MÉTODO AUTOMÁTICO PARA LA PREDICCIÓN DEL AVALÚO COMERCIAL DE UN INMUEBLE EN LA CIUDAD DE BOGOTÁ

VIVIANNE JULIANA TÉLLEZ BUITRAGO

DUYBER NICOLÁS MARTÍNEZ SÁNCHEZ

Trabajo de Investigación Tecnológica para optar al título de Ingeniero e Ingeniera de sistemas y computación

Juan Carlos Barrero Calixto

Docente

UNIVERSIDAD CATÓLICA DE COLOMBIA

FACULTAD DE INGENIERÍA

PROGRAMA INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

BOGOTÁ D.C

2021

Nota de Aceptación

_____

_____

_____

_____

_____

Asesor

_____

Jurado 1

_____

Jurado 2

Bogotá, 17 de mayo de 2021

https://creativecommons.org/licenses/by-nc/4.0/deed.es

“Le quiero dedicar este trabajo y sus resultados a toda mi familia, desde mis tías y abuelas, que desde pequeña siempre me han cuidado, a mi hermanita para que siga viendo en mi un ejemplo a seguir y sobre todo a mis padres, que han dedicado toda su vida a que sus hijas siempre tengan lo mejor y no puedo estar más feliz de ser su hija, espero esta meta alcanzada los haga sentirse orgullosos y felices de la hija que tienen. Los amo con todo mi corazón”
Juliana

AGRADECIMIENTOS

Queremos agradecerle al profesor Juan Carlos Barrero, por su infinita colaboración a lo largo del proyecto, por su paciencia y por su sabiduría a la hora de guiarnos, muchas gracias por todo profe. A mí papá y arquitecto Jhonny Hernán Téllez Zamora, el cual nos ayudó a definir el rumbo del proyecto gracias su trabajo como perito avaluador y arquitecto, sus aportes de conocimiento en el área contribuyeron a que la realización de este fuera coherente con el campo en el que estábamos trabajando. También al profesor Alexander Aldana de humanidades, el cual nos ayudó con una de las tareas que suelen dificultarse más a los ingenieros, que es la redacción, muchas gracias por ayudarnos a pulir el trabajo profesor. Por último y no menos importante, al ingeniero Roger Guzmán, aunque ya no se encuentre en la universidad, sus aportes y conocimientos durante la electiva de Deep Learning fueron valiosos para el desarrollo del proyecto.

De mi parte, quiero agradecerle infinitamente a mi compañero de trabajo Nicolás Martínez, muchas gracias por apoyarme hasta el final en el desarrollo de este proyecto, no sé qué hubiera sido de el sin tus aportes y de tu increíble esfuerzo, no sé qué nos depare la vida, pero espero poder seguir compartiendo muchos proyectos a tu lado. Muchas gracias por todo. Por otra parte quiero agradecer a mi papá Alexander Martinez quien me hizo posible el estudio en esta universidad, agradezco a mi mamá Dora Nely Sanchez Espinoza quien me apoyó todo este tiempo universitario y en la vida académica, gracias por haberme facilitado este computador portátil con el cual fue posible el desarrollo de esta tesis, infinitos agradecimientos a mis padres por todo su apoyo afectivo y económico.

Tabla de contenido

LISTA DE FIGURAS

LISTA DE TABLAS

Resumen

En Bogotá habitan 7.5 millones de personas, que viven en aproximadamente 2 millones de inmuebles. El avalúo de los inmuebles es realizado por peritos, registrados en el Registro Nacional de Avaluadores, que se rigen por la resolución 620 del 2008 del IGAC. Aunque los avalúos siguen una metodología dada en esta resolución, siempre hay un nivel de subjetividad en la apreciación de las variables que se consideran para definir el precio de un inmueble. Adicionalmente es un proceso lento ya que el avaluador debe ir hasta el inmueble, investigar características de la zona, comparar precio con otros inmuebles y generar un informe. Así mismo, es un proceso con altos honorarios. Ante esto, el campo de los avalúos comerciales representa el lugar perfecto para desarrollar técnicas de aprendizaje de máquina, ya que son muchas las personas interesadas en conocer de forma inmediata el valor de sus viviendas en el mercado inmobiliario. Además, aportar a las mejoras tecnológicas en el mercado inmobiliario pueden impactar significativamente en la economía, ya que este mercado es de los que más aporta al PIB del país. Al mismo tiempo, el país está viviendo un proceso de transformación hacia la industria 4.0 que se enfoca en el uso de estas tecnologías para aportar en los distintos campos de la industria.

Por tal motivo, en este proyecto se realizó el diseño, desarrollo e implementación de un método automático mediante técnicas de Machine Learning con el cual se puede predecir el valor de inmuebles de vivienda en la ciudad Bogotá a través de la extracción de datos de publicaciones de venta de inmuebles en páginas web. Con la herramienta de web Scraping “Dexi”, se extrajo la información de la página web de *Finca Raíz*. Luego, se realizó un proceso de limpieza de datos con la ayuda de Python y Excel. Después, se implementaron las técnicas de: Árboles de decisión, regresión lineal, Random Forest y redes neuronales profundas. Por último, se calcularon las medidas de desempeño de Coeficiente de determinación (R^2), Error cuadrático medio (RMSE) y Coeficiente de Variación (CV) respecto a la varianza, para así seleccionar el mejor método.

Como resultado se obtuvo un conjunto de datos organizados con la información extraída de la página web de *Finca Raíz*. Sin embargo, por las condiciones de heterogeneidad de estos datos, se segmentó el Dataset por los estratos y se implementaron los algoritmos para cada uno de ellos. Así mismo, para Random Forest se obtuvo una media aritmética del coeficiente de determinación del 91% y un coeficiente de variación del 17%. Igualmente, para la regresión lineal se obtuvo una media aritmética del coeficiente de determinación del 77% y un coeficiente de variación de 33%. De la misma manera, para árboles de decisión se obtuvo una media aritmética del coeficiente de determinación del 82% y para el coeficiente de variación de 24% y, por último, para las redes neuronales profundas se obtuvo una

media aritmética del coeficiente de determinación del 43% y para el coeficiente de variación de 86%. En conclusión, para este proyecto el mejor modelo fue el de Random Forest, dado que sus medidas de desempeño fueron superiores respecto a los demás modelos.

Palabras clave: Inteligencia Artificial, Avalúo Comercial, Industria Inmobiliaria, Recopilación de datos, Procesamiento de datos.

Abstract

In Bogota dwells 7.5 million people, who live in 2 million of property approximately. The valuation of property is made by auditors, which are registered in the National Register of Auditors, which are governed by resolution 620 of 2008 from IGAC. Even if valuation follows a methodology given by already said resolution, there always exists a subjective level in the appreciation of the variables that are considered to define the price of a property. Furthermore, it is a slow process because the auditor must go to the property, to investigate the area's characteristics, to compare prices with other properties, and to make a report. In addition, it is a process with a high fee. The field of commercial appraisal represents the perfect place to develop techniques of Machine Learning because there are a lot of people interested to know immediately the price of their houses in the real estate industry. Moreover, contributing to the improvement of technology into the real estate industry can impact the economy, because this market is one which invests more in the PIB of Colombia significantly. In addition, the country is undergoing a transformation process around 4.0 Industry, which is focused on the use of this technology to support in the different fields of the industry.

For these reasons, in this project was made the design, development, and implementation of an automatic method with techniques of Machine Learning, which predict prices of property in Bogota through the extraction of data from sale publications property in web pages. With the tool DEXI of web scraping, information was extracted from the web page *Finca Raíz*. Then, with this data, was transformed and cleansed the process with Python and Excel as tools. After, was implemented the techniques of: Decision tree, linear regression, random forest, and deep neural network. For last, was calculated the following measurements: coefficient of determination (R2), root-mean-square error (RMSE), and coefficient of variation (CV) with respect to variance, for which these select the best method.

As a result, there is a dataset with the information extracted from the web page *Finca Raíz*. Due to the conditions of heterogeneity of data, the dataset was segmented by social stratum and the different algorithms were implemented in each one. For random forest, the mean of coefficient of determination was 91% and for the coefficient of variation 17%, for lineal regression, the mean of coefficient of determination was 77% and for the coefficient of variation 33%, for the decision tree, the mean of coefficient of determination was 82% and for the coefficient of variation 24%, and for last, for the deep neuronal network the mean of coefficient of determination was 43% and for the coefficient of variation 83%. For that reason, for this project, the best model was random forest given that performance measurements were better than the other models.

Keywords: Data collection, Real estate industry, Commercial Appraisal, Artificial Intelligence, Data processing.

1 INTRODUCCIÓN

Los avalúos comerciales de inmuebles de vivienda son de vital importancia ya que hacen parte de la industria inmobiliaria la cual tiene un gran impacto en la economía del país. Adicionalmente, los avalúos definen una referencia para los precios de compra-venta de un inmueble, siendo utilizados tanto por las personas naturales para realizar su compra o su venta, como por los bancos, para definir el valor a prestar en un crédito de vivienda, según el precio que se defina en el avalúo.

De acuerdo con investigaciones recientes, distintos autores como lo son Bin Ja, Gardiner en su estudio ”Multi-source urban data fusion for property value assessment: A case study in Philadelphia”[1] y Rotimi Boluwatife Abidoye,Albert P.C. Chan en su estudio “Predicting property price index using artificial intelligence techniques Evidence from Hong Kong”[2], se está en la búsqueda de algoritmos que faciliten la predicción del valor de un avalúo comercial, que se adapten a las necesidades y características de los lugares diferentes en donde se realicen. En años recientes se ha visto un aumento en estudios que buscan diseñar estos métodos, teniendo muy buenos resultados con la implementación de redes neuronales y técnicas de aprendizaje de máquina. Por ejemplo, se encuentran estudios realizados para algunas ciudades, como lo son Philadephia[3], Hong Kong[4], Los Ángeles[5], entre otras, que demuestran la importancia de implementar estos algoritmos.

En Colombia, el proceso para obtener el avalúo comercial de un inmueble suele ser realizado por un perito, proceso que involucra la visita al inmueble en cuestión, el análisis del sector, el análisis de la vivienda, entre otras actividades que pueden llegar a ser dispendiosas, costosas y demorar el proceso de obtención del valor de la vivienda, además, de estar sujetas a la subjetividad del evaluador, dado que es el perito quien da su juicio respecto al valor que debe de tener la vivienda.

Por esta razón, en el presente proyecto se propone la implementación de un método automático para la predicción del avalúo comercial de los inmuebles de vivienda en la ciudad de Bogotá, tomando la información de páginas web que se dedican a la publicación de venta de viviendas tal como la página web *Finca Raíz*, a través de técnicas de raspado web conocido en inglés como web scraping. Con los datos

[1] Bin, J.a, Gardiner, B.b, Li, E.c, Liu, Z.a, 2019, Multi-source urban data fusion for property value assessment: A case study in Philadelphia, ScienceDirect, p.16

[2]Rotimi Boluwatife Abidoye, Albert P.C. Chan and Funmilayo Adenike Abidoye, Olalekan Shamsideen Oshodi | 2018 | Predicting property price index using artificial intelligence techniques Evidence from Hong Kong | International Journal of Housing Markets and Analysis. P.16

[3] Bin, J.a, Gardiner, B.b, Li, E.c, Liu, Z.a, Op.Cit, p.16

[4] Rotimi Boluwatife Abidoye, Albert P.C. Chan and Funmilayo Adenike Abidoye, Olalekan Shamsideen Oshodi, Op.Cit. p.16

[5] Junchi Bin, Bryan Gardiner, Zheng Liu | 2019| Attention-based multi-modal fusion for improved real estate appraisal: a case study in Los Angeles. P.16

obtenidos, se realiza una limpieza y creación de un Dataset, para así realizar el diseño y desarrollo de distintos modelos de Machine Learning que podrán facilitar la predicción del valor del avalúo, utilizando medidas de desempeño de regresión como lo son el coeficiente de determinación (R^2) y la raíz del error cuadrático medio (RMSE) para determinar cuál de los modelos desarrollados es el mejor para la ciudad de Bogotá.

Dentro de los impactos analizados en este proyecto, se puede observar el impacto tecnológico que se genera, dada la necesidad del país por adentrarse más en los procesos propios de la industria 4.0, teniendo en cuenta la transformación en la industria gracias a todas las tecnologías ofrecidas dentro de la cuarta revolución industrial. También se pueden incluir la automatización de procesos y la implementación de inteligencia artificial para facilitar esta transformación, ya que los avalúos son un proceso perfecto para ser automatizados. Otro de los impactos de este proyecto apunta al campo socioeconómico, ya que se ayuda a las personas que necesitan vender sus inmuebles a definir el precio de una manera más sencilla y evitando gastos de contratación de peritaje en el avalúo comercial del inmueble. Por otra parte, dentro de los impactos ambientales del proyecto se puede observar una reducción de los gases que afectan a la capa de ozono, al no tener que desplazarse los peritos para realizar el avalúo, junto a una reducción del papel que éstos usan a la hora de entregar los mismos. Por último, dentro de los impactos socio-culturales, se enfatiza en el hecho de hacer que la industria inmobiliaria adopte como propias las tecnologías de la industria 4.0, cambiar la mentalidad de las personas hacia la tecnología buscando que vean en las herramientas un apoyo y complemento a sus trabajos diarios y no como algo que reemplacen su labor dentro de una empresa.

Finalmente, el presente trabajo se cataloga como trabajo de investigación según el acuerdo 265 de la Universidad Católica de Colombia[6], en donde se identifica una necesidad tecnológica y se da una propuesta de solución a través del método automático obtenido en el proyecto, como un avance para la solución de la misma.

[6] Acuerdo 265 del 2018, Consejo superior de la Universidad Católica de Colombia, “Por el cual se aprueban los lineamientos y las opciones de grado para los programas de la Facultad de ingeniería de la Universidad Católica de Colombia, 12 de diciembre del 2018, Consejo superior de la Universidad Católica de Colombia. P.17

2 DESCRIPCIÓN DEL PROBLEMA

La industria 4.0 o la cuarta revolución industrial consiste, según el documento del ministerio de las telecomunicaciones MinTIC del año 2019[7], en la digitalización de la industria y todos los servicios relacionados con una empresa, lo cual involucra el uso de nuevas tecnologías como lo son la implementación de inteligencia artificial en los procesos, el uso de la robótica a favor de las empresas o el famoso Internet de las cosas o conocido en inglés como Internet of Things (IoT). Esta industria está tomando cada vez más fuerza en Colombia, pues las diferentes áreas en el sector económico deben adaptarse a estas nuevas tecnologías e iniciar con la innovación para estar a la altura de la competencia a nivel mundial.

De esta manera, el sector inmobiliario es uno de los más influyentes en la economía del país, de hecho, para el primer trimestre del 2019 representó un 9.40% de participación dentro del PIB nacional[8]. Además, invertir en un inmueble puede generar grandes ganancias a largo plazo, dada a la valorización que tienen. Por ejemplo, en Bogotá, la valorización está en un rango promedio del 10 al 15 % por propiedad adquirida en un lapso de 5 años. Sin embargo, a pesar de su importancia, este sector se encuentra atrasado con respecto a la implementación de tecnologías innovadoras con relación a otras industrias. Es decir, se encuentra la necesidad de adaptarse a la industria 4.0.

Dentro de los sectores del área inmobiliaria que están empezando a involucrarse con tecnologías propias de la industria 4.0, sobresalen las áreas de los avalúos comerciales y la predicción de valores de los inmuebles en el país, principalmente en Bogotá. Actualmente ya existen diversas herramientas que utilizan técnicas de Big Data y Machine Learning, utilizados tanto por el consumidor normal como por los peritos avaluadores que trabajan con los bancos para la predicción de estos valores. Lo anterior se ejemplifica con la página web *Properati*[9], la cual utiliza *redes neuronales* para la predicción de estos avalúos utilizando ciertas características comunes de los inmuebles para predecir estos valores. A pesar que esta aplicación implementa una de las muchas técnicas de Machine Learning que existen, hay otros tipos de algoritmos como los usados por Deep Learning que pueden aportar mucho a esta área.

---

[7] MINISTERIO DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES, Aspectos básicos de la industria 4.0, República de Colombia, OFICINA ASESORA DE PLANEACIÓN Y ESTUDIOS SECTORIALES, 2019. P.18

[8]DIRECCIÓN DE MINERÍA EMPRESARIAL, Análisis del comportamiento del PIB minero primer trimestre de 2019, Bogotá D.C, 2019. P.18

[9] PROPERATI, Disponible en: (www.properati.com). p.18

A pesar de las innovaciones y de la existencia y aplicación de algunas herramientas tecnológicas, el proceso para realizar un avalúo comercial de un inmueble de vivienda en Bogotá, puede llegar a ser dispendioso. Esto debido a que, según la resolución 620 de 2008 del IGAC[10] por la cual se establecen los procedimientos para los avalúos ordenados dentro del marco de la Ley 388 de 1997[11], existen distintos métodos y factores al momento de realizar un avalúo comercial en el país que pueden alterar el precio de esa vivienda. Esto se suma a que el proceso de realizar un avalúo puede resultar subjetivo, puesto que la percepción que tiene cada persona es diferente al momento de realizar un avalúo. Por ejemplo, según el registro nacional de avaluadores por sus siglas RNA, en Bogotá existen, para el mes de agosto del 2019, 265 personas registradas y calificadas para realizar avalúos, aparte de los que están registrados en lonjas.

Además de la subjetividad de los avalúos es importante revisar algunas características propias de la ciudad y del inmueble que pueden afectar el valor asignado por el perito. En este sentido, vale recordar que diversos estudios han demostrado que la selección de las variables para realizar investigaciones en esta área es crucial para la implementación de técnicas que involucren Machine Learning y Deep Learning, pues los estudios tienden a variar según el lugar de su realización. Así, no es lo mismo estudiar el comportamiento inmobiliario de Philadelphia a realizarlo en un lugar como Hong Kong o Corea del sur, la selección de características para estos estudios puede variar en algunos aspectos, pero pueden compartir muchas cosas en común. INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI SEDE CENTRAL, RESOLUCIÓN 620 DE 2008

Cabe agregar que Bogotá es una ciudad con características demográficas y distribuciones de población variadas, por lo que uno de los factores más importantes para realizar el avalúo en Bogotá es el estrato de la vivienda, teniendo en cuenta que el Decreto Distrital 551 de 2019[12] es la que ordena la estratificación de los inmuebles residenciales. Bogotá tiene unas características de estratificación muy únicas, teniendo en una misma localidad diferentes estratos, y lo más impactante, es que estos estratos, incluso los más separados como el 1 y 6, se hallan uno junto al otro, fenómeno observable en múltiples barrios de las localidades de Suba y Usaquén. Sin embargo, existen otros sectores de la ciudad donde se puede ver una distribución por estratos más pareja, como es el caso de Los Mártires o Antonio

[10] INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI SEDE CENTRAL, RESOLUCIÓN 620 DE 2008, Bogotá, Secretaría Jurídica Distrital, 2008. P.19
[11] Ley 388 de 1997. Por la cual se modifica la Ley 9 de 1989, y la Ley 2 de 1991 y se dictan otras disposiciones. 18 de julio de 1997. D.O. No. 43091. P.19
[12] SECRETARIA DISTRITAL DE PLANEACIÓN, Decreto Distrital 551 de 2019, Bogotá, Secretaria Distrital de planeación, 2019. P.19

Nariño. En la siguiente figura se puede observar de mejor manera como es la distribución de los estratos en algunas localidades, tomando como ejemplo la localidad de Antonio Nariño como una localidad homogénea en estratos y la localidad de Usaquén como heterogénea en la distribución de los estratos.

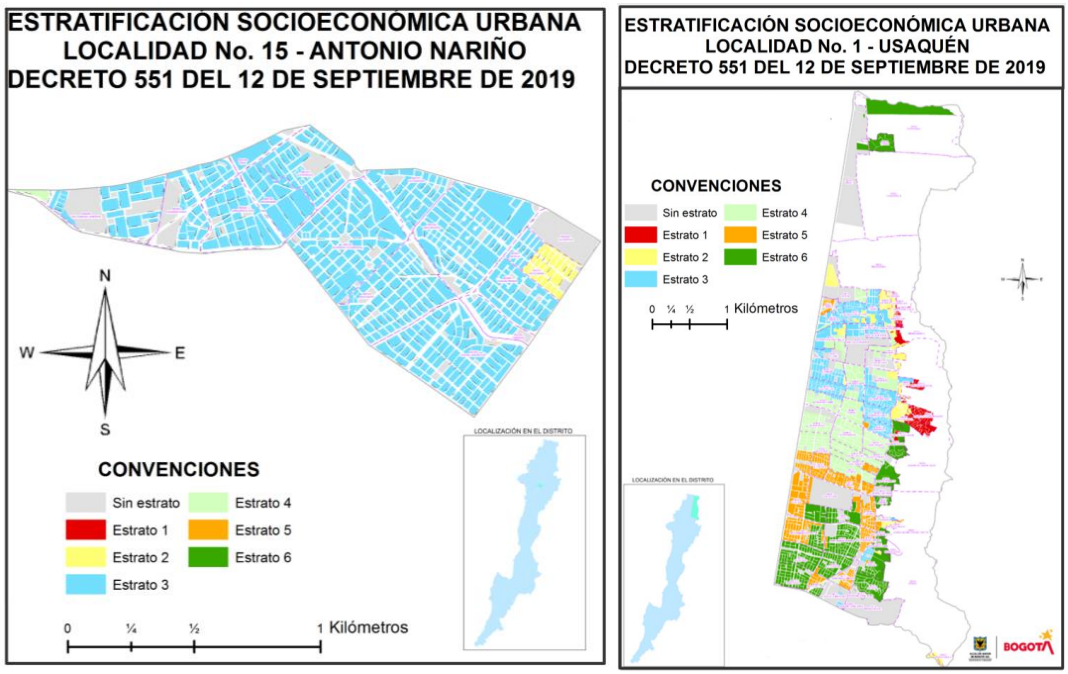

Figura 1, Comparativo entre localidades, Fuente: http://www.sdp.gov.co/gestion-estudios-estrategicos/estratificacion/estratificacion-por-localidad

También, se observa que el proceso de avalúo en Bogotá obedece a múltiples factores que en ocasiones pueden resultar engorrosos, confusos, subjetivos y que terminan perjudicando a los usuarios. Así pues, es menester desarrollar un modelo más objetivo, sencillo y directo para establecer un peritaje inmobiliario que se adapte a las demandas de los procesos de industria 4.0. En este sentido, se propone un método automático que implemente los algoritmos de Machine Learning dentro de la predicción de valores para avalúos comerciales de inmuebles de vivienda en Bogotá, con esto se busca la obtención de los mejores resultados de forma rápida y precisa, a la vez, se pretende el fácil acceso y la usabilidad de la información por parte de las personas. Además, se quiere reducir en los gastos de los avalúos, incluso, se propende por su realización gratuita, evitando la contratación de peritaje. Todas estas implementaciones están en concordancia con la transformación 4.0 que se está viviendo actualmente en el país.

