., & Carrillo, B. (2013). Indice de precios espacial para la vivienda urbana en Colombia:. *Revista de Economı´ıa del Rosario, 16*(1), 25-59.
- Hunger, M., & Lyon, W. (s.f.). *Analyzing the Panama Papers with Neo4j: Data Models, Queries & More.* Obtenido de https://neo4j.com/blog/analyzing-panama-papers-neo4j/
- Instituto Geográfico Agustín Codazi. (23 de Mayo de 2017). *¿Cuál es la diferencia entre un avalúo catastral y uno comercial?* Recuperado el 06 de Junio de 2019, de https://www.igac.gov.co/es/contenido/cual-es-la-diferencia-entre-un-avaluo-catastral-y-uno-comercial
- Kay, M. (2008). *XSLT 2.0 and XPath 2.0 Programmer's Reference.* Indianapolis.
- Kho, J. (26 de Septiembre de 2018). *How to Web Scrape with Python in 4 Minutes.* Recuperado el 03 de 05 de 2019, de https://towardsdatascience.com/how-to-web-scrape-with-python-in-4-minutes-bc49186a8460
- Lee, Y., & Kang, S.-J. (2019). Web Scraping Crawling-based Automatic Data Augmentation for Deep Neural Networks-based Vehicle Classifications. *2019 IEEE International Conference on Consumer Electronics, ICCE 2019.* Las Vegas.
- Lewis, H. (19 de Marzo de 2019). *12 First-Time Home Buyer Mistakes and How to Avoid Them.* Obtenido de https://www.nerdwallet.com/blog/mortgages/first-time-home-buyer-mistakes-that-are-easy-to-avoid/
- Li, S., Ye, X., Lee, J., Gong, J., & Qin, C. (2017). Spatiotemporal Analysis of Housing Prices in China: A Big Data Perspective. *Applied Spatial Analysis and Policy*, 421-433.
- Lim, W., Wang, L., Wang, Y., & Chang, Q. (2016). Housing price prediction using neural networks. *12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery.* Changsha; China.
- López de Mesa, A. M. (29 de Abril de 2019). Así se pinta el desarrollo urbanístico en Antioquia. *El colombiano*.

Referencias

- Mehak, S., Zafar, R., Aslam, S., & Bhatti, S. (2019). Exploiting filtering approach with web scrapping for smart online shopping : PPenny Wise: A wise tool for online shopping. *2nd International Conference on Computing, Mathematics and Engineering Technologies*. Sukkur, Pakistan: iCoMET 2019.
- Metro Cuadrado. (11 de Enero de 2019). *¿Cómo definir el precio de venta de una vivienda?* Recuperado el 01 de Junio de 2019, de https://www.metrocuadrado.com/noticias/guia-de-compra-y-venta/como-definir-el-precio-de-venta-de-una-vivienda-252
- Murillo, D., & Saavedra, D. (2017). Web Scraping de los Perfiles y Publicaciones de una Afiliación en Google Scholar utilizando Aplicaciones Web e implementando un Algoritmo en R. *4to Congreso Internacional AmITIC 2017*. Popayan, COlombia.
- Park, B., & Bae, K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*, 2928-2934.
- Phan, T. (2019). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. *Proceedings - International Conference on Machine Learning and Data Engineering*. Sydney, Australia: iCMLDE 2018.
- Propiedades Oriente Raíz. (14 de Enero de 2019). *Valor del metro cuadrado en el Oriente Antioqueño 2019.* Recuperado el 1 de Junio de 2019, de https://www.orienteraiz.co/blog/valor-metro-cuadrado-oriente-antioqueno-2019/
- Rubio, J., Guzmán, F., & Otero, J. (2019). Una base de datos de precios y características de vivienda en Colombia con información de Internet. *Revista de Economía del Rosario*, 75-100.
- Van Rossum, G. (1989). Python. Paises Bajos.
- Villa, M. (28 de Abril de 2018). *¿Por qué es importante invertir tu dinero?* Recuperado el 01 de 05 de 2019, de http://forbes.es/business/42248/por-que-es-importante-invertir-tu-dinero/
- Wang, J., Hu, S., Zhan, X., Luo, Q., Yu, Q., Liu, Z., . . . Liu, Y. (2018). Predicting House Price with a Memristor-Based Artificial Neural Network. *IEEE Access*, 16523-16528.


¡Soy orgullosamente UPB! • Sede central Medellín