3 PREGUNTA DE INVESTIGACIÓN

¿De qué manera se puede construir un método automático que a partir de técnicas de Machine Learning y la adquisición de datos de páginas web permita realizar la predicción del avalúo comercial de un inmueble residencial en la ciudad de Bogotá?

4 OBJETIVOS

4.1 Objetivo General

Implementar un método automático para la predicción de los avalúos comerciales de inmuebles en la ciudad de Bogotá a partir de técnicas de Machine Learning y de adquisición de datos de páginas web.

4.2 Objetivos específicos

- Construir el Dataset a través de adquisición datos de páginas web inmobiliarias de Bogotá para registrar las variables a utilizar en el entrenamiento de los algoritmos.
- Diseñar un método automático a partir de técnicas de Machine Learning para predecir el valor de los avalúos comerciales de un inmueble en la ciudad de Bogotá.
- Desarrollar el método automático aplicando algoritmos de Machine Learning al Dataset para evaluar cuál de ellos se ajusta mejor a la naturaleza de los datos.
- Evaluar el desempeño del método automático, a partir de las métricas de modelos de regresión para su implementación.

5 MARCO REFERENCIAL

5.1 Marco conceptual

5.1.1 Inmuebles

Los bienes inmuebles son aquellos que como su nombre lo indica son inmóviles, que no se pueden separar del terreno en el que se encuentran, según Pérez y Gardey, estos bienes forman parte de lo que se conocen como bienes raíces, dada a su naturaleza arraigada al suelo, en donde se pueden encontrar lo que son las casas, los edificios o terrenos [13], usualmente el término de inmueble suele asociarse al concepto de vivienda, que será explicado en el siguiente término. Un inmueble es lo que se conoce como bienes, pero de manera más específica de los bienes raíces, de tal modo, son terrenos o construcciones que se caracterizan como bienes física o jurídicamente.

Actualmente existen dos tipos de inmuebles, el inmueble rústico y el inmueble urbano, el inmueble rústico se puede identificar por ser de carácter más rudimentario y agrario, estos están ubicados en terrenos que son utilizados para el desarrollo de actividades agrícolas, ganaderas o forestales. Por otro lado, el inmueble urbano se puede encontrar en zonas más de ciudad con casas, edificios y locales. Como dato adicional, según el blog “rebajatuscuentas.com” existe otro tipo de inmueble poco conocido llamado inmueble mostrenco, caracterizado por diversas causas como la defunción, abandono o ausencia del propietario[14].

5.1.2 Vivienda

Según Pérez y Gardey, se define como vivienda aquel lugar cerrado y cubierto que es construido para que habiten en ellas personas, ofrece refugio a estas y los protege de condiciones climáticas, ofreciendo intimidad y espacio seguro para guardar sus cosas[15]. Una casa, apartamento, residencia piso, entre otros son usados como sinónimos para vivienda, la utilización de cada concepto depende generalmente de características asociadas a su construcción.

---

[13] Julián Pérez Porto y Ana Gardey. Definicion.de: Definición de inmueble, {En línea}. {23 de octubre del 2020} disponible en: (https://definicion.de/inmueble/). P.23

[14] Rebajatuscuentas.com, ¿Qué es un inmueble y qué tipo de inmuebles existen?, {En línea}, {8 de mayo del 2021} disponible en: (https://rebajatuscuentas.com/pe/blog/que-es-un-inmueble) p.23

[15] Julián Pérez Porto y Ana Gardey Definicion.de: Definición de vivienda, {En línea}. {23 de octubre del 2020} disponible en: (https://definicion.de/vivienda/) p.23

De la misma manera se le puede definir a una vivienda como el lugar donde vive una persona, grupo de personas o familia que constituye un hogar[16]. Actualmente el diseño de todo tipo de viviendas es una manera de competencia para los arquitectos e ingenieros, pero también existen personas que pueden diseñar sus propias plantas, pero de la misma manera deben tener legalmente la firma de un profesional de construcción en los planos del diseño, sea un maestro mayor de obras, arquitecto o ingeniero civil.

5.1.3 Avalúos comerciales y avalúos catastrales.

Empezando con el objeto principal de estudio de este trabajo, ¿Qué es un avalúo? Es un proceso que permite determinar el valor de venta de un bien de acuerdo con sus características físicas, ubicación, uso, entre otras[17]. Existen dos clases de avalúos: el avalúo comercial y el avalúo catastral o fiscal. el objetivo de este proyecto apunta a trabajar con los avalúos comerciales los cuales ofrecen el valor de compra-venta de un inmueble en momento determinado, lo cual difiere del avalúo catastral, ya que este se utiliza para cobros de impuestos y suele representar entre 50% o 75% del valor de avalúo comercial además que es hecho exclusivamente por el estado colombiano.

Aunque el avalúo comercial es utilizado en los procesos de compra y venta, este no define el valor total del negocio, ya que este depende de la decisión de los compradores y vendedores. De la misma manera Grajales define un avalúo catastral como aquel que se obtiene mediante la investigación y análisis estadístico del mercado inmobiliario y el cual puede ser tomado como referencia para determinar el precio real de una vivienda[18], este precio de la vivienda no representa el 100% del avalúo comercial como se expresa anteriormente, pues el precio de la vivienda también depende de cómo se está comportando el mercado en determinado sector y tiempo, es por esto que no siempre el valor catastral refleja el verdadero precio de la vivienda.

5.1.4 Mercado Inmobiliario.

El mercado inmobiliario es el conjunto de las acciones de oferta y demanda de bienes inmuebles, entendiendo que la oferta es el número de inmuebles disponibles

[16] Diccionarioactual.com, ¿Qué es vivienda?, {En línea}, {8 de mayo del 2021}, disponible en: (https://diccionarioactual.com/vivienda/) p.24

[17] SUÁREZ, Alexandra, Antes de vender o comprar, piensa en hacer un avalúo. {En línea}, {27 de octubre 2020}, disponible en: (https://www.metrocuadrado.com/noticias/guia-de-compra/antes-de-vender-o-comprar-piensa-en-hacer-un-avaluo-2863) p.24

[18] GRAJALES Yuri, Modelo predicción precios viviendas proyecto Medellín, Medellín, 2020, 74 Págs, Modelo predicción precios viviendas proyecto Medellín, Facultad de Ingeniería, Departamento en Tecnologías++ de la Información y Comunicación. p.24

en el sector, y la demanda son las personas que están dispuestas a comprar un sector determinado. La naturaleza de estos bienes puede ser muy distinta, diferenciándose entre bienes de naturaleza residencial, comercial, industrial, urbano, etc. Todas las operaciones que se produzcan relacionadas con la compra y venta de este tipo de inmuebles forman el sector inmobiliario, esencial para el desarrollo de una economía sostenible de un país. La naturaleza de estos bienes puede ser muy distinta, diferenciándose entre bienes de naturaleza residencial, comercial, industrial, urbano, etc. Todas las operaciones que se produzcan relacionadas con la compra y venta de este tipo de inmuebles forman el sector inmobiliario, esencial para el desarrollo de una economía sostenible de un país.[19]

En Colombia a pesar de la situación mundial respecto al tema de salud, según los expertos, el comportamiento del sector inmobiliario presenta condiciones favorables este año, ya que es uno de los sectores económicos más estables del país, lo que ha hecho que el Gobierno Nacional proporcione incentivos para que los ciudadanos inviertan en su vivienda propia. Por otro lado, el panorama se muestra como una oportunidad única dado que los precios no han mostrado un aumento significativo desde la emergencia sanitaria, y se espera que este año la dinámica se mantenga igual.[20]

5.1.5 Conjunto de Datos (“Dataset”)

Un conjunto de datos es un término utilizado comúnmente para la investigación de casos, análisis de estadística o la ciencia de los datos, presentado como una agrupación de información que, luego de ser extraída, manipulada y transformada será de eficaz utilidad para ser trabajada. Según Thais Balaguero, un Dataset es un término extranjero, que se ha incorporado a nuestra lengua como un término más en los países hispanohablantes. Su traducción Dataset a conjunto de datos es una colección de datos habitualmente tabulada[21].

Este es un concepto ya bastante conocido como muestra Thais, los datos son un grupo de información que nutre proyectos de investigación los cuales necesitan de estos conjuntos de datos para su análisis y de esto sacar una conclusión para un tema en específico, estos Dataset están compuestos por una gran cantidad de datos

[19] REALIA. ¿Qué es el mercado inmobiliario?, {En línea}. {3 de mayo 2021} Disponible en (https://www.realia.es/que-es-mercado-inmobiliario) p.25

[20] Equipo de redacción de OIKOS, Perspectivas del sector inmobiliario para este 2021, {En Línea}.{3 de mayo 2021} Disponible en (https://www.oikos.com.co/inmobiliaria/noticias-inmobiliaria/como-va-el-sector-inmobiliario) p.25

[21] BALAGUERO Thais, ¿Qué son los Dataset y los dataframe en el Big Data?, {En Línea} {13 de mayo de 2021}, Disponible en (https://www.deustoformacion.com/blog/programacion-tic/que-son-datasets-dataframes-big-data). P.25

o así mismo por un pequeño conjunto, pero siempre tienen la misma función de sustentar y fortalecer un trabajo, tarea, o proyecto de investigación.

5.1.6 Inteligencia Artificial

Según García, [22] La Inteligencia Artificial (IA) es un subcampo de la informática que se creó en la década de 1960, y que trata de solucionar tareas que son sencillas para los seres humanos, pero difíciles para las computadoras. Se trata de un concepto bastante genérico e incluye todo tipo de tareas tales como la planificación, el reconocimiento de objetos y sonidos, hablar, traducir, realizar actividades creativas (como por ejemplo crear obras de arte, o la poesía), etc. A continuación, se encuentra una imagen (figura 2) donde están las distintas técnicas de inteligencia artificial aplicadas hoy en día.

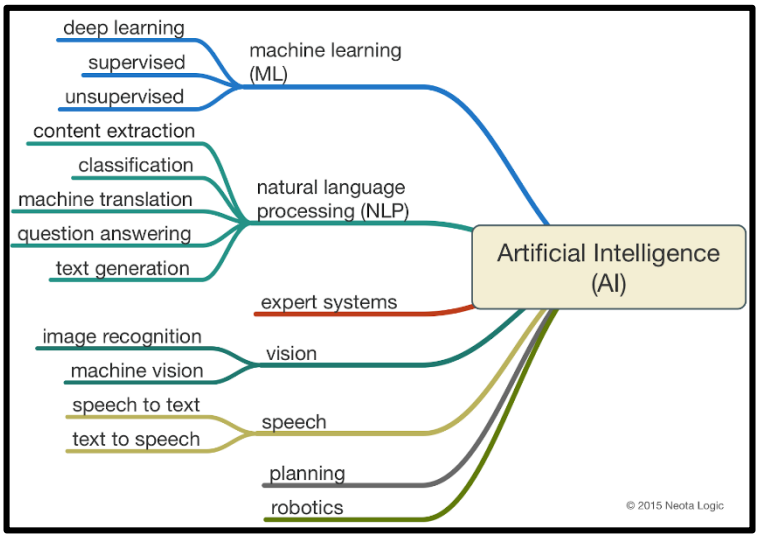

Figura 2, Ramas de la inteligencia artificial, Fuente: https://planetachatbot.com/historia-de-la-inteligencia-artificial-relacionada-con-los-chatbots-41a6cda22906

Así como se observa en la figura 2, la inteligencia Artificial se compone de otras subramas en la cual se observa el Machine Learning, el procesamiento de lenguaje natural, entre otras más. De la misma manera, como subramas del Machine Learning (ML) se evidencia el Deep Learning y el aprendizaje supervisado y no supervisado.

[22] GARCÍA, Carlos, ¿Qué es el Deep Learning y para qué sirve? {en línea}, {23 de octubre de 2020}, Disponible en (https://www.indracompany.com/es/blogneo/deep-learning-sirve) p.26

5.1.7 Aprendizaje de máquina (Machine Learning)

Según García[23], El Machine Learning, se describe a menudo como un tipo de técnicas de Inteligencia Artificial donde las computadoras aprenden a hacer algo sin ser programadas para ello. El aprendizaje de máquina se caracteriza por implementar distintas técnicas como lo es la lógica difusa o las redes neuronales para que aprenda a hacer alguna acción específica, existen dos clases de aprendizaje de máquina, supervisado en donde, según Zambrano[24], se entrena al algoritmo otorgándole las preguntas, denominadas características, y las respuestas, denominadas etiquetas, por otra parte existe el aprendizaje no supervisado, donde la máquina debe aprender por sí sola a realizar las acciones que se le pida sin un conocimiento previo de las etiquetas, creando solamente grupos que tienen características similares.

Es de esta manera como el Machine Learning y sus algoritmos son capaces de encontrar patrones a partir de un conjunto de datos suministrados y de esta manera crear un modelo de predicción y toma de decisiones como se evidencia en este proyecto de investigación como apoyo a la toma de decisiones en los bienes raíces.

5.1.8 Aprendizaje Supervisado

Es un tipo de aprendizaje que se presenta en los modelos de Machine Learning como árboles de decisión, las máquinas de soporte vectorial, la regresión por mínimos cuadrados y métodos “Ensemble” como el Random forest. En este aprendizaje los algoritmos o modelos trabajan con un conjunto de datos ya etiquetados o definidos donde tienen como objetivo encontrar una función que, según las variables de entrada se les asigne una etiqueta de salida[25], trayendo como ejemplo el presente proyecto de investigación, de unas variables ya etiquetadas como lo es x número que corresponde al número de habitaciones, baños, parqueaderos, entre otras más, el algoritmo tendrá como salida un valor el cual como etiqueta corresponderá al precio de una vivienda.

El aprendizaje se suele usar en problemas de regresión o clasificación, estos dos tipos de problemas se pueden diferenciar por el tipo de variable objetivo a buscar, es decir, en los casos de clasificación esta variable objetivo es de tipo categórico o

---

[23] GARCIA Carlos, ¿Qué es el Machine Learning? Ibíd. P.27

[24] ZAMBRANO, Juan, ¿Aprendizaje supervisado o no supervisado? Conoce sus diferencias dentro del Machine learning y la automatización inteligente, {En línea}, {23 de octubre de 2020}, Disponible en (https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b) p.27

[25] SANTOS Paloma, Tipos de aprendizaje en Machine Learning: supervisado y no supervisado, {En Línea}, {10 de mayo de 2021}, Disponible en https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/) p.27

cualitativo, mientras que para la regresión dicha variable es de tipo cuantitativo o numérico[26], por ejemplo como problema de regresión se puede evidenciar en el proyecto actual al trabajar con predicciones de vivienda de bienes raíces en Bogotá.

5.1.9 Aprendizaje profundo (Deep Learning)

Según García[27] El Deep Learning lleva a cabo el proceso de Machine Learning usando una red neuronal artificial que se compone de un número de niveles jerárquicos. En el nivel inicial de la jerarquía la red aprende algo simple y luego envía esta información al siguiente nivel. El siguiente nivel toma esta información sencilla, la combina, compone una información algo un poco más compleja, y se lo pasa al tercer nivel, y así sucesivamente.

Así mismo como el Machine Learning, el Deep Learning busca dar solución a un problema por medio de algoritmos matemáticos y con la ayuda de una gran variedad de redes neuronales las cuales pueden ser aplicadas de distintas maneras según el tipo de problema que este propuesto, desde el análisis de imágenes con ayuda de las redes neuronales convolucionales, o el análisis de datos históricos para predecir datos futuros por medio de las redes neuronales recurrentes entre otros tipos de redes neuronales que componen la área del aprendizaje profundo.

5.1.10 Redes neuronales artificiales

Según García[28] Las redes neuronales artificiales están basadas en el funcionamiento de las redes de neuronas biológicas, según la figura que se muestra a continuación, la suma de las entradas multiplicadas por sus pesos asociados determina el “impulso nervioso” que recibe la neurona. Este valor, se procesa en el interior de la célula mediante una función de activación que devuelve un valor que se envía como salida de la neurona.

[26] SANTOS Paloma, Op, Cit, p.28

[27] GARCÍA, Carlos. Op. Cit. P.28

[28] GARCÍA, Oscar, Redes Neuronales artificiales: Qué son y cómo se entrenan, {En línea} {23 de octubre del 2020}, Disponible en (https://www.xeridia.com/blog/redes-neuronales-artificiales-que-son-y-como-se-entrenan-parte-i) p.28

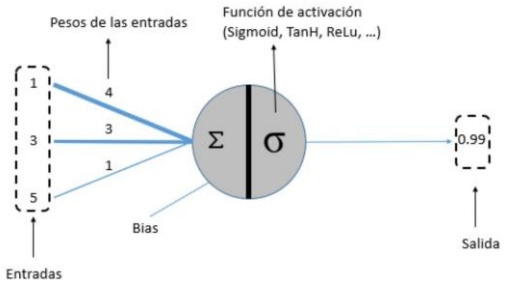

Figura 3, Estructura básica de una neurona, Fuente: https://www.xeridia.com/blog/redes-neuronales-artificiales-que-son-y-como-se-entrenan-parte-i

Como se puede apreciar en la anterior imagen (Figura 3), una red neuronal se podría definir como un grafo dirigido, en donde la capa entrada, contiene las características iniciales a trabajar, ya sea los pixeles de una imagen a clasificar, o las características que se quieran trabajar de un conjunto de datos, cada valor es una neurona dentro de la capa de entrada. Las capas ocultas contienen diferentes conjuntos de neuronas que influyen en el valor final en la capa de salida, cambiando su valor según los pesos y sesgos que le sean asignados y utilizando una función de activación para definir su valor final, y la capa de salida, que serían la clasificación que se está buscando.

Del mismo modo, García[29] afirma que el cerebro humano está compuesto por neuronas interconectadas entre sí, una red neuronal artificial está formada por neuronas artificiales conectadas entre sí y agrupadas en diferentes niveles que se denominan capas:

" Una capa es un conjunto de neuronas cuyas entradas provienen de una capa anterior (o de los datos de entrada en el caso de la primera capa) y cuyas salidas son la entrada de una capa posterior."

5.1.11 Páginas web

Según Pérez y Merino[30], una página web es un documento que forma parte de un sitio web y que suele contar con enlaces (también conocidos como hipervínculos o links) para facilitar la navegación entre los contenidos. Las páginas web están desarrolladas con lenguajes de marcado como el HTML, que pueden ser interpretados por los navegadores. De esta forma, las páginas pueden presentar

[29] GARCÍA, Oscar. Op. Cit. P.29

[30] Julián Pérez Porto y María Merino. Definicion.de: Definición de página web, {En línea}. {23 de octubre del 2020} disponible en: (https://definicion.de/pagina-web/) p.29

información en distintos formatos (texto, imágenes, sonidos, videos, animaciones), estar asociadas a datos de estilo o contar con aplicaciones interactivas.

Hasta el día de hoy se puede evidenciar cómo han evolucionado las páginas web, incluso más unas que otras, se puede observar cómo páginas como *Finca Raíz* ofrecen una amplia base de datos de sus viviendas ofertadas con una variedad de características cada una, a diferencia de otras páginas web, de esta manera se puede observar la variedad de páginas web que existen y cómo se diferencian unas de otras.

5.1.12 Web Scraping

El web scraping es una técnica en la cual se busca navegar automáticamente por los sitios web y extraer información de estos sitios, se utiliza un Bot que se configura manualmente por los usuarios para poder navegar por los sitios, es por eso que a la hora de llenar un formulario sale la pregunta de “no soy un robot”, previniendo precisamente de un Bot[31]. El web scraping es una técnica muy popular hoy en día para la obtención de datos de distintas páginas web, partiendo de una sencilla programación se puede lograr obtener un considerable conjunto de datos con el cual trabajar. Estos datos son extraídos y algunos webs scraping tienen la posibilidad de ser almacenados en una base de datos.

De esta manera los webs scraping tienen la posibilidad de ser configurados para la extracción de información diaria y de manera automática, hoy en día existen grandes plataformas y portales web que brindan herramientas para diseñar estos web scraping de manera más sencilla, algunos de manera gratuita y limitada y otros con licencias de pago los cuales ofrecen más que un simple robot de extracción, como la posibilidad de ampliar la cantidad de información extraída y las páginas web a las cuales van dirigida, de este modo los robots cuentan con herramientas para la limpieza de datos como la eliminación de formato HTML en la información.

5.1.13 Bases de datos no relacionales

Las bases de datos no relacionales son las que, a diferencia de las relacionales, no tienen un identificador que sirva de relación entre un conjunto de datos y otros. La información se organiza normalmente mediante documentos y es muy útil cuando no se tiene un esquema exacto de lo que se va a almacenar[32]. Estas bases de datos

---

[31] LAFUENTE, Ainhoa, Que es el web Scraping, {En Línea}, {23 de octubre 2020} Disponible en (https://aukera.es/blog/web-scraping/) p.30

[32] LAFUENTE, Ainhoa, Bases de datos relacionales vs. no relacionales: ¿qué es mejor?, {En Línea},{23 de octubre 2020 }Disponible en (https://aukera.es/blog/bases-de-datos-relacionales-vs-no-relacionales/) p.30

no relacionales son un sistema de almacenamiento de información que se caracteriza por no usar el lenguaje SQL para las consultas. No significa que no lo utilicen del todo, pero se utiliza más como apoyo, de aquí el término NoSQL o no solo SQL según el blog español “ayudaleyprotecciondatos.es” el cual también afirma que lo que caracteriza una base de datos no relacional es que la información no es almacenada en tablas sino en documentos, son bases de datos con un alto grado de escalabilidad y están diseñadas para soportar grandes volúmenes de datos, pero como contramedida las bases de datos no relacionales carecen de un sistema de estandarizado debido a que son bases de datos relativamente nuevas[33].

5.1.14 MongoDB

Dentro de las varias opciones que existen para implementar una base de datos no relacional se encuentra MongoDB la cual está orientada a documentos. Esto quiere decir que, en lugar de guardar los datos en registros, guarda los datos en documentos. Estos documentos son almacenados en BSON, que es una representación binaria de JSON[34].
Al hablar de bases de datos es muy común que el siguiente pensamiento sea SQL y en las bases de datos relacionales, MongoDB no es partícipe de ellas, al ser una base de datos no relacional que guarda los datos en tablas con un esquema dinámico, MongoDB se caracteriza por su gran velocidad la cual alcanza un balance perfecto entre rendimiento y funcionalidad, según Ángel Robledano quien presenta a MongoDB como una base de datos en donde se pueden realizar todo tipo de consultas, desde la búsqueda por campos, consultas por rangos y expresiones regulares[35], esto debido a que MongoDB es una base de datos orientada a documentos, es decir, que en lugar de almacenar los datos en registros, guarda los datos en documentos.

[33] Ayudaleyprotecciondatos.es, Base de datos no relacional. ¿Qué es? Características y ejemplos, {En línea}, {9 de mayo. de 2021}, Disponible en (https://ayudaleyprotecciondatos.es/bases-de-datos/no-relacional/) p.31
[34]FERNÁNDEZ, Rubén, MongoDB: qué es, cómo funciona y cuándo podemos usarlo (o no), {En Línea},{23 de octubre 2020 }Disponible en (https://www.genbeta.com/desarrollo/mongodb-que-es-como-funciona-y-cuando-podemos-usarlo-o-no) p.31
[35] ROBLEDANO Ángel, Qué es MongoDB, {En línea}, {9 de mayo de 2021} Disponible en (https://openwebinars.net/blog/que-es-mongodb/) p.31

5.2 Estado del arte

En este apartado se muestran los distintos casos de estudios que están más relacionados con este proyecto, donde se presentan técnicas de Machine Learning para la predicción de precios de vivienda en diferentes países y ciudades del mundo. Es importante mencionar que, estos artículos o casos de estudios fueron encontrados con las siguientes palabras clave: “neural network + real state” y “Deep Learning + real state” en la base de datos de Scopus[36] proporcionada por la universidad Católica de Colombia.

En razón a que desde el punto de vista técnico el componente más importante en el desarrollo del proyecto se basa en la búsqueda inteligente de patrones de acuerdo con ciertos criterios y variables, en este apartado se presenta el resultado de la búsqueda de artículos científicos que ofrecen una gran cantidad de casos de estudio relacionados con técnicas muy similares de aprendizaje de máquina, como la extracción de características y los modelos implementados. En uno de estos artículos Chao Maab Zhenbing, Liub Zhiguang, Caoc Wen Songd y JieZhange Weiliang en el año 2019 mencionan cómo aplicar métodos de discretización como una forma de clasificar y aplicar los modelos de Random Forest, Deep Forest y el algoritmo de K means, dichos modelos y algoritmos son modificados o aplicados a la discretización de los datos para un aprendizaje sensible aplicado al costo de las propiedades residenciales[37].

Por otro lado, se encuentra un nuevo artículo realizado por el estudiante miembro de la IEEE Junchi Bin en 2019 en el cual se trabaja el estudio de bienes raíces en la ciudad de Filadelfia, en este artículo se emplean los árboles de decisiones como técnicas de clasificación en métodos de aprendizaje supervisado del Machine Learning debido a su alta precisión, estabilidad y facilidad de interpretación. Estos árboles son mejorados o adaptados para estimar los valores de mercado aplicando como metodología la fusión de los datos. En dicho artículo también se aplica como método de Deep Learning las redes neuronales profundas (DNN), familia de los métodos del Machine Learning, dichas redes están basadas en las redes neuronales artificiales (ANN). Dicho esto, las redes neuronales profundas son construidas con 3 capas ocultas en donde se concatenan como variables la fusión de los datos (a la neurona ingresan las variables como los atributos de las casas, las actividades humanas y las características de la vivienda) y los pisos esperados en la propiedad (a la neurona ingresa el formato de imagen que revela el número de niveles con la que esta cuenta). Finalmente, con la ayuda

[36] Base de Datos de Scopus. Disponible en: (https://www-scopus-com.ucatolica.basesdedatosezproxy.com/search/form.uri?display=basic#basic) p. 32

[37] Chao Ma, Zhenbing Liu, Zhiguang Cao, Wen Song, Jie Zhang, Weiliang Zeng, Cost-sensitive Deep Forest for Price Prediction, Pattern Recognition (2020), doi: https://doi.org/10.1016/j.patcog.2020.107499. P.32

de los árboles de regresión potenciados se realiza una estimación del valor en el mercado[38].

Por otra parte, en un artículo de predicción del costo de la vivienda mediante redes neuronales, elaborado por los científicos A.N Boyko, G.A Gladyshev y otros del año 2020, se presenta una posibilidad de implementar una red neuronal de Kohonen con métodos de binarización para la predicción de precios en el mercado inmobiliario. La red neuronal de Kohonen es un tipo de red neuronal artificial entrenada mediante un aprendizaje no supervisado para producir una representación discreta del espacio de las muestras de entrada, llamado mapa[39].

De este modo, en gran parte de los artículos se presenta el desarrollo de métodos en la identificación de precios de bienes raíces, implementando desde árboles de regresión hasta métodos de binarización para la extracción de características. De esta forma, se encuentra cómo se realizan diversos estudios con ayuda del Deep Learning para su aplicación en diferentes países, de la misma forma como se presenta en Filadelfia aparece un artículo científico reciente, en el cual se expone el uso de datos por sistemas inteligentes para la previsión de precios en el contexto de la construcción de relaciones con los clientes en el mercado inmobiliario de Lublin. Este escrito realizado por el científico J Lipski en el 2019 señala que un conjunto de datos (incluso con tan solo 50), la selección de características mediante encuestas de precios inmobiliarios para los clientes y las redes neuronales artificiales, se pueden pronosticar los precios inmobiliarios a partir de datos históricos[40].

De este modo, se encuentran múltiples artículos científicos relacionados con el mismo tema o con el mismo enfoque, al realizar la predicción de precios en los bienes raíces, se encuentran distintas aplicaciones con las redes neuronales, DNN, ANN o CNN. Haciendo énfasis en esta última, el científico Yong Piao en el 2019 presentó un artículo científico en el cual se muestra cómo predecir el precio de vivienda a partir de CNN mediante el método del cálculo del gradiente “Backpropagation” son entrenadas las redes neuronales convolucionales o CNN, considerando que varios factores pueden tener influencias entrelazadas inesperadas al momento de la predicción del precio de la vivienda, dicho esto se presenta un modelo basado en CNN para lidiar con la complejidad[41].

[38] Bin, J.a, Gardiner, B.b, Li, E.c, Liu, Z.a, Op. Cit. P.33
[39] Boyko, A.N. Email Author, Gladyshev, A.G. Email Author, Kadyrova, G.M. Email Author, Barmenkova, N.A. Email Author, Zybenko, A.V. 2020 Predicting the cost of housing using neural networks. P.33
[40] A Bojanowska1 and J Lipski | 2019 | The use of data by smart systems for price forecasting in the context of building customer relationships on the Lublin real estate market | IOP. P.33
[41] Yong Piao, Ansheng Chen, Zhendong Shang | 2019 | Housing Price Prediction Based on CNN | ICIST. P.33

Ahora bien, gran parte de los artículos encontrados están basados en técnicas y modelos similares al anterior, como en el siguiente documento, el científico Yao Sun en el año 2019 presentó como modelo de evaluación de bienes raíces basado en algoritmo genético de red neuronal optimizada por medio del cual se exponen métodos de backpropagation en un modelo de redes neuronales, aunque en el caso de este documento, las redes neuronales son optimizadas por un algoritmo genético. Dicho algoritmo, en general, cuenta con una excelente capacidad de búsqueda global el cual puede obtener eficazmente la solución óptima global y compensar el defecto de la red neuronal, es decir, el algoritmo genético tiene la capacidad de exploración macroscópica y optimización global en la red neuronal[42].

De manera análoga en el siguiente artículo se presenta un modelo similar a los mencionados anteriormente, pues el científico Junchi Bin en el 2017 presentó un modelo de regresión para el avalúo de inmuebles mediante redes neuronales recurrentes y un árbol optimizado (como los presentados anteriormente) en donde se presenta un modelo bastante importante en el área inmobiliaria, pues se trata del AVM o la valoración automatizada de inmuebles la cual es una herramienta útil para evaluar las carteras de forma global, pero tiene sus limitaciones y no siempre se utiliza bien. Los AVM más comunes empleados por la industria de los avalúos se basan en el análisis de regresión múltiple. Se encuentran otras herramientas analíticas, como el aprendizaje estadístico y los algoritmos difusos los cuales se han vuelto más populares debido a la capacidad cada vez mayor de recopilar un gran volumen de datos y al avance del Machine Learning. En el artículo se propone un modelo de árbol de impulso facilitado con una Red Neuronal Recurrente (RNN) para pronosticar el precio promedio de un área, los autores presentan dicho modelo el cual dan a entender como un modelo más avanzado que los existentes en la industria de los avalúos[43].

Como contrapartida ahora se presenta un artículo poco distinto a los anteriores pues los científicos estadounidenses Cornell Omid Poursaeed, Tomáš Matera y Serge Belongie presentan una estimación de precios inmobiliarios basada en la visión. En este artículo, se evalúa el impacto de las características visuales de una casa en su valor de mercado. Usando redes neuronales convolucionales profundas (Deep CNN) en un gran conjunto de datos de fotos de interiores y exteriores de casas, se desarrolla un método para estimar el nivel de lujo de las fotos de bienes

[42] Yao Sun, | 2019 | Real Estate Evaluation Model Based on Genetic Algorithm Optimized Neural Network | Data Science Journal. P.34

[43] Junchi Bin, Shiyuan Tang, Yihao Liu, Gang Wang, Bryan Gardiner, Zheng Liu, Eric Li | 2017 | Regression model for appraisal of real estate using recurrent neural network and boosting tree | IEEE Intternational Conference. P.34

raíces. También se desarrolla un marco novedoso para la evaluación de valor automatizada utilizando fotos ya tomadas, además de las características del hogar, incluido el tamaño, precio ofrecido y cantidad de dormitorios. Finalmente, se aplica dicho método propuesto para la estimación de precios a un nuevo conjunto de datos de fotos y metadatos inmobiliarios[44].

Ahora bien, como previamente se ha mencionado en artículos con investigaciones aplicadas a distintos países, ahora se presenta este nuevo caso en Hong Kong, por el profesor australiano Rotimi Boluwatife Abidoye, el profesor Albert P.C Chan, de la universidad de Hong Kong y los científicos Funmilayo Adenike Abidoye y Olalekan Shamsideen Oshodi en el año 2018 donde se expone la predicción del índice de precios de la propiedad utilizando técnicas de inteligencia artificial, este estudio tiene como objetivo investigar el uso de la inteligencia artificial (IA) para la predicción del índice de precios de la propiedad (IPP). Mediante la comparación de modelos como ARIMA, redes neuronales artificiales (ANN) y máquinas de soporte vectorial (SVM), los autores generaron la predicción de precios inmobiliarios, y finalmente se reveló que el modelo ANN supera a los modelos SVM y ARIMA en la predicción de índice de precios de las propiedades[45].

Igualmente como se ha observado en distintas investigaciones acerca del área inmobiliaria en países como Hong Kong, se encuentran también predicciones de *Finca Raíz* en América del Norte, en este caso se evidencia como el científico Yitong Huang realiza la predicción de costo de vivienda en el estado de California apoyándose con el campo del Machine Learning y distintos métodos de aprendizaje como las máquinas de soporte vectorial o árboles decisión y la comparación de resultados de los mismos para de esta manera determinar cuál de dichos métodos son de los más eficientes, y es de esta manera como se realiza para este trabajo de investigación la selección de cuáles son los mejores métodos para aplicarlos durante el diseño del método automático para su posterior desarrollo. Así como lo realizaron los científicos Quansheng You, Ran Pang, Liangliang Cao, y Jiebo Luo en el año 2017 cuando presentaron su artículo científico en el cual presentan como se realiza un avalúo de una propiedad a partir de una imagen y pocos datos específicos de la propiedad a evaluar, con esta gran variedad de artículos científicos con el mismo enfoque a este proyecto se busca tomar soporte para el apoyo y sustentación del mismo.

Después de analizar una gran variedad artículos científicos se puede evidenciar que el tema o enfoque de este trabajo ya ha sido tratado por otros científicos, ingenieros o investigadores quienes han aplicado estas técnicas de la Inteligencia Artificial en diferentes países y ciudades alrededor del mundo. Como lo demuestran los

[44] Poursaeed, O. Matera, T. Belongie, S.| 2018 | Vision-based real estate price estimation | Machine Vision and Aplications. P.35

[45] Rotimi Boluwatife Abidoye, Albert P.C. Chan and Funmilayo Adenike Abidoye, Olalekan Shamsideen Oshodi. Op. cit. P.35

ingenieros Xibin Wanga, Junhao Wena, Yihao Zhanga y Yubiao Wang de la universidad de Chongqing en China del año 2013 en el cual con tan solo máquinas de soporte vectorial y con la selección de los correctos parámetros se logra la predicción del precio de viviendas. Es importante mencionar que como se ha evidenciado en los casos de estudio previamente mencionados, se han utilizado una gran variedad de modelos del Machine Learning, así como el random forest, el Deep forest, los árboles de decisión y las máquinas de soporte vectorial, etc. Dan a entender el gran conjunto de oportunidades con las cuales se puede desarrollar un método automático, mediante experimentos y pruebas de rendimiento de cada modelo, se busca elegir cuál es el que destaca entre todos. Es decir, de la variedad de casos de estudio relacionados, se puede observar como el random forest es uno de los modelos de Machine Learning más empleado en casos de regresión o de predicción.

Ahora bien por lado del método en el que se realiza la extracción de los datos y de esta manera obtener un conjunto de datos con el cual trabajar, estos pueden ser extraídos y obtenidos de diferentes formas, desde la adquisición de licencias de un robot creado por expertos en la materia, hasta la creación de un robot de manera personalizada por medio de entorno Python y con librerías que apoyan el desarrollo del mismo, como lo implemento la magister en tecnologías de la información y comunicación, Yuri Vanesa Grajales Álzate el 24 de Febrero del 2020, en la ciudad de Medellín, la cual presenta un robot de web scraping realizado mediante entornos de Python y con librerías como “Beautifulsoup, Selenium y Scrappy”, para la extracción de información útil en páginas web como *Finca Raíz*, mercado libre, etc. Esto demuestra para este trabajo que la forma de obtención de los datos se puede realizar por medio de distintas maneras y técnicas de raspado web[46].

[46] GRAJALES, Yuri, Modelo predicción precios viviendas proyecto Medellín, Op, Cit, p.36

5.3 Marco teórico

5.3.1 Portal web *Finca Raíz*

Finca Raíz es el Portal Web en Colombia dedicado al mercado de Bienes Raíces. fincaraiz.com.co ofrece en Bogotá, Medellín, Valle del Cauca, Cali y el resto del país, inmuebles en venta y arriendo, en donde se pueden encontrar anuncios clasificados de Constructoras, Inmobiliarias o Particulares: vivienda usada, inmuebles comerciales, proyectos nuevos de vivienda y otras propiedades[47]. Actualmente *Finca Raíz* cuenta con 208.000 anuncios publicados de venta de inmuebles en todo el país, para anuncios de apartamentos y casas en la ciudad de Bogotá cuenta aproximadamente con 55.000 anuncios. Es importante mencionar que día a día se agregan más propiedades a la venta en este Portal Web (donde incluso se encuentran mismas viviendas repetidas) y *Finca Raíz* es una página la cual destaca entre otras como mercado libre, al ofrecer una interfaz amigable y muy exacta de lo que el cliente necesita, buscando presentar las mismas características en la mayoría de inmuebles que tienen publicados en su Portal Web. Dentro de las características que ofrece *Finca Raíz* para la búsqueda de inmuebles, están los distintos filtros que ofrece, como lo son el número de habitaciones, el área del inmueble, el estrato del inmueble, la antigüedad del inmueble, la ubicación del inmueble entre otros.

Dentro de los avisos publicados se encuentra una información mínima para cada anuncio, como lo es el área del inmueble, el número de baños, la cantidad de habitaciones y el número de parqueaderos. Después cada anuncio puede o no tener ciertas características publicadas, esto depende de cada anunciador, esas características puede ser el estrato, la ubicación, la antigüedad, el número de pisos, el piso donde está ubicado si es un apartamento, el valor de la administración si es conjunto cerrado, entre otras. Luego se puede encontrar una descripción del inmueble hecha por el anunciante, en donde detalla brevemente las características más relevantes del inmueble, y luego se encuentra de una manera más sencilla, una descripción del sector, junto con otros detalles del inmueble como lo son si cuenta con jardín o terraza, por último se encuentra la ubicación del inmueble dentro del mapa de su respectiva ciudad, esta ubicación puede ser real o una ubicación aproximada, dada la interfaz que se utiliza para mostrar el mapa. Aprovechando los filtros que tiene esta herramienta, se puede filtrar para que busque solamente inmuebles de vivienda usados, colocando de filtros casas y apartamentos (que son los inmuebles más buscados por las personas), junto con la antigüedad para así evitar los proyectos que se clasifican como nuevos dentro del portal.

[47] *FINCA RAÍZ*.COM, ¿Quiénes Somos?, {en línea}, {25 de octubre 2020}, disponible en (https://www.fincaraiz.com.co/). P.37

5.3.2 DEXI.IO, Suite de inteligencia de comercio digital

Dexi cuenta con un motor ETL avanzado que administra y organiza metadatos, la configuración de dexi permite definir y construir los procesos y reglas dentro de la plataforma que, en función de requisitos de datos, instruirá a los robots en base a cómo se enlazan y controlan junto a otros robots extractores para capturar datos de fuentes de datos externas específicas[48].
Las reglas para la transformación de los datos extraídos (como la eliminación de duplicados) también se pueden definir para crear los archivos de salida unificados deseados. La definición de dónde se envían los datos hacia, desde y quién tiene derechos de acceso también se realiza dentro de la plataforma como por ejemplo Azure, Hannah, Google Drive, Amazon S3, Twitter y Google Sheets, herramientas visuales u otro entorno. Dexi es un software de rastreo web/extracción de datos avanzado. Rastrea los datos de cualquier sitio web/fuente en línea utilizando esta interfaz muy intuitiva tal que es apuntar y hacer clic y dexi envía los datos extraídos en un formato deseado.

Esta es una herramienta de rastreo web/extracción de datos para personas de nivel superior, Dexi es una herramienta de rastreo web basada en la nube que permite a empresas extraer y transformar datos desde cualquier fuente en la Web o en la nube a través de tecnología de automatización avanzada y extracción inteligente. Los robots de rastreo avanzado de Dexi.io, y la compatibilidad con el entorno de navegadores, permiten rastrear e interactuar con los datos de cualquier sitio web con precisión humana. Una vez que se extraen los datos, Dexi.io ayuda a transformarlos y combinarlos en un conjunto de datos. esta magnífica herramienta cuenta con diversas funciones tales como:
Herramientas de extracción de datos en documentos y páginas web, extracción de direcciones IP, extracción de direcciones de correo electrónico, extracción de precios, extracción números de teléfonos móviles, extracto de imágenes, recolección de datos dispares.

Dexi puede monitorear los mercados tanto de los minoristas como de las marcas para obtener visibilidad de quien vende productos, dónde se venden, quién los vende y a qué precio se venden. Dexi puede realizar más análisis en el mercado para ver cómo se venden estos productos, conocer quién gana la caja de compra y por qué[49].

- **Robots de captura de datos digitales**

Los robots de captura de datos digitales son robots inteligentes que permiten automatizar la recopilación de datos desde cualquier lugar de la web. En lugar de

[48] DEXI .IO SOLUTIONS {En Línea} {29 de octubre 2020} Disponible en (https://www.dexi.io/solutions/). P.38
[49] DEXI DATA Solutions - ETL Engine 2015. P.38

una persona que rastree manualmente miles de sitios web para recopilar información clave, se pueden construir robots que extraigan automáticamente dicha información para su uso analítico.
Ya sean datos de precios y disponibilidad impulsados por el comercio electrónico, información de reserva en un sitio de viajes o información de la empresa de un directorio de empresas; el robot captura los datos de manera automatizada y de esta manera mientras se extraen dichos datos profundizar con el diseño del método automático.

- Extractores: Los robots extractores son los robots más avanzados de Dexi gracias a que permiten elegir todas las acciones que el robot necesita realizar, como completar formularios, hacer clic en botones y extraer capturas de pantalla.
- Crawler: Los rastreadores son robots menos inteligentes que los extractores, y cuando se inician, solo visitan todos los enlaces que pueden encontrar desde la URL de inicio. Estos robots se controlan configurando una serie de "procesadores de página" que determinan qué acciones debe realizar el rastreador para cada página que visita. Se usa para extraer grandes cantidades de datos fácilmente accesibles e identificables.
- Pipes o Tuberías: Un robot de tubería es un súper robot. Los robots de tubería pueden controlar otros robots y también pueden extraer información externa de API, bases de datos y similares. Los bots de tubería no extraen datos de los sitios web en sí mismos, sino que combinan otros robots, API y conjuntos de datos para crear un único flujo de extracción y procesamiento de datos.
- AutoBot: AutoBots siempre acepta una URL como entrada y luego asigna esa URL a una lista de extractores para una variedad de sitios. Si solicita algo de un AutoBot que no sabe cómo analizar, agregará la URL a su lista interna de sitios. Esto le permite utilizar un único AutoBot para extraer productos de cientos de sitios utilizando URL para los productos individuales.

5.3.3 Anaconda

Anaconda es una distribución de Python para Cálculo Numérico, Análisis de Datos y Machine Learning. Contiene las librerías más usadas por los científicos de datos. Además, hace muy fácil la instalación de otras librerías que se pueda necesitar con comandos cortos y fáciles de implementar. Con Anaconda también es posible crear varios entornos de trabajo en dado caso de estar desarrollando varios proyectos. Esto puede ser útil, como por ejemplo, si uno de los proyectos necesitan diferentes versiones de Python, por ejemplo, Python 3 para un proyecto en específico y Python 2 para algún otro, esta acción es posible gracias a la cantidad de entornos que se

puede configurar en Anaconda, sea desde el “anaconda prompt” por medio de comandos, o simplemente desde la interfaz de “Anaconda Navigator”, en donde ofrece una interfaz intuitiva y fácil de utilizar para de esta manera instalar las librerías de manera más sencilla, de la misma manera con los diferentes entornos que Anaconda ofrece a los programadores, tales como Jupyter Notebook o Spyder. Así mismo como anaconda ofrece la oportunidad de trabajar en los diferentes entornos con distintas versiones de Python, esta distribución da la oportunidad de trabajar en un proyecto que necesite librerías específicas o que tengan una versión específica, por ejemplo, tener un entorno con una determinada versión de TensorFlow o de Numpy. Anaconda es un administrador de paquetes, un administrador de entorno, una distribución de ciencia de datos y una colección de más de 8000 paquetes de código abierto. Anaconda es de uso gratuito, fácil de instalar, y ofrece soporte comunitario gratuito. La forma más fácil de instalar librerías para programación o desarrollo en ciencias de datos es utilizando Anaconda. Inicialmente Anaconda instala librerías básicas, posteriormente se pueden instalar manualmente otras librerías cuando se necesiten.

5.3.4 Spyder

Spyder es un entorno gratuito de programación proporcionado por la distribución de Anaconda, así como otros entornos de programación en Python, Spyder maneja características únicas que le diferencian de otros, con una interfaz comprensible para el programador y funciones para el análisis de gráficas como lo es el perfilador, a diferencia de otros entornos, Spyder tiene un espacio para la visualización de las variables que se están trabajando durante la compilación, aspectos como estos convierten a Spyder como uno de los mejores para la programación en Python.

5.3.5 Python.

Python no sólo es una multiplataforma, sino que también sirve para el desarrollo de cualquier tipo de algoritmo, como lo es en casos de computación, móviles y redes. Para que Python pueda desarrollar dicha programación este lenguaje debe contar con frameworks de gran nivel, dichos frameworks auxilian desde el desarrollo web, hasta el desarrollo de juegos o algoritmos científicos de niveles avanzados[50].
Matplotlib, Basemap y Seaborn son algunos de los plugin o paquetes que tiene Python en su estructura, permitiendo mayor eficacia al momento de generar información y el diseño en sí de los proyectos, además de la organización que ofrece para contar con estos agregados[51].

[50] MARTINEZ, José, “15 Librerías de Python para Machine Learning” {En Línea} {29 de octubre 2020} Disponible en (https://www.iartificial.net/librerias-de-python-para-Machine-learning/) p.40
[51] Ibíd. P. 40

- **Librerías de Python para Cálculo Numérico y Análisis de Datos**

Otra de las fases del proceso de Machine Learning que más tiempo consumen es la preparación de datos y el cálculo de atributos relevantes o características (features). Numpy, SciPy y Pandas son las librerías de Python ideales para análisis de datos y computación numérica.

- **Numpy**

Numpy es una librería numérica que proporciona estructuras de datos universal y funciones matemáticas de alto nivel. Numpy proporciona una estructura de datos universal que posibilita el análisis de datos y el intercambio de datos entre distintos algoritmos. Las estructuras de datos que implementa son vectores multidimensionales y matrices con capacidad para gran cantidad de datos. Además, esta librería proporciona funciones matemáticas de alto nivel que operan en estas estructuras de datos.

- **SciPy**

SciPy es una librería que maneja secuencias numéricas intuitivas y fáciles de implementar, SciPy opera de manera similar con Numpy donde se pueden realizar integraciones, optimización de datos, etc. SciPy es más comúnmente utilizado en el procesamiento de datos[52].

- **Pandas**

Pandas es una librería de Python para la manipulación de datos y el análisis de datos, Pandas es una de las librerías de Python más útiles para los científicos de datos. Las estructuras de datos principales en pandas son Series para datos en una dimensión y DataFrame para datos en dos dimensiones. Estas son las estructuras de datos más usadas en muchos campos tales como finanzas, estadística, ciencias sociales y muchas áreas de ingeniería. Pandas destaca por lo fácil y flexible que hace la manipulación de datos y el análisis de datos[53].

- **Numba**

Numba es una librería más de Python, así como las demás cuenta con funcionalidades y herramientas que facilitan la programación en el ámbito del Deep

[52] MARTINEZ, José. Op, Cit. P.41

[53] Python Data Analysis Library – pandas: Python Data Analysis Library». pandas. {En Línea} {29 de octubre 2020} disponible en (https://pandas.pydata.org/pandas-docs/stable/index.html). P.41

Learning según Larano Gabriel[54], Numba traduce funciones escritas en Python a código máquina optimizado a la hora de ejecutarse. Lo consigue usando el estándar industrial LLVM como librería de compilación. Los algoritmos numéricos compilados con Numba pueden alcanzar velocidades de ejecución tan altas como las de C o FORTRAN.

- **Scikit-learn**

Scikit-learn es una librería de Machine Learning para el lenguaje de programación Python, cuenta con algoritmos para modelos de regresión y es una de las mejores librerías para los científicos de datos y el manejo de los mismos

- **Librerías de Python para Deep Learning**

Teniendo presente que el Deep Learning es una rama del Machine Learning, las librerías de Python para aprendizaje profundo están al mismo nivel. Esto se debe a que últimamente, el Deep Learning es el mayor responsable del incremento en el conocimiento del Machine Learning[55] gracias a su variedad de redes neuronales.

- **TensorFlow**

TensorFlow es una librería de código abierto para Deep Learning y aprendizaje automático, es una librería de Python, desarrollada por Google, para realizar cálculos numéricos mediante diagramas de flujo de datos. Esto puede ser confuso un poco al principio, porque en vez de codificar un programa, se codifica un grafo. Los nodos de un grafo son operaciones matemáticas y las aristas representan los tensores. Con esta computación basada en grafos, TensorFlow puede usarse para Deep Learning y otras aplicaciones de cálculo científico[56].

- **Keras**

Con Keras es una librería fácil de experimentar en el Deep Learning y obtener resultados rápidamente, es una interfaz de alto nivel para trabajar con redes neuronales. Respecto a TensorFlow, la interfaz de Keras es mucho más intuitiva

---

[54] LANARO, Gabriele, Python high performance: build robust application by implementing concurrent and distributed processing techniques (Second edition edition). ISBN 978-1-78728-243-8. OCLC 990086907. {En Línea} {29 de octubre 2020} Disponible en (https://www.worldcat.org/title/python-high-performance-build-robust-application-by-implementing-concurrent-and-distributed-processing-techniques/oclc/990086907). P.42

[55] MARTÍNEZ, José. Op. Cit. P.42

[56] BUHIGAS, Javier, Todo lo que necesitas saber sobre TensorFlow, la plataforma para Inteligencia Artificial de Google 2018, {En línea}, {29 de octubre 2020} Disponible en: (https://puentesdigitales.com/2018/02/14/todo-lo-que-necesitas-saber-sobre-tensorflow-la-plataforma-para-inteligencia-artificial-de-google/). P.42

esta facilidad de uso es su principal característica. Con Keras es más fácil comprobar si las ideas tendrán buenos resultados rápidamente. Keras utiliza otras librerías de Deep Learning (TensorFlow, CNTK o Theano) de forma transparente para realizar el trabajo que se necesite[57].

- **PyTorch**

PyTorch es una librería para Deep Learning en Python, desarrollada por Facebook, que permite el cálculo numérico eficiente con valores de CPU y GPUs. Se puede pensar en PyTorch como una librería que incluye las capacidades de Numpy, pero desde una GPU. En otras palabras, si una tarjeta gráfica tiene un procesador gráfico (por ejemplo, una tarjeta gráfica NVIDIA de última generación), el código se puede ejecutar mucho más rápido, duplicando mucho la velocidad de procesamiento. El aprendizaje profundo (Deep Learning) usa cálculos matriciales y de derivadas masivas y paralelizables en GPUs es por esto que PyTorch también se especializa en Deep Learning.

5.3.6 Jupyter Notebook

Jupyter Notebook es una aplicación web para crear documentos que contienen código, ecuaciones, visualizaciones y texto. Se puede usar Jupyter notebooks para limpiar datos, transformarlos, realizar simulaciones numéricas, modelos estadísticos, visualizaciones de datos, Machine Learning y mucho más. A efectos prácticos es como una consola interactiva de Python en un navegador que permite la ejecución de código Python, visualización de datos y gráficos.

Jupyter Notebook no es una librería, es un entorno web que facilita el desarrollo de programación. Con Jupyter se pueden probar nuevas ideas y posteriormente ver los resultados de forma muy intuitiva, a la vez que dichos resultados pueden documentarse.

5.3.7 Redes Neuronales Artificiales

Como bien se conoce una red neuronal se define como un conjunto de nodos que procesan y envían información entre sí. Dichas redes pueden identificarse con una composición bastante simple o bastante compleja, pues su cantidad de neuronas es independiente y según como este desarrollado o cómo esté orientado el modelo o el algoritmo red neuronal puede estar compuesta por muchas neuronas interconectadas. Ahora bien, se encuentran las redes neuronales artificiales, estas redes están inspiradas en las redes neuronales biológicas, en la cual se representa

[57] MARTÍNEZ, José. Op. Cit. P.43

como ejemplo concreto el cerebro humano. Estas redes están construidas por elementos que se comportan de forma similar a una neurona biológica dado a sus funciones más comunes, su organización se representa de forma similar a la de un cerebro humano.

5.3.8 Redes Neuronales Profundas

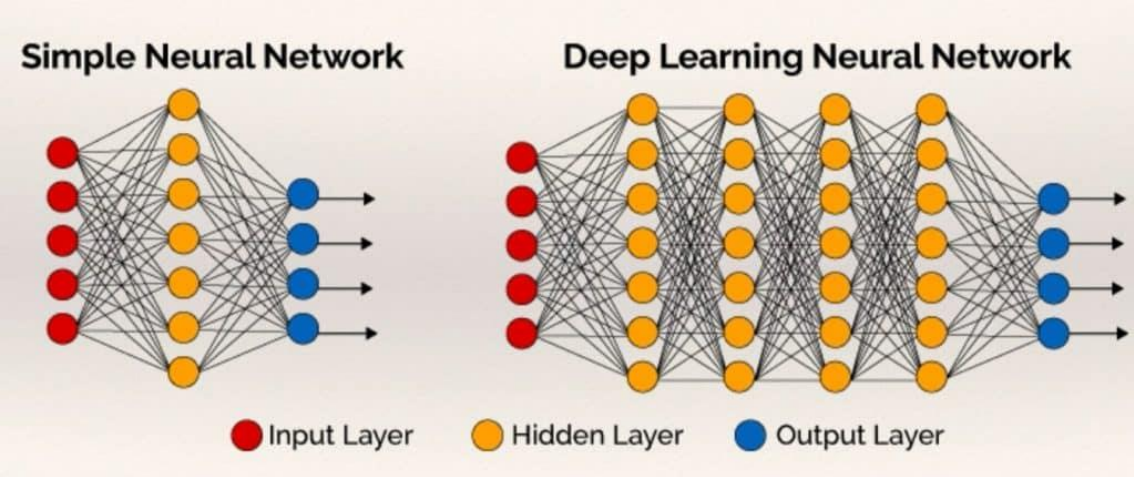

Figura 4, Red Neuronal Profunda, Fuente: https://www.iartificial.net/redes-neuronales-desde-cero-i-introduccion/

Así como se presentan en la figura 4, las redes neuronales artificiales se componen de varias capas ocultas con una **n** cantidad de neuronas a diferencia de una red neuronal simple, esta red neuronal está relacionada a las redes neuronales artificiales, dado que las redes neuronales profundas son una red neuronal artificial con varias capas ocultas entre las capas de entrada y salida, estas redes pueden modelar relaciones NO lineales complejas.

El propósito principal de estas redes radica en recibir un conjunto de entradas, después realizar los cálculos previamente configurados o diseñados en las neuronas y dar una salida para resolver problemas del mundo real como la clasificación, normalmente se está acostumbrado a trabajar con varias neuronas dado a que con más neuronas el trabajo o algoritmo durante su análisis puede desarrollarse de manera más sencilla o menos compleja. Estas neuronas se utilizan comúnmente en el aprendizaje supervisado y en los problemas de aprendizaje por refuerzo. En las redes neuronales profundas se suele comúnmente trabajar con una gran cantidad de capas ocultas, se puede suponer alrededor de unas 2000 capas, esto debido a su sencillez a la hora de analizar los datos entrantes.

5.3.9 Redes Neuronales Convolucionales

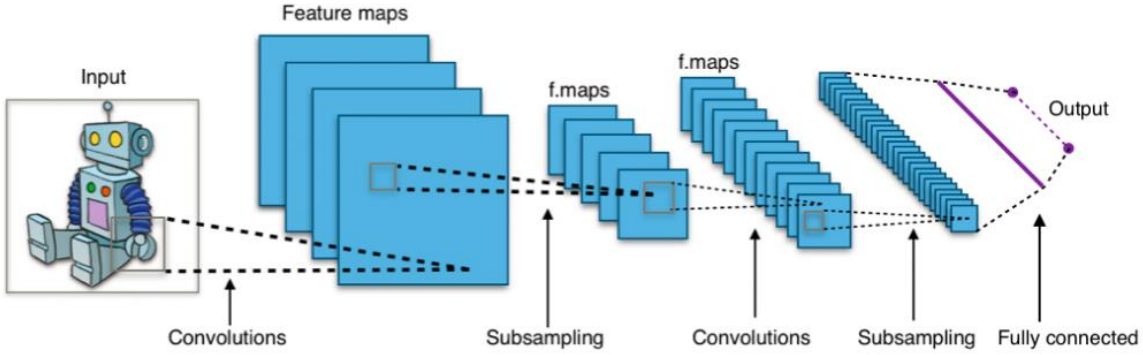

Figura 5, Red Neuronal Convolucional, Fuente: https://pochocosta.com/podcast/redes-neuronales-convolucionales-explicadas/

Así como se presentan las redes neuronales profundas, las Redes Neuronales Convolucionales son un tipo de Red Neuronal Artificial con aprendizaje supervisado que en sus distintas capas procesa de manera similar a la del ojo humano, pues trata de identificar distintas características en las entradas que en definitiva hacen que la red neuronal pueda identificar los datos y procesarlos como esta dicha neurona programada. Como se puede observar en la imagen anterior una característica especial de estas neuronas es que están especializadas para el procesamiento de imágenes, así como se muestra en la figura 5, pues durante su ejecución trabaja con los pixeles de las imágenes en forma de matriz, dichas matrices están constituidas por valores de 0 a 255 pero para la red neuronal se normaliza de 0 a 1. Una red toma como entrada estos píxeles, tomando por valores el ancho y alto de la imagen, vale aclarar que la imagen se puede presentar en diferentes colores, dado esto se presentaran 3 distintos canales RGB, más respectivamente como rojo, verde y azul; estos valores ingresan por las neuronas en la capa de entrada y posteriormente como las demás redes neuronales continúa con su procesamiento de desarrollo según los algoritmos que estas contengan.

5.3.10 Redes Neuronales Recurrentes

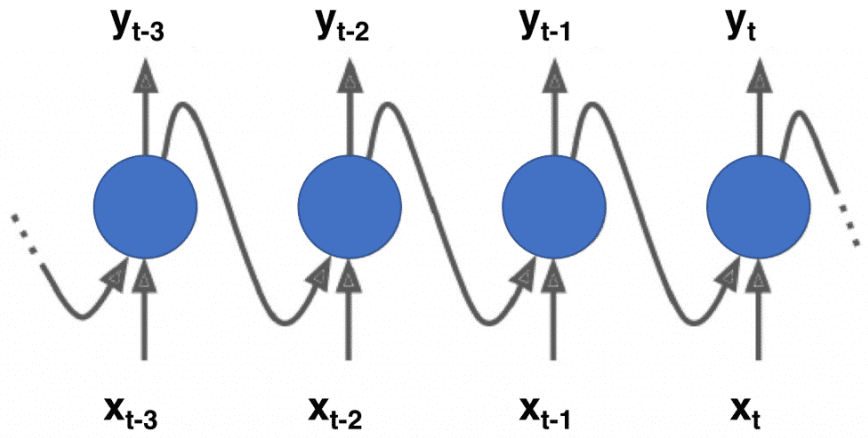

Figura 6, Red Neuronal Recurrente, Fuente: https://torres.ai/redes-neuronales-recurrentes/

De la misma manera como se han presentan distintas métodos, algoritmos y modelos en el Deep Learning, se encuentran las Redes Neuronales Recurrentes, también conocidas como las redes LSTM, estas redes son una clase de redes para el análisis de datos en series de manera temporal, de esta manera permiten el trabajo mediante la dimensión del tiempo, factor poco común en las demás redes neuronales. Durante cada instante de tiempo las neuronas recurrentes reciben distintas entradas de la capa anterior, así mismo su propia salida en un tiempo menor para de esta forma generar una nueva salida, debido al factor tiempo en las neuronas recurrentes, dichas neuronas pueden recibir hasta dos entradas durante el mismo instante de tiempo. Estas cuentan con un conjunto de parámetros, aplicados a la entrada de datos que reciba la capa anterior y también aplicados a la entrada de datos correspondientes al vector salida de la neurona emisora de los datos.

5.3.11 Excel

Excel es un programa informático desarrollado por Microsoft y forma parte de Office, una suite informática que incluye otros programas como Word y PowerPoint[58]. Excel es un programa esencial hoy en día para el análisis y tratamiento de datos gracias a sus numerosas funciones que incluye, es una potente hoja de cálculo que a pesar de sus años aún no se queda atrás, con diferentes licencias de Office, ofrece un set de herramientas para el análisis de datos numéricos o categóricos y que incluso, incluye un espacio para la programación Visual Basic por medio del lenguaje de

[58] ORTIZ Moisés, Que es Excel y para qué sirve, {En Línea], {10 de mayo de 2021} {Disponible en: https://exceltotal.com/que-es-excel/} p.46

macros de Microsoft, por otra parte Excel ofrece la oportunidad de realizar transformación y limpieza de datos de manera manual para el tratamiento de grandes volúmenes de datos, estos datos también son muy utilizados en Excel gracias a las tablas dinámicas que este programa ofrece al análisis de datos, permitiendo ordenar y visualizar los datos de manera más ordenada y con el uso de filtros que permiten el reordenamiento, recuento, suma y más funciones para los datos, para de esta manera, realizar varios análisis y obtener diferentes conclusiones gracias a la lectura de los datos con las tablas dinámicas de Excel.

5.3.12 Regresión Lineal.

Para el modelo de regresión lineal es ideal aplicar el método de mínimos cuadrados ordinarios (MCO), pues es un método donde se encuentran los parámetros poblacionales del modelo. Este método será consistente siempre y cuando se cumplan los 7 supuestos o axiomas para poder llevar a cabo una regresión lineal según el libro de econometría de Gujarati y Porter [59].

- **Supuesto 1. Linealidad en los parámetros:**

El modelo de regresión es lineal en los parámetros, aunque puede o no ser lineal en las variables.

- **Supuesto 2. Valores de X independientes del término error:**

Las variables X entre ellas deben ser independientes con el error, es decir los valores que toma X pueden considerarse fijos en muestras repetidas.

- **Supuesto 3. El valor medio del error es igual a cero:**

Este supuesto establece que el valor de la media de la perturbación que dependen de la variable X dada es cero, este supuesto representa mediante la siguiente gráfica que algunos valores de la variable X y las poblaciones Y asociadas a cada uno de ellos.

[59] GUJARATI Damodar, PORTER Dawn. Econometría: Mínimos cuadrados ordinarios. Quinta Edición. México: Mc Graw Hill, 2010. p.47

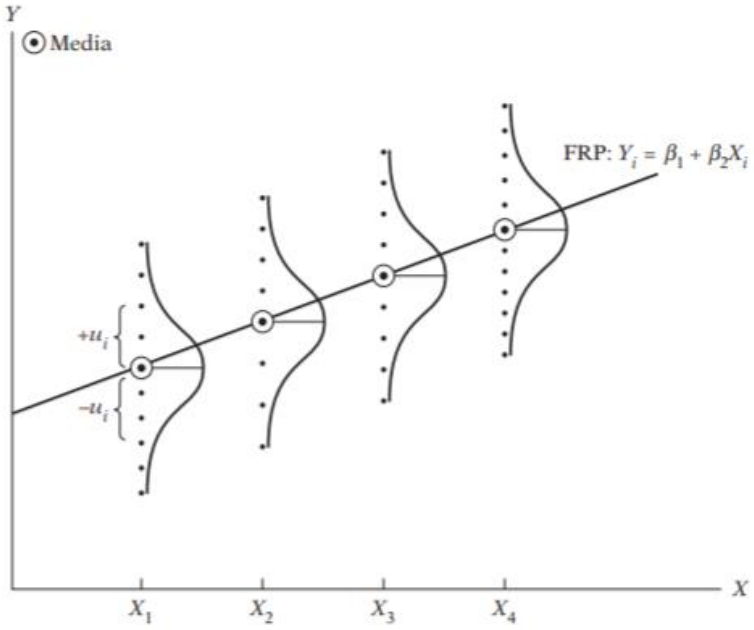

Figura 7 Distribución condicional de las perturbaciones u_i, *Fuente libro: GUJARATI Damodar, Econometría: mínimos cuadrados ordinarios. Quinta edición, México, 63p.*

- **Supuesto 4. Homocedasticidad o varianza constante:**

La varianza del término de error, o de perturbación, es la misma sin importar el valor de X, es decir, en estadística se dice que un modelo predictivo presenta "Homocedasticidad" cuando la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones[60].

- **Supuesto 5. No hay autocorrelación entre las perturbaciones.**

El valor entre dos perturbaciones cualesquiera debe ser cero al no estar correlacionadas, es decir, estas observaciones se muestran de manera independiente.

- **Supuesto 6. El número de observaciones n debe ser mayor que el número de parámetros a estimar:**

De la misma manera, el número de observaciones debe ser mayor que el número de variables explicativas.

[60] Damodar N. Gujarati. Econometría, Op, cit. P.48

- **Supuesto 7. La naturaleza de las variables de X:**

No todos los valores X en una muestra determinada deben ser iguales, es decir las variables de X deben ser un número mayor que cero y además no pueden habar valores atípicos en la variable X, es decir, valores muy grandes en relación con el resto de las observaciones.

5.3.13 Árboles de decisión

Un árbol de decisión es un método analítico que a través de una representación esquemática facilita la toma de decisiones en el uso de resultados y probabilidades asociadas. Estos árboles de decisión se pueden emplear para generar sistemas expertos o búsquedas binarias, una de las ventajas de los árboles de decisión es que reduce el número de variables independientes y proporciona un alto grado de comprensión para la toma de decisiones, los árboles de decisión son demasiado útiles para la toma de una decisión partiendo de un gran conjunto de opciones, muestra las opciones que tenemos y las consecuencias de cada una de estas, está estructurada de una forma visual comprensible, en donde se evidencia todo el proceso para la toma de una decisión.

5.3.14 Random Forest

El Random Forest al igual que el árbol de decisión es un modelo de aprendizaje supervisado para la clasificación, pero también puede ser empleado para la regresión. Al Random forest se le conoce comúnmente como un tipo de “Ensamble” debido a que se combinan diversos árboles y de todos estos al final como salida de cada uno, se toma el resultado más votado el cual será la respuesta del Random forest como se observa en la siguiente figura.

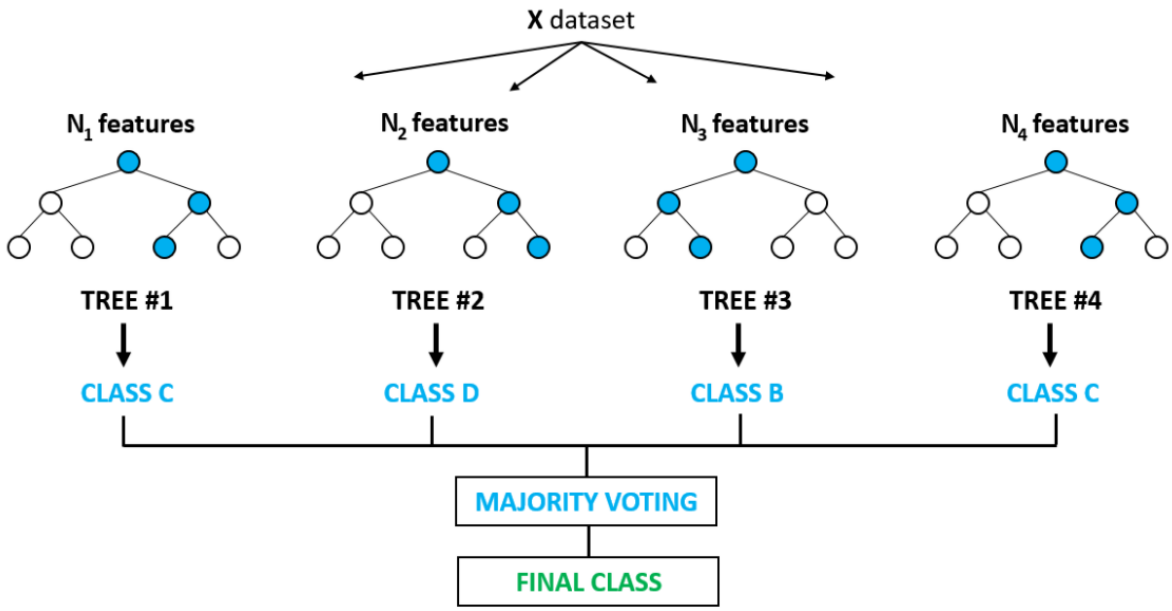

Figura 8 Estructura Random Forest, Fuente: https://rpubs.com/Avalos42/randomforest

Como se evidencia en la figura anterior, se puede observar la estructura de un modelo de Random Forest. Este modelo está compuesto por varios árboles de decisión y al final como resultado o decisión final, se obtiene el resultado más votado de cada uno de los árboles de decisión.

5.3.15 Máquinas de soporte vectorial para la regresión (SVMR)

Según Grajales, las máquinas de soporte vectorial empleadas para la regresión con una mejora a la máquina de soporte vectorial (SVM), la cual utiliza los mismos principios que la SVM para la clasificación, aunque con algunas diferencias debido a que la salida es un número real, entonces se hace complicado predecir la información disponible que tiene infinitas posibilidades[61]. Estas máquinas de soporte vectorial también se pueden evidenciar en uno de los casos de estudio presentados en el estado del arte, en donde, con una buena extracción de características del conjunto de datos objetivo, la máquina de soporte vectorial puede lograr un buen resultado, es importante mencionar que, las máquinas de soporte vectorial presentan una mayor rentabilidad de los problemas de clasificación, es por esto que se presentan estas SVMR las cuales están más adaptadas a la regresión aunque no con el mismo desempeño respecto a un árbol de decisión.

[61] GRAJALES Yuri, Modelo predicción precios viviendas proyecto Medellín, Op, Cit, p.50

5.4 Marco Legal

5.4.1 Leyes

- **Ley 388 de 1997**

“Reglamentada por los Decretos Nacionales 150 y 507 de 1999; 932 y 1337 de 2002; 975 y 1788 de 2004; 973 de 2005; 3600 de 2007; 4065 de 2008; 2190 de 2009; Reglamentada parcialmente por el Decreto Nacional 1160 de 2010”

“*Por la cual se modifica la Ley 9 de 1989, y la Ley 2 de 1991 y se dictan otras disposiciones*”

Esta ley se conoce como la ley de desarrollo territorial (diferente a ordenamiento territorial), que establece un mandato para que todos los municipios del país formulen sus respectivos Planes de Ordenamiento Territorial[62].

Dentro de los objetivos de esta ley se encuentran[63]:

- Armonizar y actualizar las disposiciones contenidas en la Ley 9 de 1989 con las nuevas normas establecidas en la Constitución Política, la Ley Orgánica del Plan de Desarrollo, la Ley Orgánica de Áreas Metropolitanas y la Ley por la que se crea el Sistema Nacional Ambiental.
- El establecimiento de los mecanismos que permitan al municipio, en ejercicio de su autonomía, promover el ordenamiento de su territorio, el uso equitativo y racional del suelo, la preservación y defensa del patrimonio ecológico y cultural localizado en su ámbito territorial y la prevención de desastres en asentamientos de alto riesgo, así como la ejecución de acciones urbanísticas eficientes.
- Garantizar que la utilización del suelo por parte de sus propietarios se ajuste a la función social de la propiedad y permita hacer efectivos los derechos constitucionales a la vivienda y a los servicios públicos domiciliarios, y velar por la creación y la defensa del espacio público, así como por la protección del medio ambiente y la prevención de desastres.

Estos objetivos son claves para desarrollo urbanístico de las ciudades ya que representan los usos y los objetivos que se debe de tener dentro de una ciudad a la hora de definir donde pueden estar ubicadas las viviendas, los usos de suelo de

[62] CAMARA DE COMERCIO DE BOGOTÁ, Como ubicar tu empresa en Bogotá desde el punto de vista administrativo, {En línea} {9 de mayo 2021}, disponible en (http://recursos.ccb.org.co/ccb/pot/PC/files/ley388.html#:~:text=En%20el%20a%C3%B1o%201997%20el,respectivos%20Planes%20de%20Ordenamiento%20Territorial.) p.51

[63] Ley 388 de 1997.Op. Cit . P.51

cada sector de dicha ciudad entre otras cosas, lo cual a la hora de realizar avalúos influye en el valor dado por el perito al tener en cuenta dicha ley.

- **Ley 1266 de 2008**

"*Por la cual se dictan las disposiciones generales del hábeas data y se regula el manejo de la información contenida en bases de datos personales, en especial la financiera, crediticia, comercial, de servicios y la proveniente de terceros países y se dictan otras disposiciones"*

En este proyecto, se tiene en cuenta que existe la ley en la cual se regula el manejo de la información por parte más que todo de entidades financieras, sin embargo, los datos que se obtienen a lo largo de este proyecto no compromete la integridad de las personas, ya que son publicados en plataformas públicas y no son datos personales de alguna persona en específico, sino información que la gente está dispuesta a compartir para la venta de sus inmuebles

- **Ley 1581 de 2012**

"*Por la cual se dictan disposiciones generales para la protección de datos personales."*

Similar a la ley del Habeas Data, en este proyecto no se vulnera la información ni los datos privados de los individuos que hayan publicado sus viviendas a la venta dentro de las páginas web utilizadas, ya que estas personas al haber publicado sus viviendas ahí son conscientes de que la información es de manejo público y cualquiera puede acceder a ella, es responsabilidad del ofertante saber qué información puede publicar y cual no.

5.4.2 Decretos

- **Decreto 190 del 2004**

"*Por medio del cual se compilan las disposiciones contenidas en los Decretos Distritales 619 de 2000 y 469 de 2003"*

Este decreto compila los Decretos 619 de 2000 por el cual expidió el Plan de Ordenamiento Territorial, y 469 de 2003, por el cual se Revisó el Plan de Ordenamiento Territorial de Bogotá[64].

Este decreto es utilizado por los peritos ya que en él se describen los usos del suelo y las unidades de planeamiento zonal (UPZ) dentro de la ciudad de Bogotá, las

[64] ALCADÍA DE BOGOTÁ. Documentos para PLAN DE ORDENAMIENTO TERRITORIAL :: Estatutos Orgánicos. {En línea}. {10 de mayo 2021}. Disponible en: (https://www.alcaldiabogota.gov.co/sisjur/listados/tematica2.jsp?subtema=21178#:~:text=El%20Decreto%20Distrital%20190%20de,es%20decir%2C%20no%20ha%20sido) p.52

cuales agrupan diferentes barrios en cada una. Esta información tiene que ser tenida en cuenta para poder determinar el valor de una casa ya que dependiendo de donde está ubicada y como es la construcción puede aportar información de valor para definir el precio de la misma.

- **Decreto 159 del 2004**

“"*Por el cual se adoptan normas urbanísticas comunes a la reglamentación de las Unidades de Planeamiento Zonal.”*

Este decreto adopta normas urbanísticas comunes a la reglamentación de las Unidades de Planeamiento Zonal –UPZ. Señala disposiciones sobre densidad, habitabilidad, equipamiento comunal privado, estacionamientos, usos, uso dotacional, antejardines, retrocesos, sótanos, semisótanos, rampas, escaleras, voladizos, alturas, aislamientos, edificaciones permanentes, adosamiento, pareamiento, englobe, tratamiento de consolidación, modalidad urbanística, normas volumétricas, cerramientos, construcciones provisionales, tratamiento de mejoramiento integral, subdivisiones, obras nuevas, adecuaciones, ampliaciones y modificaciones[65].

Los peritos avaluadores utilizan este decreto para visualizar a mayor profundidad los detalles que deben ser tenidos en cuenta en cada construcción de vivienda, dependiendo de su ubicación dentro de las unidades de planeamiento zonal, por tal motivo es consultado por los peritos para ver si las viviendas cumplen con dichas reglamentaciones

- **Decreto Distrital 551 de 2019, Estratificación en Bogotá**

“*Por medio del cual se adopta la actualización de la estratificación urbana de Bogotá D.C. para los inmuebles residenciales de la ciudad”*

Para el área urbana del Distrito Capital está determinada por el Decreto Distrital 551 de 2019. Según el Departamento Nacional de Estadística -DANE[66]- ¨La

[65] ALCADÍA DE BOGOTÁ. Documentos para POT UNIDADES DE PLANEAMIENTO ZONAL -UPZ- :: Reglamentación. {En línea}. {10 de mayo 2021}. Disponible en: (https://www.alcaldiabogota.gov.co/sisjur/listados/tematica2.jsp?subtema=21291&cadena=p#:~:text=Decreto%20159%20de%202004%20Alcald%C3%ADa,Unidades%20de%20Planeamiento%20Zonal%20%E2%80%93UPZ.&text=Relaci%C3%B3n%20de%20las%20Unidades%20de,uno%20de%20los%20decretos%20respectivos.) p.53

[66] Departamento Administrativo Nacional de Estadística – DANE. Estratificación socioeconómica. {En línea}, {9 de mayo 2021}. Disponible en (https://www.dane.gov.co/index.php/69-espanol/geoestadistica/estratificacion/468-estratificacion-socioeconomica). P.53

estratificación socioeconómica es el mecanismo que permite clasificar la población en distintos estratos o grupos de personas que tienen características sociales y económicas similares, a través del examen de las características físicas de sus viviendas, el entorno inmediato y el contexto urbanístico o rural de las mismas". Los municipios y distritos pueden tener entre uno y seis estratos, dependiendo de la heterogeneidad económica y social de sus viviendas. Bogotá se clasifica en seis (6) estratos. Esta herramienta de focalización del gasto se emplea para cobrar los servicios públicos domiciliarios con tarifas diferenciales por estrato y para asignar subsidios y contribuciones a los hogares en esta área. De esta manera, quienes tienen más capacidad económica pagan más por los servicios públicos y contribuyen para que los hogares de estratos bajos puedan pagar sus tarifas[67].

La estratificación en Bogotá no solamente sirve para lo anteriormente mencionado, si no que resulta un factor importante a la hora de realizar un avalúo comercial sobre un inmueble, ya que como bien se describe en la definición se tienen en cuenta las características físicas de las viviendas que es un factor relevante para definir el precio de la vivienda, el entorno inmediato y el contexto urbanístico también son factores relevantes dentro del estrato también lo son a la hora de realizar un avalúo, siendo de esta manera el estrato fundamental a la hora de realizar los procesos de avalúos por la información que proporcionan.

5.4.3 Resoluciones

- **RESOLUCIÓN 620 DE 2008**

Dentro del estado colombiano, existe una normativa específica para la realización de los avalúos del país, esta se encuentra dentro de la resolución 620 de 2008 del IGAC (Instituto Geográfico Agustín Codazzi) Por la cual se establecen los procedimientos para los avalúos ordenados dentro del marco de la Ley 388 de 1997.

Esta resolución se encuentra dividida en 38 artículos separados en 7 capítulos en donde se describen desde las definiciones de los distintos métodos para realizar los avalúos, como lo son el método por comparación de mercado o el método residual; las distintas etapas para tener en cuenta a la hora de realizar un avalúo hasta las fórmulas necesarias en el proceso.

En ansías por realizar el presente trabajo lo más apegado a la normatividad vigente, se busca implementar una de las técnicas descritas en el mencionado documento, la comparación de mercado, la cual consiste en establecer el valor comercial del bien, a partir del estudio de las ofertas o transacciones recientes, de bienes semejantes y comparables al del objeto de avalúo. Tales ofertas o transacciones

[67] SECRETARIA DISTRITAL DE PLANEACIÓN, Decreto Distrital 551 de 2019, Op. Cit. P.54

deberán ser clasificadas, analizadas e interpretadas para llegar a la estimación del valor comercial[68].

Dada la descripción anterior este método es el que mejor se adapta para el objetivo que se quiere realizar en este trabajo, ya que se orienta a la comparación entre distintas ofertas dentro de la misma zona, cosa que sería posible de implementar si se utiliza una arquitectura de redes neuronales para desarrollar el método, además que tener los datos del mercado, puede resultar sencillo, teniendo en cuenta que en el país se encuentran con una gran variedad de páginas web en donde encontrar este tipo de información.

[68] INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI SEDE CENTRAL. Op. Cit, p.55

5.5 Estudio de Benchmarking

A continuación, se presentan distintas aplicaciones que se encuentran en el mercado colombiano para calcular el valor de un avalúo comercial en Bogotá. La mayoría de estas aplicaciones utilizan un proceso similar al que se quiere plantear en el presente proyecto, pues según unas características dadas por el usuario, se utiliza inteligencia artificial para determinar el precio del avalúo del inmueble. Las principales características que suelen resaltar estas aplicaciones para realizar este proceso son el uso de inteligencia artificial y Big data, aunque no explican cómo las utilizan o qué técnicas usan específicamente. Además, utilizan las funciones de Google Maps para poder ubicar de manera más fácil los inmuebles, algunas pueden ser más específicas en las características que piden. Muchas de estas aplicaciones son gratuitas, aunque existen otras aplicaciones que, a pesar de que utilizan tecnologías similares, cobran para obtener el resultado del avalúo.

Vale la pena aclarar, que no son muchos los portales que ofrecen al cliente la posibilidad de realizar avalúos en línea por él mismo, por el contrario, gran parte de las páginas web ofrecen en su sistema la opción “contáctenos” para el momento de solicitar un avalúo.

5.5.1 Properati

Properati es una plataforma web que trabaja con propiedades y que tiene como objetivo cambiar la forma en que se venden y arriendan los inmuebles en Latinoamérica. Aquellas personas que buscan un nuevo hogar o que busquen inversiones en propiedades pueden acceder a las que están registradas en Properati para una mejor toma de decisiones. Además de esto, Properati cuenta con valiosa información y herramientas como la realización de promedios de precios. Esta herramienta para la predicción de precios estimados surge de algoritmos desarrollados en Properati que consideran características esenciales de un inmueble, tales como el área en metros cuadrados, la cantidad de habitaciones, baños o el nivel de estrato. Así, un rango de precios más alto ayuda a indicar una mayor variabilidad de precios en la zona donde se está ubicada la propiedad cuyo precio se desea estimar. La interfaz de Properati se puede evidencia en la siguiente imagen (Figura 9).

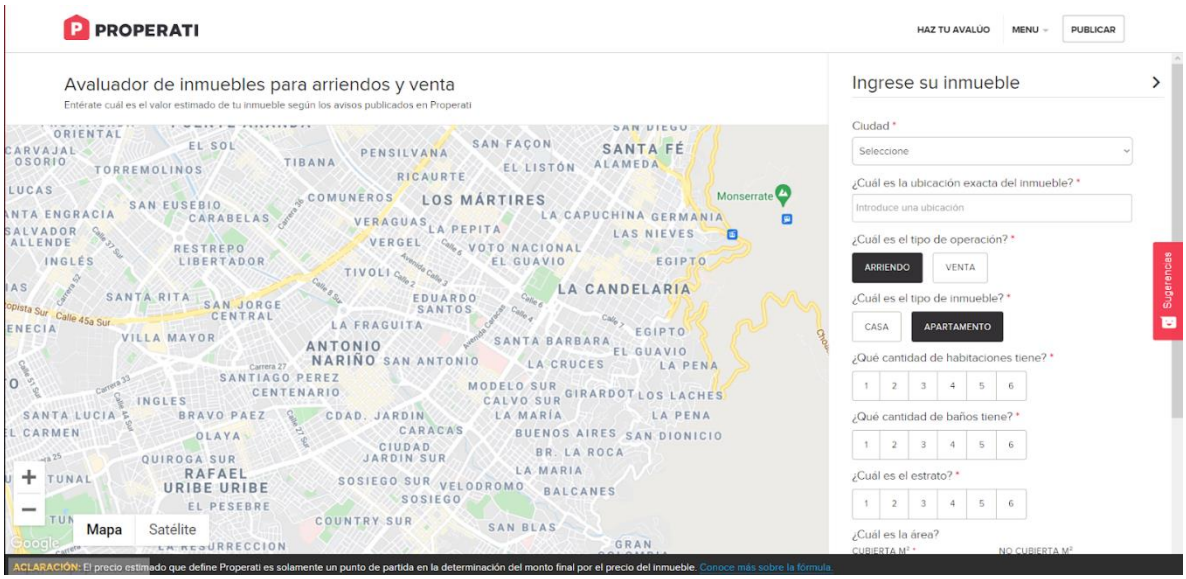

Figura 9, Interfaz página web Properati

5.5.2 Finco

Finco.co, es una plataforma de datos inmobiliarios que ayuda a la toma de decisiones inmobiliarias, sea para la venta, compra, arrendamiento e inversiones. Finco estima en pocos segundos el precio de venta, metro cuadrado y arriendo en cualquier apartamento de Bogotá, explicando los argumentos que validan ese precio, como, por ejemplo, las características deL sector del inmueble valuado. Esto quiere decir, que Finco no es una herramienta que solo ayuda a estimar el precio de una vivienda, sino que muestra también las características que justifican los precios predichos, características tales como la seguridad de la zona, las zonas verdes, centros comerciales, etc.

Finco lanzó esta plataforma de información para el mercado inmobiliario el 20 de octubre del año 2020, la cual utiliza herramientas de Machine Learning y ciencia de datos y cuenta con más de 50 fuentes de información y cientos de millones de datos únicos, los cuales, mediante procesos de ETL, limpia, transforma y modela para hacerlos entendibles. Es de esta manera como Finco ofrece una herramienta eficaz para los usuarios para tener una idea del precio de venta de una casa o apartamento, junto con una justificación del porqué ese precio. Finco por el momento cuenta con una cobertura para casas y apartamentos en la ciudad de Bogotá, Colombia, pero se está desarrollando para cubrir prontamente lotes, oficinas, bodegas y más ciudades dentro del país. La interfaz de Finco se puede evidenciar en la siguiente imagen (Figura 10).

Figura 10, Interfaz página web Finco, **Fuente: https://www.finco.co/?queryStep=1.**

5.5.3 OIKOS

Oikos empezó en el mercado inmobiliario buscando ofrecer a los clientes un servicio integral y de esta forma lograr posicionarse como una empresa especializada en la administración y comercialización de inmuebles. Cuenta con un portafolio de gestión en *Finca Raíz* y servicios como avalúos inmobiliarios, manejo de propiedad horizontal, comercial e industrial. Oikos es una inmobiliaria estructurada, la cual ofrece la oportunidad de realizar el avalúo de un inmueble completamente en línea, obteniendo como resultado un archivo PDF del avalúo realizado al inmueble.

El proceso que realiza el simulador de avalúo de vivienda o avalúo comercial de propiedad comienza con la selección de un perfil de usuario, siendo persona natural, jurídica, o empresa inmobiliaria, seguido a esto se completa un formulario del inmueble a realizar el avalúo. Oikos cuenta con dos tipos de avalúos, un avalúo avanzado el cual se expide un PDF con la información del inmueble y el avalúo certificado, en el cual un perito certificado por la RNA (registro nacional de avaluadores) visitará el inmueble y se le enviará en físico el avalúo firmado del inmueble. La interfaz de la página web de Oikos en su sección avalúos se puede evidenciar en la siguiente imagen (Figura 11).

Figura 11, Interfaz página web Oikos, ***Fuente: https://www.finco.co/?queryStep=1***

5.5.4 Avalúo en línea

Avalúo en línea es una plataforma creada por Jaime Miranda y Alejandro Gaviria, nació después de una larga trayectoria de análisis de inteligencia de datos y algoritmos matemáticos para los usuarios de inmuebles en la ciudad de Bogotá. Esta herramienta permite a los colombianos acceder a un servicio de avalúos en línea donde las personas pueden conocer el valor real de su inmueble, sin la necesidad de agendar visitas de peritos. Así mismo, se reconoce que un avalúo puede ser un proceso demorado, costoso y subjetivo, lo cual justifica la creación de Jaime Miranda y Alejandro Gaviria de su la plataforma de avaluoenlinea.com, el cual mediante el uso del poder de la Data y de la tecnología permite estimar el valor comercial de un inmueble de manera objetiva a partir de las características y el entorno en el que se ubica la propiedad. La interfaz de la página web de Avalúo en línea se puede observar en la siguiente imagen (Figura 12).

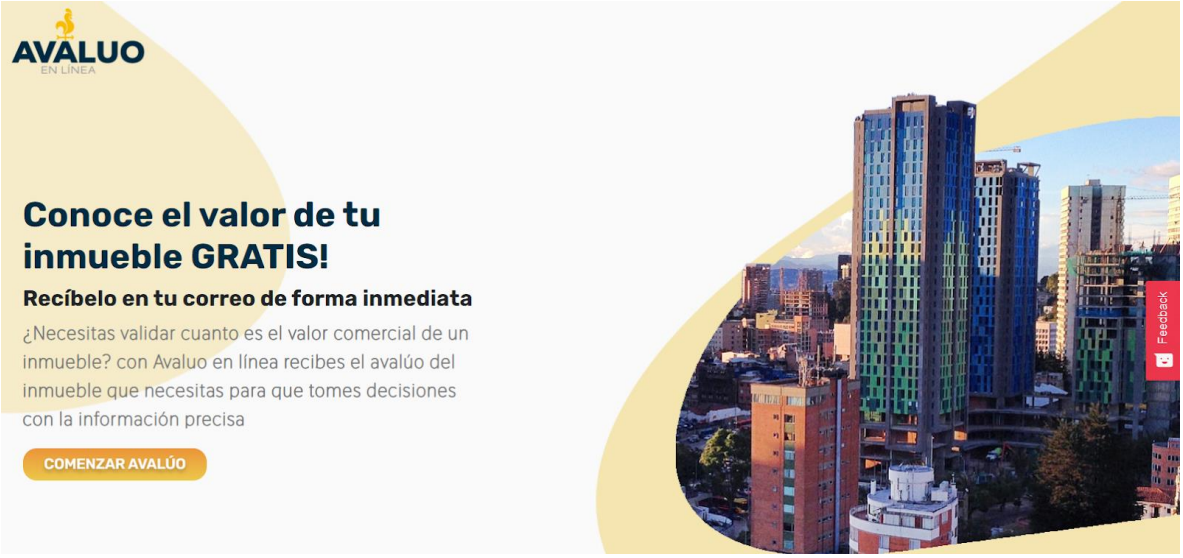

Figura 12, Interfaz página web Avaluoenlinea, , Fuente: https://www.avaluoenlinea.com

5.5.5 HABI

Así como las plataformas previamente mencionadas, Habi se suma a las herramientas tecnológicas para la estimación de precios de vivienda en Colombia. Pero cuenta con características que la hacen una plataforma orientada al sector inmobiliario de una manera distinta, pues esta plataforma compra las viviendas como casas o apartamentos que son publicadas u ofertadas por los usuarios, de esta manera Habi las remodela y las ofrece a la venta a precios competitivos.

Ahora bien, Habi maneja una herramienta tecnológica para la estimación de precios de vivienda, cuenta con un algoritmo diseñado para analizar más de 25 variables en una propiedad en venta. A través de esta herramienta se compran viviendas y se remodelan con todas las características que ofrece el mercado, dando como resultado un precio basado en la competencia del inmueble. Con esta herramienta los usuarios pueden saber el valor comercial, catastral, de arriendo y mucho más de un inmueble con lo que se llega a una descripción muy parecida a un avalúo comercial, pero de manera virtual en línea. La interfaz de la página web de Habi se puede evidenciar en la siguiente imagen (Figura 13).

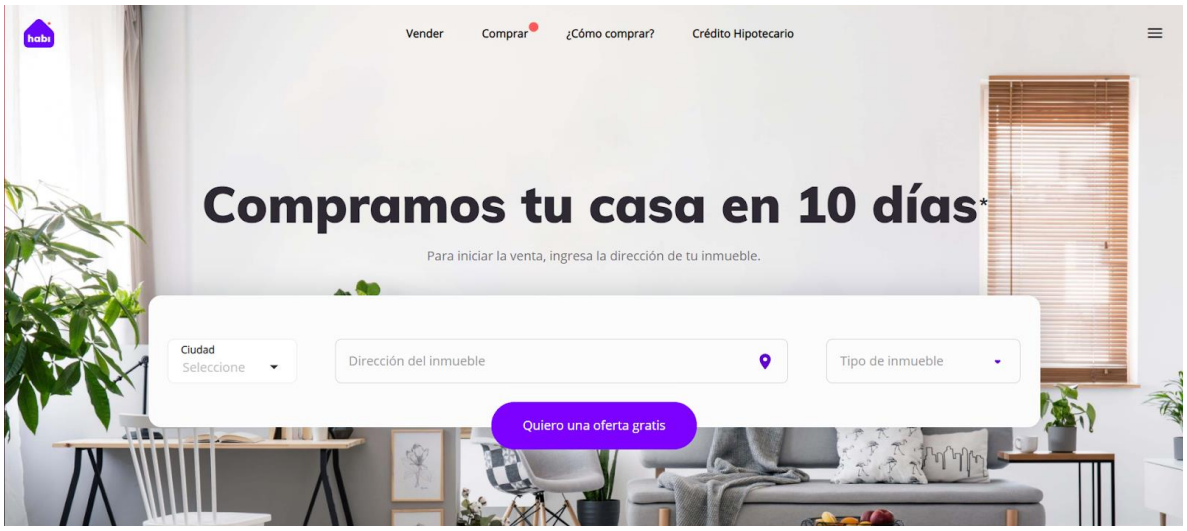

Figura 13, Interfaz página web Habi , *Fuente: https://www.habi.co*

6 ALCANCES Y LIMITACIONES

Este proyecto se enfocará únicamente en la predicción de avalúos comerciales de vivienda, dado que los avalúos catastrales tienen otros fines diferentes a los que se buscan en este proyecto. Por otra parte, el área de estudio será únicamente en la ciudad de Bogotá, debido a que es la ciudad en la cual mayores búsquedas de compra y venta en el país se realizan, teniendo facilidad para la obtención de los datos. Por último, con respecto a temas inmobiliarios, los inmuebles a tener en cuenta durante el estudio serán únicamente apartamentos y casas, dado que realizar la predicción para otro tipo de inmuebles, como lo son los lotes, requieren de otro tipo de técnicas que difieren a las de aplicar en los inmuebles anteriormente descritos, por eso se prefiere trabajar con estos inmuebles, además, estos son los más buscados por las personas.

Igualmente, en el área de desarrollo del método, se deben tener en cuenta ciertas consideraciones, como la cantidad de modelos a desarrollar según las arquitecturas y algoritmos encontrados, para este trabajo se seleccionarán, trabajarán y compararan 3 modelos, en dado case que se considere, se estudiará un 4 modelo.

Por último, este proyecto solo presenta el método que mejor se adapta a la naturaleza de los datos de Bogotá, no se realizará ningún desarrollo de software con el método obtenido.

7 METODOLOGÍA

Para la implementación de este método automático se necesitó conocer cuál es el proceso que los peritos realizan con los avalúos comerciales, uno de los métodos registrados en la resolución 620 de 2008 del IGAC, es el método de comparación de mercado, el cual consiste en comparar el inmueble a elección con diferentes ofertas similares en el mismo sector para poder determinar su precio, basado en esta premisa, para la realización de este proyecto se presenta una metodología diseñada a partir de los procedimientos comúnmente utilizados en aprendizaje de máquina, teniendo en cuenta los procesos de realización de los avalúos comerciales, siendo dividida en un total de 6 fases como se muestra a continuación:

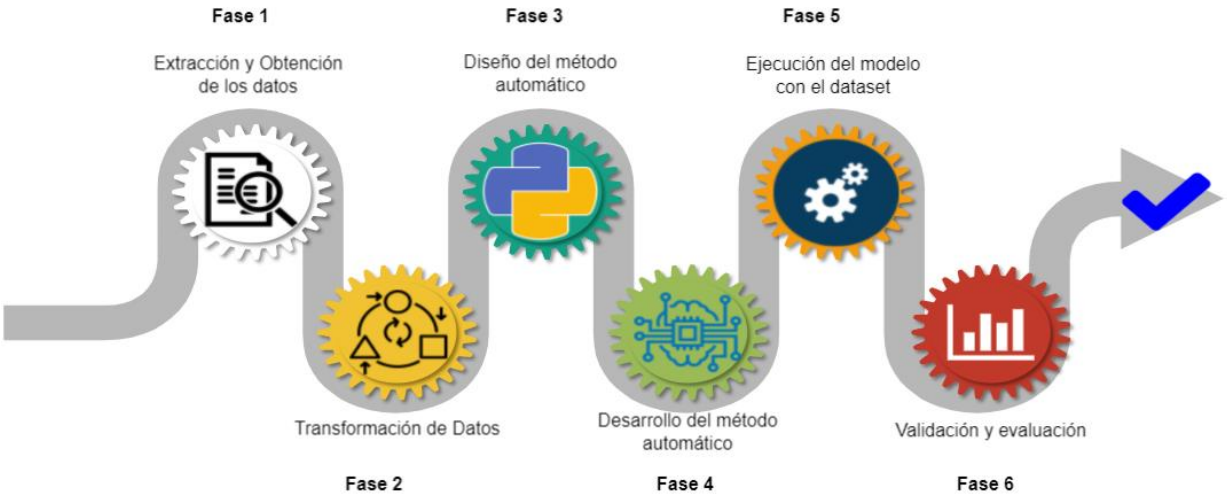

Figura 14, Estructura de la metodología a utilizar, Fuente: Propia.

7.1 Fase 1: Extracción y obtención de los datos.

Seleccionar las páginas web inmobiliarias según la información de vivienda que estas ofrezcan y de esta manera con las que se va trabajar e identificar una técnica o software que permita la captura de los datos de una web seleccionada y finalmente realizar la configuración del software para la extracción.

7.2 Fase 2: Transformación de los datos.

Analizar el archivo con la información exportada por el software de raspado web y transformar los datos mediante la búsqueda de patrones matemáticos para una organización y eliminación de información sobrante y posteriormente verificar que el dataset sea adecuado para su uso en el entretenimiento de los algoritmos.

7.3 Fase 3: Diseño de los métodos automáticos.

A partir de la naturaleza de las variables del Dataset y de los algoritmos seleccionados en trabajos similares en la literatura científica, se realizó el diseño del método automático para cada uno de ellos.

7.4 Fase 4: Desarrollo de los métodos automáticos.

Con los diseños realizados en la fase anterior se procede a buscar las librerías en Python en el entorno Spyder para posteriormente hacer la codificación de cada uno de los métodos y de la misma manera se codifican para generar las medidas de desempeño.

7.5 Fase 5: Ejecución del modelo con el Dataset.

Una vez codificados los modelos se procede a su ejecución con el Dataset ya funcional.

7.6 Fase 6: Validación y Evaluación

Después de ejecutar el Dataset en los modelos, se recogen las medidas de desempeño y se realiza el análisis comparativo de las mismas para seleccionar el mejor modelo.

8 CONSTRUCCIÓN DEL DATASET

A continuación, en esta fase se describirán todos los pasos que se tuvieron en cuenta para la construcción del Dataset en este proyecto.

8.1 Selección de la fuente de los datos.

De la variedad de páginas web que se pueden encontrar en el internet para el negocio de mercado inmobiliario se seleccionó el portal web de *Finca Raíz* por su gran catálogo de viviendas que tiene registradas en su base de datos. A diferencia de otras páginas web, *Finca Raíz* ofrece al cliente una interfaz amigable y con toda la información necesaria que hay que conocer para la compra y venta de viviendas, además. Además, *Finca Raíz* cuenta con un moderno sistema de filtrado categórico para una búsqueda más exacta como la búsqueda de casas y apartamentos en la ciudad de Bogotá. Aspectos como estos hacen que fincaraiz.com sea el portal web elegido para la extracción y recolección de datos para este proyecto.

8.2 Extracción de los datos

Para la construcción del Dataset se trabajó con la herramienta de rastreo web o extracción de datos para profesionales DEXI.IO, se adquirió una licencia por un mes a costo de 119 dólares (USD). Gracias a esta suscripción se tuvo acceso al soporte de 24 horas de dexi y con su ayuda se comenzó la construcción del robot. Dicha construcción partió del direccionamiento URL, es decir, se comienza ingresando a la URL objetivo, para que el robot realice la extracción de datos. En este caso, se ingresó la URL de *Finca Raíz,* pero modificada, para que una vez dentro se puedan visualizar ya las casas y apartamentos que están en venta y filtrados de tal manera que solo se presentan opciones en la ciudad de Bogotá. Acto seguido en el algoritmo se encuentran funciones para hacer clic sobre elementos o “Click Element”, estos Click están programados para que una vez se ingrese a la página se dé clic en los letreros emergentes que salen, como el de aceptar las cookies del navegador y otro más para aceptar los derechos de datos de la página de *Finca Raíz*. Sin embargo, la aparición de estos letreros emergentes no es siempre segura durante cada ejecución, puesto que son mensajes aleatorios que se presentan una vez cargue o se actualice la página.

Una vez accedida a la URL, sin mensajes emergentes y cuando solo se visualicen los datos de las diferentes casas y apartamentos que estén publicados en *Finca Raíz* y filtrados para la ciudad de Bogotá, entonces, el paso a seguir será configurar el “Page Iteration”, que como su nombre lo indica, repetirá una secuencia de páginas. Como se podrá visualizar en la página de *Finca Raíz*, se encuentran alrededor de 16000 a 20000 datos entre apartamentos y casas y solo por cada página se encuentran alrededor de 50 datos, entonces, gracias a la funcionalidad

del “Page Iterator” el robot accederá a las siguientes páginas de manera repetitiva, una vez termine con las demás funciones a completar dentro de la primera página y así de manera sucesiva. Ahora bien, con la página visualizada y con el iterador de páginas configurado, se agregaron dos funciones usando “Click Element” nuevamente, con la funcionalidad de cerrar las ventanas emergentes, solo con la diferencia que estas dos funciones se repetirán nuevamente una vez se actualice la página o se cambie a la siguiente nueva página.

Después de estas dos funciones se encontró una gran función que abarca todas las ofertas que se pueden visualizar en la página de *Finca Raíz*, esta es la función “Loop through elements” que se encarga de seleccionar todos los nombres de las viviendas que se encuentran publicadas y, una vez seleccionadas se continuará con otras funciones. Esta función representa el núcleo del robot, puesto que es donde estarán todas las funciones para realizar la extracción o captura de datos, todo esto teniendo en cuenta que la función “Loop through elements” se contiene dentro de la función del iterador de páginas. Además, dentro de la función de “Loop through elements” se comenzará con una función de “Click Element”, esté clic será la función encargada de acceder a cada una de las viviendas y para de esta manera acceder a su información principal y todas sus características que *Finca Raíz* ofrece como información para los clientes y la cual es el objetivo principal de extracción de cada vivienda. Debe decirse que este procedimiento no estuvo exento de problemas, ya que después de realizar ejecuciones se observó que el robot fallaba luego de cierto tiempo. Con el soporte técnico de Dexi se encontró que el robot fallaba porque la página tardaba mucho tiempo en ingresar a la página de la vivienda, pero, gracias al soporte técnico de Dexi, se implementó una nueva funcionalidad después de este primer Click, la función denominada “Wait for page load”, la cual ayudó a que el robot esperará más tiempo del predeterminado hasta que la página cargará por completo.

Una vez accedida a la página de la vivienda donde se encuentra toda su información de calidad, se procede con la extracción de los datos, esto se realizó en un orden específico gracias a la función de extracción de dexi “Extract Here”. Entonces, con la página visualizada, simplemente, se aplicó dicha función en cada una de las secciones de interés, que son: El Nombre, el precio, si es una vivienda usada, el tamaño, el número de habitaciones, de baños, de parqueaderos, el área construida, el precio por metro cuadrado, el precio de la administración en caso de los conjuntos residenciales, el estrato, la antigüedad, el número de piso en caso de ser una vivienda de propiedad horizontal, el sector, la descripción general, las características interiores, exteriores y del sector de la vivienda. Siendo todas estas las características que se pueden observar en las viviendas que están publicadas en *Finca Raíz*, entre todas estas variables, sin importar si son datos numéricos o cadenas de texto, el robot ayudó a la extracción de cada dato y una vez finalizada

la extracción de esta vivienda, la función “Loop through elements” detectó automáticamente la siguiente vivienda que estaba en la página. Esta página, al estar definida por el page iterator, al completar este ciclo de viviendas en una página, continuó con las siguientes páginas de manera sucesiva hasta lograr un tope de datos. Estas ejecuciones se realizaron más de una vez para lograr la extracción de un gran volumen de datos hasta que el robot falló por algún motivo.

Durante el mes de suscripción se realizó la extracción de datos con la ayuda del robot programado personalmente que previamente fue explicado, también, se preparó para la extracción de los datos del sitio web *Finca Raíz*, portal web líder para la búsqueda de inmuebles en cualquier ciudad de Colombia, pero centrándose en la ciudad capital Bogotá D, C. Para la extracción se manejaron un total de 18 parámetros o características que para *Finca Raíz* son las más esenciales al momento de describir una casa o apartamento, tales como el área de la vivienda, el número de habitaciones, baños y parqueaderos. Con parámetros como estos el robot estaba programado para la captura de dichas características iterando y explorando cada casa y apartamento que estaba registrado en la página de FincaRaiz.com. De esta manera, durante el mes de licencia con la herramienta Dexi.io el robot logró la captura de 3174 datos en los que se encontraban datos de apartamentos y casas con las 18 características y por otro lado se logró la captura o extracción de 4000 datos de apartamentos solo con las 8 características principales que para fincaraiz.com son las que están a primera vista del usuario, tales como el nombre, el precio, si es un apartamento usado o no, su tamaño en metros cuadrados, y el número de habitaciones, baños y parqueaderos. Una vez con los datos a disposición gracias a la herramienta Dexi, estaba completa la extracción de datos, lo que traía como siguiente paso la transformación de dichos datos mediante los procesos de transformación, aplicando una serie de funciones sobre los datos ya extraídos para así tener los datos limpios y de manera organizada.

8.3 Transformación de los datos

Para la transformación de los datos lo primero que se realizó fue un análisis en Excel sobre los datos recibidos, se utilizó el archivo Excel generado por parte de la herramienta de DEXI y se escogió el que traía más información de cada vivienda pero que tenían una menor cantidad de registros de vivienda. A partir de aquí se siguieron los siguientes pasos para la transformación de los datos:

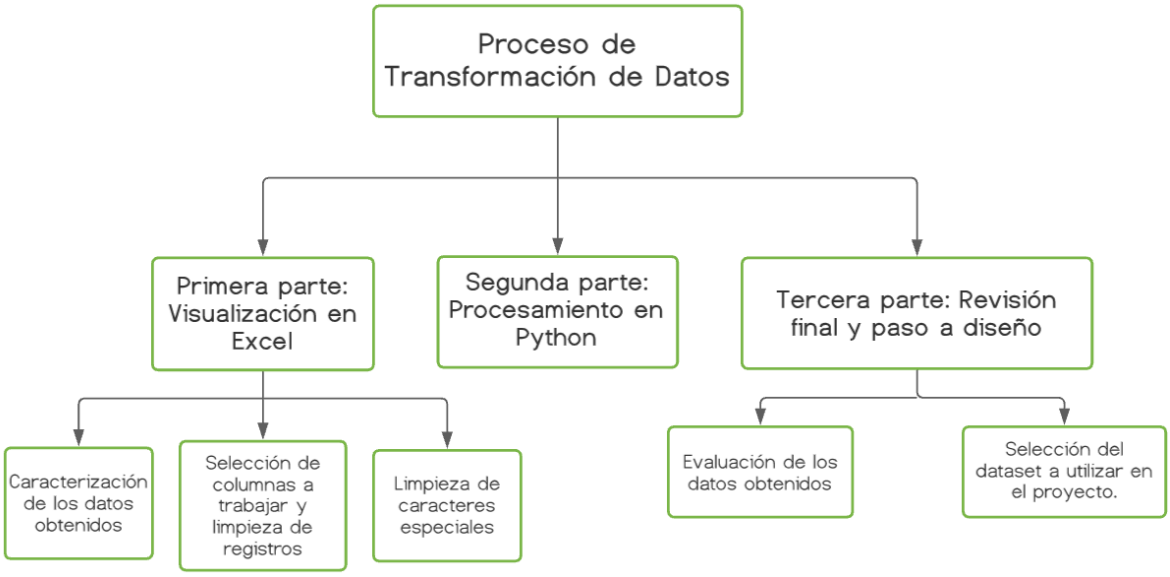

Figura 15, Proceso Transformación de Datos, Fuente: Propia

8.3.1 Primera parte: Visualización en Excel.

- **Caracterización de los datos obtenidos**

Para este primer análisis se encontraron que las columnas que venían en este Dataset eran: “Nombre”, "areaConst” “preciom2” “Admon” “estrato” “Antigüedad” “noPiso” “sector” “descripción” “caracteristicasInt” “caracteristicasExt” “caracteristicasSect” “Precio” “Usado” “Tamano” “Habitaciones” “Banos” “Parqueadero” “error”.

Lo primero que se logró identificar fue que desde las columnas “areaConst” hasta la columna “Sector”, la información que venía en los campos venía con su propio nombre de variable y valor correspondiente. Sin embargo, muchas veces ese nombre de variable no corresponde con la columna en la cual se encuentra almacenado además que se encontraron nuevas variables que deberían de tener su propia columna como lo son: “Estado” y “Area Privada”. Para esta sección del Dataset se estableció que era necesario agregar las columnas para estas 2 variables identificadas, es imperioso que el valor de cada columna corresponde únicamente al valor de la variable numérica, es decir sin que esté precedido por el nombre de dicha variable. Para este último proceso descrito, se reconoció un patrón que facilitó la separación de esta información. Cada variable antes de mencionar el valor correspondiente está separada por el carácter dos puntos (:) lo cual será utilizado en el código de Python para lograr el objetivo propuesto.

Por otra parte, las columnas de “descripción” “caracteristicasInt” “caracteristicasExt” “caracteristicasSect”, no tienen un valor numérico en sí, puesto que estas

almacenan descripciones en cadenas de texto lo cual dificulta la segmentación de esta en información útil para ser implementada en los algoritmos. Para extraer dicha información, sería necesario realizar un procesamiento del lenguaje natural y este proceso se encuentra fuera del alcance de este proyecto. Las columnas de “tamaño” “Habitaciones” “Banos” “Parqueadero”, son similares a las primeras columnas identificadas, pero con la ventaja que el nombre de la variable que almacena si corresponde a la columna en donde está ubicada, lo cual facilita el tratamiento de los datos en esta sección ya que solo se necesitó eliminar toda la palabra que antecede el valor de la variable para que estas columnas quedarán solamente con el valor numérico requerido

La columna de “Usado” se utilizó para identificar los registros de viviendas que eran catalogadas como usadas en *Finca Raíz*, ya que en la página se ofertan tanto viviendas nuevas como usadas. Si la columna tenía como valor la palabra “Usado” significaba que efectivamente la información es de una vivienda que se cataloga como usada, si se encontraba vacío, es decir, sin nada dentro de esa columna lo más probable es que se tratase de una vivienda nueva, que son promocionadas dentro de la página.

Por último, para las columnas de “Nombre” y “precio”, se identificó que no había gran tratamiento que hacer, ya que los datos estaban ubicados correctamente. Para la columna “Nombre” la mayoría de estas empezaban por “Bogotá” y el barrio o sector en donde se ubica la vivienda, lo cual tiene coherencia ya que el proyecto está enfocado únicamente en la ciudad de Bogotá, por otro lado, la columna “precio” no tenía nombres de variables que le antecedieron en la mayoría de los casos, ya que existían situaciones en donde almacenaban registros como “Desde 20000000” para la columna de “precio” y la columna de “nombre” estaba vacía. Lo cual tiene amplia relación con que la columna “usado” también estuviera vacía.

- **Selección de columnas y limpieza de registros.**

Una vez realizada la caracterización de los datos se escogieron las columnas con las que se iban a seleccionar para procesar dentro del código de Python. Estas columnas fueron: ‘Nombre’, ‘Área construida', 'Área privada', 'Precio metro cuadrado', 'Precio administración', 'Estrato', 'Antigüedad', 'Sector', 'Estado', 'Piso No.', 'Tamaño', 'Habitaciones', 'Baños', 'Parqueaderos', 'Precio'. Las columnas de “descripción” “caracteristicasInt” “caracteristicasExt” “caracteristicasSect” se descartaron del Dataset, ya que involucra realizar procesamiento natural del lenguaje para extraer su información, así que se optó para dejarlos como opción en trabajos futuros. Por otro lado, la columna de “error” también se descartó del Dataset final, ya que esta solo aportaba información que era relevante durante la ejecución del robot, no aportaba nada al estudio realizado, por tal motivo se decide eliminar.

Para la limpieza de registros del Dataset lo primero que se realizó fue filtrar en la columna “usado” por vacío, es decir aquellas filas en donde dicha columna no tuviera información almacenada. Como se mencionó anteriormente estas filas coincidían en que las columnas de “nombre” estuvieran también vacías y que en la columna “precio” existieran las palabras “desde”, cuando en esta columna se espera solo el valor numérico. Estos registros se deciden eliminar del Dataset ya que almacenan información de viviendas nuevas y estas no están cubiertas dentro del alcance de este proyecto, ya que involucran otras técnicas de avalúos distintas a las que se utilizan en las viviendas usadas. Luego se decide filtrar en la columna “nombre” cualquier registro que el inicio fuera diferente a Bogotá y alguna otra palabra (que podía ser el sector o el barrio) que le precedieron. Dentro de estos registros filtrados se encontraron casos atípicos como números, campos vacíos, campos que repetían el valor de las etiquetas, los cuales al no aportar valor al estudio se eliminaron. Pasando de tener 3175 datos a 2350 datos, la mayoría siendo eliminados de aquellas viviendas que eran catalogadas como nuevas en *Finca Raíz*.

- **Limpieza de caracteres especiales.**

Para poder pasar el Excel a Python es necesario eliminar ciertos caracteres especiales que si bien sirven para referenciar las unidades de medidas ya sea metros (m^2) o que es un precio ($), dentro del método a realizar esta información resulta estorbosa, ya que solo se necesita el valor numérico. Utilizando la función buscar y reemplazar de Excel, se buscan los siguientes caracteres “m^2”,”/”,”$”,” ª” ,”$\bar{a}$”y se reemplazan con vacío. Para la columna “precio” se transforma la columna a general dentro de las categorías de Excel, ya que este mismo lo identificaba como moneda y le agregaba el signo “$”. Por último, se identificó que en algunas las columnas los separadores de miles y los separadores de decimales no eran los mismos. En algunas columnas como “Preciom2” utilizaban el punto (.) como separador de miles y en “Admon” utilizaban la coma (,) como el separador de miles, lo cual dificulto el manejo de estos caracteres y que se soluciona en el desarrollo del código.

8.3.2 Segunda parte: Procesamiento en Python.

Una vez realizada esta limpieza se procede a utilizar Python para la organización de los datos. El código utilizado en Python se presenta a detalle en el Anexo 1, siendo propiedad intelectual de los autores. Ahora bien, durante la ejecución de dicho código se encontraron varios datos atípicos, se encontró una fila que tenía dentro de sus campos el valor de “Estrato: campestre”, como solo se encontró una se eliminó dicha fila para que no afectará el desarrollo del estudio.

Del mismo modo se encontró una fila que en la columna “parqueaderos” tenía el valor de “Más de 10 parqueaderos”, así que también se decidió eliminar dicha fila.

Se encontró también que había filas en el campo donde estaba “Piso No” decían “otros” así que se cambió esa palabra por “-1” para que así se pudieran procesar y luego tener en cuenta ese valor como algo no relevante de esas filas. Por último, se encontró una fila cuyo valor en precio era “A consultar”, por tal motivo esta fila es eliminada del Dataset para el estudio. Una vez realizados estos cambios que surgieron al momento de la ejecución, el programa finalizó sin problemas, generando un archivo tipo hoja cálculo de Excel en donde se realizó un nuevo análisis de los datos obtenidos. Las nuevas columnas obtenidas en este Dataset son:
‘Nombre','Area construida', 'Area privada','Precio metro cuadrado','Precio administr acion','Estrato','Antiguedad','Sector','Estado','Piso No.','Tamaño','Habitaciones','Ba ños','Parqueaderos','Precio'.

8.3.3 Tercera parte: Revisión final y paso a diseño.

- **Evaluación de los datos obtenidos.**

Para el presente estudio las columnas de nombre y sector no pueden ser utilizadas ya que se requieren datos numéricos tipo entero para trabajar, sin embargo, para trabajos futuros se considera realizar una categorización por UPZ según el sector, pero, por ahora se decidió eliminar estas columnas del Dataset.

Por otra parte, se encuentra que con el nuevo Dataset organizado todas las filas tienen “Área construida”, “Precio metro cuadrado”, “Precio”, “Tamaño”,” Habitaciones”,” Baños”,” parqueaderos”, pero 1569 tienen “área privada”, 1599 tienen “Precio administración”, 2295 tienen “estrato”, 2012 tiene “Antigüedad”, 1609 tienen “estado” y por último solo 820 tienen “Piso No”. Como muy pocas filas tienen el campo “Piso no”, se decide eliminar esa columna del estudio, además que esto solo debería de aplicarse para apartamentos y en muchos casos los vendedores se confunden y colocan la cantidad de pisos que tienen una casa en venta. Sin embargo, se deciden eliminar las filas que no poseen antigüedad y estrato, ya que comparado con las de “Piso no” la cantidad es menor y además se ha encontrado que el estrato puede influir fuertemente en el precio, quedando así con 2102 datos.

- **Selección del Dataset a utilizar en el proyecto.**

Para aprovechar al máximo la información obtenida se decide crear 3 Dataset para probar con los algoritmos a diseñar, el primero es el que tiene la mayor cantidad de filas para el estudio (2102), pero eliminando las columnas de “área privada”, “precio administración” y “Estado”, el segundo es dejando la columna del “estado”, eliminando “área privada” y “Precio administración”, para evaluar si el estado de la vivienda tiene correlación directa por el precio, quedando un total de 1549 datos

para realizar el estudio. Por último, se deja un Dataset que tenga valores en todas las columnas quedando con un total de 881 datos para implementar en los algoritmos. Los Dataset que se generaron se pueden consultar en el anexo 2.

9 DISEÑO DEL MÉTODO AUTOMÁTICO

En este capítulo para el diseño del método automático se tomaron en cuenta que algoritmos de Machine Learning fueron los utilizados para dar respuesta al objetivo propuesto y qué métricas o medidas de desempeño se utilizaron para la evaluación del mismo.

9.1 Algoritmos a utilizar

Es importante aclarar que existen una gran variedad de técnicas de aprendizaje automático, para clasificar, agrupar o predecir datos. Es por esto que gracias al estudio realizado con el estado del arte y los diversos casos de estudio que aquí se presentan, se escogieron los modelos de Machine Learning por su simplicidad de implementación, interpretación y que son los más frecuentes en estos casos, para de esta manera ponerlos a prueba con el conjunto de datos obtenido en el capítulo anterior. Los modelos a utilizar fueron:

9.1.1 Regresión Lineal

Es el primer modelo implementado debido a su simplicidad de entendimiento, este algoritmo busca encontrar de manera automática la recta que mejor se ajuste a la naturaleza de los datos, para esto se interpreta la relación que se puede presentar entre varias variables, dicha relación puede demostrar una tendencia fuerte o débil la cual puede ser cuantificada por medio del coeficiente de correlación y de determinación. Para poder identificar esta ecuación o recta, se pueden estimar una o varias rectas, pero para esto se utiliza el método de mínimos cuadrados ordinarios.

- **Mínimos Cuadrados Ordinarios (MCO)**

El término de mínimos cuadrados ordinarios MCO está ligado a la regresión y a la correlación, pues ambas determinan la existencia de relación entre dos o más variables como se presenta en el método automático donde como variable dependiente (Y) se tiene el precio de las viviendas siendo los targets u objetivos en el modelo de regresión, y como variable independiente (X) se encuentran almacenados todos los parámetros y características de vivienda que describen dicho precio, en otras palabras el método de mínimos cuadrados ordinarios permite encontrar los mejores estimadores lineales.

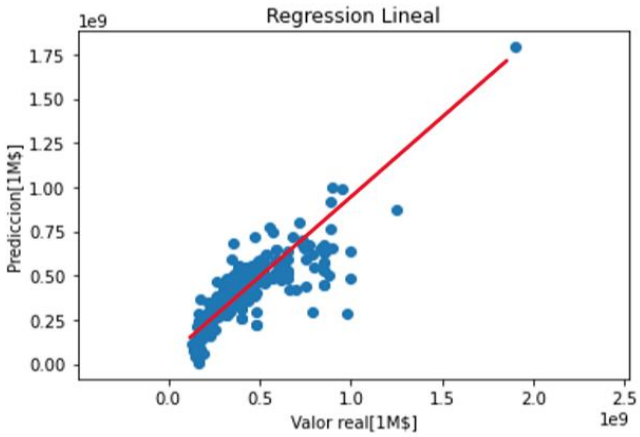

Figura 16, Modelo de Regresión Lineal, Fuente: Propia

Así como se observa en la figura anterior (Figura 16), se observa como un conjunto de datos predichos con un modelo de regresión lineal se ajusta linealmente al conjunto de datos de viviendas en Bogotá. El resultado y calidad de estas predicciones se puede acomodar de una mejor manera mediante el ajuste correcto de dichos datos, buscando una correlación entre las características de la vivienda con el precio.

9.1.2 Random Forest

El modelo de Random Forest, según José Martínez Heras[69], es una técnica de aprendizaje automático popular y cuentan con una capacidad de generalización muy alta por su aplicabilidad a muchos problemas.

Este modelo presenta bastantes ventajas respecto a otros modelos de aprendizaje supervisado al aplicar un método de aprendizaje por conjuntos para la regresión, esta técnica combina las predicciones de múltiples algoritmos de Machine Learning para realizar una predicción más precisa con un solo modelo, es por esto que se decidió implementarlo al método automático.

[69] MARTINES H José, Random Forest (Bosque Aleatorio): combinando árboles, {En Línea} {11 de mayo de 2021} {Disponible en: https://www.iartificial.net/random-forest-bosque-aleatorio/#Diferencia_intuitiva_entre_un_arbol_de_decision_y_un_random_forest}. P.74

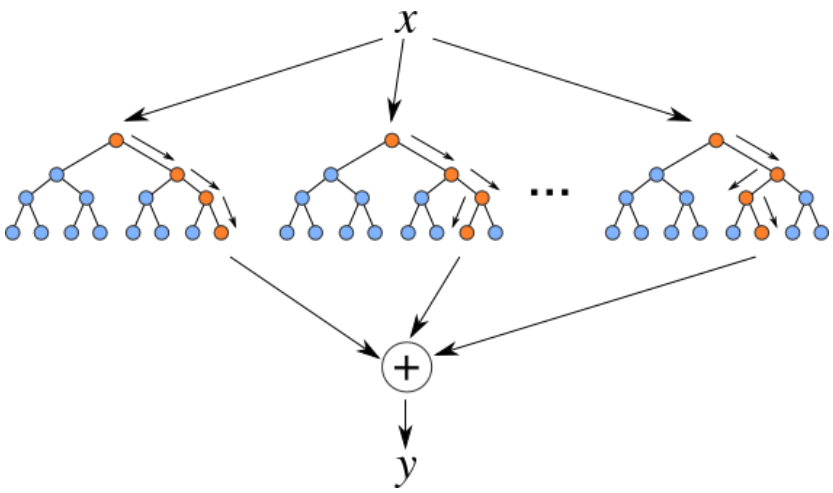

Figura 17, Estructura del Random Forest, Fuente: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Como se observa en la figura 17, el Random forest está compuesto por varios árboles de decisión, a cada árbol le corresponde una parte del conjunto de datos de entrenamiento de manera aleatoria y el cual realiza el proceso de regresión. El árbol parte de unas **x** características seleccionadas de unas totales y de esto se crean n árboles de decisión como se observa en la figura. En cada uno de estos árboles se realiza el proceso de regresión y el resultado de cada árbol se almacena obteniendo n salidas y de estos resultados se toman los más votados por el bosque (conjunto de árboles de decisión) obteniendo el resultado del Random forest en **y**.

9.1.3 Árboles de Decisión (Decision Tree)

Los árboles de decisión son uno de los modelos de predicción más utilizados en el Machine Learning, debido a su alta confiabilidad en resultados, su facilidad de implementación e interpretación, por tal razón es uno de los modelos elegidos para aplicar en el método automático propuesto. Los árboles de decisión están compuestos por nodos y su lectura se realiza de arriba hacia abajo como se muestra en la figura 18, el árbol de decisión parte de un nodo raíz y de este se producen las primeras divisiones en función de la variable más importante, de esta división se encuentran nuevos nodos que dividen nuevamente el conjunto de datos en más nodos.

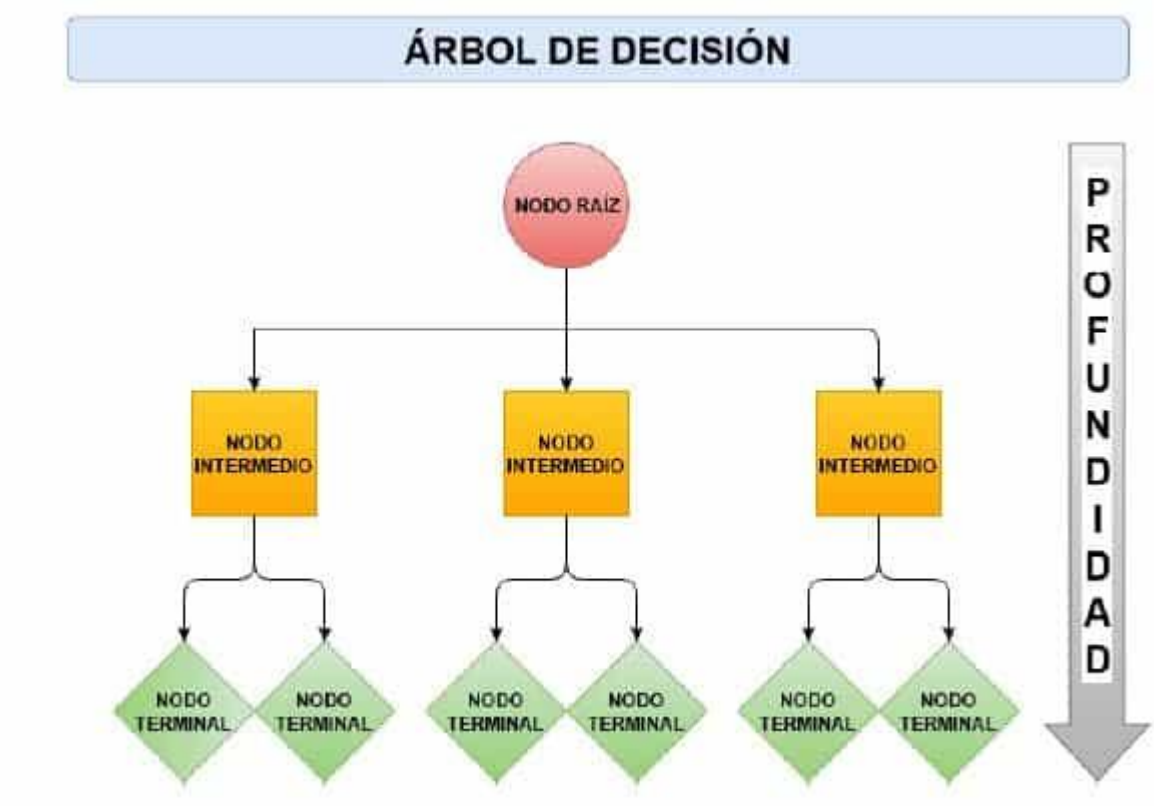

Figura 18, Estructura Árbol de Decisión, Fuente: https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/

9.1.4 Red Neuronal Profunda.

Como cuarto modelo se implementa una red neuronal profunda debido a su popularidad actual en la predicción y el desempeño en los modelos de regresión lineal. Presentando de esta manera un modelo de Deep Learning y sus técnicas de aprendizaje con ayuda de las redes neuronales, para este modelo se utiliza una red neuronal profunda, esta red neuronal, por medio de experimentos y el ajuste manual de sus híper parámetros, se busca que la red neuronal se adapte al problema que se está resolviendo y lograr un buen desempeño, algunos de los híper parámetros son los optimizadores, la función de activación, las épocas, la tasa de aprendizaje y la función de pérdida.

De esta manera, se tendrá como entradas a la red, todas aquellas variables que describen el precio de una vivienda como se evidencia en la figura 19, presentadas en el modelo de regresión como las variables independientes que se correlacionan con la variable dependiente que corresponde a la variable de la vivienda.

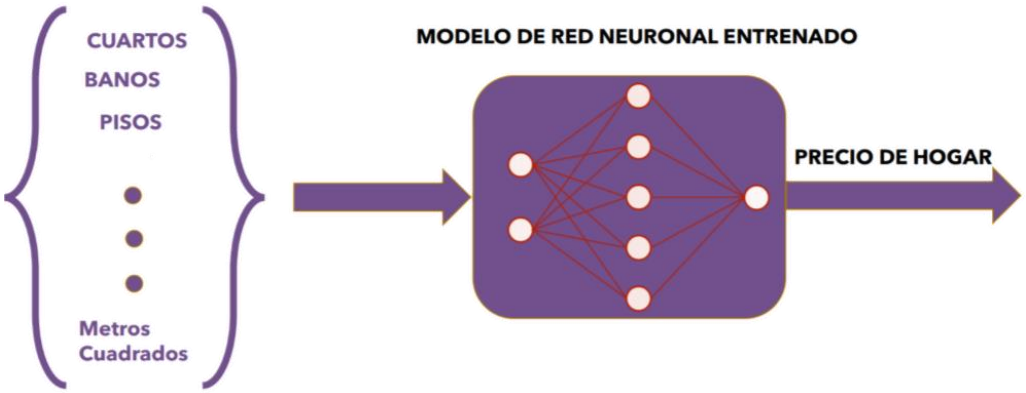

Figura 19, Estructura Red Neuronal, Fuente: Propia.

La validación de esta red se evalúa según el progreso del modelo obtenido durante su entrenamiento según sus pérdidas y los híper parámetros otorgados a la red neuronal, estás pérdidas miden la salida obtenida por el sistema, frente a la salida deseada.

Es por esto que para este método automático se evalúa su calidad de implementación en función de las predicciones realizadas por los modelos vs los datos reales que se presentan en el Dataset con las siguientes medidas de desempeño.

9.2 Medidas de Desempeño:

9.2.1 Coeficiente de Determinación (R^2).

El coeficiente de determinación tiene el objetivo de medir que tan bueno es un modelo y se representa de la siguiente manera. Este coeficiente es un indicador que permite determinar que tanto se ajustan los modelos al precio de las viviendas.

A. $$\frac{\sum_{t=1}^{T}(\hat{Y}_t - \underline{Y_t})^2}{\sum_{t=1}^{T}(Y_t - \underline{Y_t})^2} \quad (1)$$

Donde:

Y_i =Precio de la vivienda conocido.

$\hat{Y}_i$ =Precio de la vivienda predicho por el modelo.

n =número total de predicciones.

Como se observa en la ecuación 1, el coeficiente de determinación se calcula con la sumatoria de los precios de vivienda conocidos, menos el precio que ha sido generado por cada modelo. Esto se divide de igual manera, con la sumatoria de los precios que se conocen, menos el precio que ha sido generado por los modelos.

9.2.2 Error cuadrático medio RMSE.

El error cuadrático medio es una medida la cual indica el error entre un conjunto de datos, se compara un valor predicho con un valor real o ya conocido, el RMSE cuantifica qué tan diferente es un conjunto de resultados, cuanto más pequeño es el valor de RMSE, más cercanos son los resultados de la predicción a los reales u observados, esta medida es utilizada para la evaluación del método automático por su comprensión y su utilidad que demuestra ante modelos de regresión lineal.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}} \quad (2)$$

Donde:

P_i =Precio de vivienda predicho por el modelo

O_i =Precio de vivienda

n =número total de predicciones

De esta manera, como se observa en la ecuación 2, el RMSE se calcula con el valor de los precios que cada modelo ha predicho de la vivienda, menos el valor de la vivienda que ya se conoce en el conjunto pruebas. Con esto se puede calcular que tanto se ha acercado o alejado al precio real, es decir, el error que representan estos dos valores.

9.2.3 Coeficiente de Variación con el RMSE (CV)

Según Francisco Javier Marco Sanjuán, el coeficiente de variación o también conocido como coeficiente de variación de Pearson es una medida estadística que da a entender la dispersión relativa de un conjunto de datos[70]. Este coeficiente forma parte de las medidas de dispersión en la regresión, este coeficiente se obtiene a partir de dividir la desviación típica entre el valor absoluto de la media del conjunto y para su mejor comprensión se expresa en porcentaje, en otras palabras, este coeficiente es útil cuando se requiere comparar la dispersión de dos o más conjuntos de datos.

$$\mathrm{Cv} = \frac{RMSE}{|\tilde{\mathrm{x}}|} \qquad (3)$$

Donde:

$RMSE$: Raíz del error cuadrático medio.

$|\tilde{\mathrm{x}}|$: Es la media de la variable Y, es decir, el conjunto de precios de todas las viviendas.

Como se observa en la ecuación 3, este coeficiente se calcula a partir del valor del RMSE obtenido de cada modelo sobre el valor de la media de la variable Y, el cual corresponde al conjunto de precios de las viviendas que están en el conjunto de datos. Una vez como resultados, se obtiene el valor del coeficiente de variación para cada modelo y en cada una de las ejecuciones por estratos. Posterior a esto, una vez con los resultados ya generados, se define cuál de estos presenta una menor y mayor dispersión en los resultados.

[70] MARCO S Francisco Javier. Coeficiente de variación {En Línea} {11 de mayo de 2021} {Disponible en: https://economipedia.com/definiciones/coeficiente-de-variacion.html} p.79

9.3 Análisis de los Dataset

Para escoger el Dataset a utilizar de los 3 obtenidos en el capítulo anterior se realizó un análisis gráfico en el cual se estudiaron las correlaciones que presentaban cada uno de los Dataset, de estos se escogió el mejor mapa de calor como se observa en la siguiente figura.

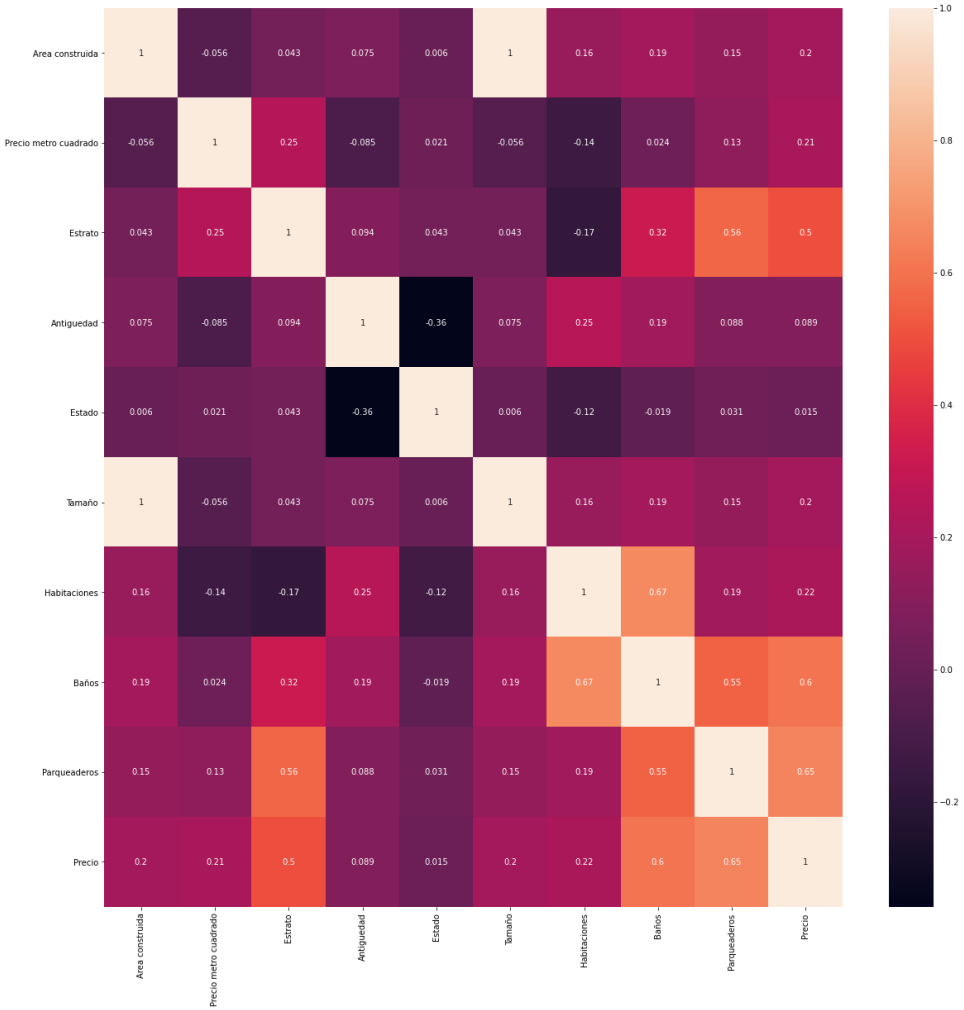

Figura 20, Dataset #3 con estado, Fuente: Propia

Como se puede evidenciar en el mapa de calor (Figura 20), se observa que la mayoría de sus variables presentan una correlación a excepción de la variable estado con la antigüedad. Acto seguido, para el gran conjunto de datos que se obtuvo, se decidió segmentar el Dataset por estratos y de esta forma trabajar con 6 conjuntos de datos diferenciados por el estrato de la vivienda del 1 al 6, de esta manera, se obtuvo una correlación más proporcional del precio con las demás características del Dataset, así como se evidencia en el siguiente mapa de calor con el estrato 6.

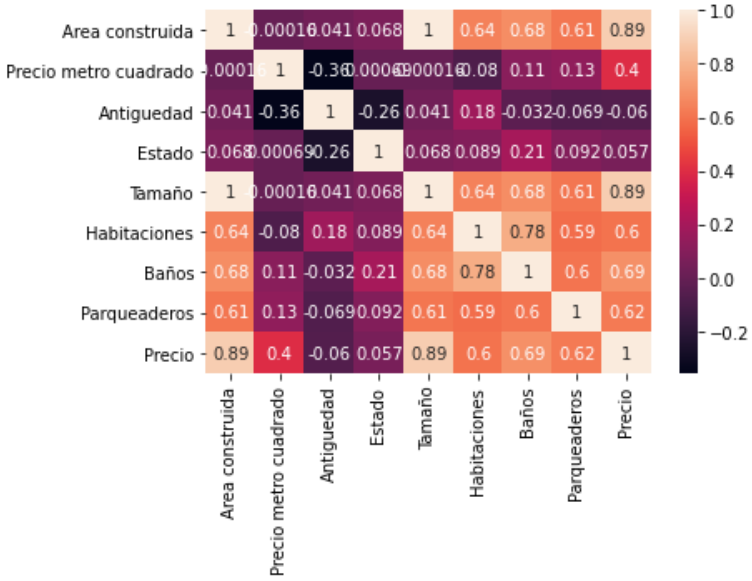

Figura 21, Dataset #3 segmentado en estratos, mapa de calor Estrato 6 ***Fuente: Propia***

Sin embargo, al realizar esta partición por estratos, se observó que el estrato 1 solo contenía un total de 4 datos como se evidencia en la figura 22, dado esto se decidió eliminar los datos del estrato 1 y trabajar con los estratos del 2 al 6. Estos fueron los datos elegidos para correr en cada uno de los modelos ya seleccionados.

	A	B	C	D	E	F	G	H	I	J
1	Area const	Precio met	Estrato	Antiguedad	Estado	Tamaño	Habitacion	Baños	Parqueade	Precio
86	72	1333333	1	4	2	72	2	1	1	96000000
221	164	7926829	1	4	2	164	3	3	2	1300000000
1462	36	2500000	1	2	2	36	3	2	0	90000000
1491	36	2500000	1	2	2	36	3	2	0	90000000

Figura 22, Dataset #3 con estado filtrado por estrato 1, ***Fuente: Propia***

9.3.1 Partición del Dataset

En cuanto a la partición de Dataset, teniendo presente que se trabajaron un total de 1537 datos, para el entrenamiento y prueba del método automático, se realizó una partición del 70% de los datos para el entrenamiento de los modelos y el 30% para las pruebas del mismo, esta partición fue la misma para cada uno de los Dataset obtenidos por estrato y de esta manera observar el comportamiento de los resultados de cada uno de los modelos y su desempeño con la predicción de precios con el conjunto de prueba.

9.4 Diseño final.

Para cerrar este capítulo de diseño en la siguiente tabla (tabla 1) se evidencia el diseño final con el cual se procedió al desarrollo del método automático.

Modelos Seleccionados	**Medidas de Desempeño utilizadas**	**Dataset Seleccionado**	**Partición del Dataset**	**Variables Utilizadas**
• Regresión Lineal • Árboles de decisión • Random Forest • Red Neuronal Profunda	• Coeficiente de Determinación(R^2) • Coeficiente de Variación(CV) • RMSE	Dataset sin la columna de área privada, precio administración, y con un total de 1550 datos. Dataset particionado en 5 por los estratos.	70% Entrenamiento 30% Prueba	Area Construida, Precio metro cuadrado, Antigüedad, Estado, Tamaño, Habitaciones, Baños, Parqueaderos, Precio.

Tabla 1, Diseño Final, *Fuente: Propia*

10 DESARROLLO DEL MÉTODO AUTOMÁTICO

Con los diseños realizados en la fase anterior se procedió al desarrollo del método automático y de los distintos modelos de Machine Learning seleccionados, estos modelos se desarrollaron en el entorno Spyder por la experiencia de programación y por la interfaz que ofrece para la visualización de gráficas y variables. Una vez en Spyder se procedió a buscar las librerías en Python las cuales proporcionan como base la construcción del método automático; las librerías empleadas se pueden evidenciar en la tabla 2.

Liberia	Utilidad
matplotlib.pyplot	Generación de gráficos a partir de listas y Arrays
Numpy	Creación de vectores y matrices
Pandas	Manipulación y análisis de datos
Seaborn	Visualización de datos y gráficas.
Scikit-learn (sklearn)	Herramienta básica para el Data Science en Python.
Math	Cálculos numéricos
TensorFlow	Convierte el desarrollo de la red neuronal de manera más sencilla
Statistics	Cálculos para análisis estadísticos.

Tabla 2, Librerías utilizadas *Fuente: Propia*

La invocación de estas librerías se puede evidenciar en la figura 23, en la cual se observa cómo se importan las librerías ya mencionadas, estas librerías son las utilizadas para la construcción de los modelos, la partición del Dataset y también se incluyen las librerías empleadas para el cálculo de las medidas de desempeño.

```
import tensorflow as tf
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, r2_score
from math import sqrt
import statistics as stats
```

Figura 23, Importación de Librerías en Spyder Fuente: Propia

Antes de proceder con la codificación de cada uno de los modelos se realiza la partición de los datos con el 70% para entrenamiento y el 30% restante de los datos para la prueba de los modelos como se mencionó anteriormente en el diseño del método automático. Esta partición se puede evidenciar en figura 24, donde se observan los porcentajes dados y la librería necesaria para su codificación.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X,y, train_size = 0.7)
X_train.shape,X_test.shape,y_train.shape,y_test.shape
```

Figura 24, Partición del Dataset entrenamiento y prueba Fuente: Propia

Posteriormente se realiza la codificación de cada uno de los modelos de Machine Learning en el entorno de programación ya mencionado, comenzando por los 3 modelos de aprendizaje supervisado, el modelo de regresión lineal, los árboles de decisión, el modelo de random forest y para finalizar el modelo de red neuronal profunda gracias al Deep Learning. A continuación, en la figura 20, para este último modelo se muestra como fue la arquitectura de la red neuronal utilizada con sus híper parámetros mencionados previamente en el capítulo del diseño, también se puede evidenciar que la arquitectura de la red que está compuesta por una capa de entrada con 100 neuronas, con la función de activación Relu al igual que las capas ocultas, y un input shape de 8 el cual corresponde al total de características que entran en la red para ser procesadas. En cuanto al entrenamiento de la red neuronal a diferencia de los demás modelos este cuenta con híper parámetros que modifican el entrenamiento y su desempeño, por ejemplo, la red fue construida de manera que entrenará durante 80 épocas y un batch size de 50 el cual es el número de muestras que entran a la red para que entrene durante cada época.

```
model = tf.keras.models.Sequential()
model.add(tf.keras.layers.Dense(units=100, activation='relu', input_shape=(8, )))
model.add(tf.keras.layers.Dense(units=100, activation='relu'))
model.add(tf.keras.layers.Dense(units=100, activation='relu'))
model.add(tf.keras.layers.Dense(units=1, activation='linear'))

model.summary()

model.compile(optimizer='Adam', loss='mean_squared_error')

epochs_hist = model.fit(X_train, y_train, epochs = 80, batch_size = 50)
```

Figura 25, Arquitectura de la red neuronal profunda Fuente: Propia

Es importante mencionar que antes de comenzar con el entrenamiento y prueba de cada uno de los modelos, se realiza una última limpieza de datos con cada uno de los Dataset por estratos para verificar y asegurar que estos datos no tengan espacios vacíos que puedan dar complejidad a la ejecución o entrenamiento de los modelos, esta limpieza se puede evidenciar en la siguiente figura 26. En ella se observa cómo se realiza una búsqueda de datos vacíos, y en caso de haberlos, se realiza la eliminación de los mismos.

```
#Busqueda de datos vacios
df.isnull().values.any()
len(df)
# Eliminacion de datos vacios
df=df.dropna()
len(df)
```

Figura 26, Búsqueda y Eliminación de Datos Vacíos Fuente: Propia

Una vez realizada esta última limpieza, se lleva a cabo el entrenamiento de cada uno de los modelos para posteriormente realizar la predicción de los precios de las viviendas que se encuentran en el conjunto de pruebas. Después de realizadas dichas predicciones, se procede a la evaluación de los modelos según los resultados de los precios obtenidos versus los precios ya conocidos del conjunto de pruebas, con ayuda de estos resultados y las medidas de desempeño mencionadas en el capítulo anterior se procede a la evaluación del desempeño de cada uno de los métodos. Para una mayor profundización acerca del desarrollo del método automático y de los distintos modelos, consultar el anexo 3.

11 EVALUACIÓN DEL MÉTODO AUTOMÁTICO

Este capítulo muestra el resultado de las medidas de desempeño generadas por la ejecución de los 4 modelos con los 5 conjuntos de datos. Es importante recordar que el entrenamiento para todos los modelos fue el mismo, es decir, se trabajaron los mismos Dataset y el mismo porcentaje de datos para entrenamiento y prueba. En la figura 27 se pueden observar el conjunto de medidas de desempeño que se pueden emplear para casos de regresión lineal. Es importante recordar que, de este conjunto de medidas, las empleadas para la evaluación de los modelos son el coeficiente de determinación (R^2), la raíz del error cuadrático medio (RMSE) y el coeficiente de variación calculado con el RMSE.

Regression	
'explained_variance'	`metrics.explained_variance_score`
'max_error'	`metrics.max_error`
'neg_mean_absolute_error'	`metrics.mean_absolute_error`
'neg_mean_squared_error'	`metrics.mean_squared_error`
'neg_root_mean_squared_error'	`metrics.mean_squared_error`
'neg_mean_squared_log_error'	`metrics.mean_squared_log_error`
'neg_median_absolute_error'	`metrics.median_absolute_error`
'r2'	`metrics.r2_score`
'neg_mean_poisson_deviance'	`metrics.mean_poisson_deviance`
'neg_mean_gamma_deviance'	`metrics.mean_gamma_deviance`
'neg_mean_absolute_percentage_error'	`metrics.mean_absolute_percentage_error`

Figura 27, Medidas de desempeño para Regresión Fuente: Propia

Los resultados obtenidos se encuentran en las siguientes tablas:

11.1 Resultado Dataset Estrato 2.

A continuación, se presenta el resultado de las medidas de desempeño generado por cada uno de los modelos con el conjunto de datos del estrato 2.

	Estrato 2			
	Regresion Lineal	**Arboles de Decision**	**Random Forest**	**Red Neuronal Profunda**
R2	0.67	0.94	0.92	0.55
RMSE	131569156	53683631	63417559	204991094
CV	0.32	0.13	0.15	0.96

Tabla 3, Resultados obtenidos con el Dataset de Estrato 2 *Fuente: Propia*

Como se observa en la tabla 3, los modelos que presentan un mejor coeficiente de determinación es el árbol de decisión y seguido de este el random forest, es por esto que dichos modelos presentan un mejor ajuste a los precios de vivienda en el conjunto de datos. Por otro lado, los resultados obtenidos del RMSE en cada modelo del estrato 2, se puede observar que los modelos que presentan un menor error son los árboles de decisión, seguido del random forest, esto da a entender que el valor de la predicción del precio de vivienda, es más cercano respecto al valor real y conocido de la vivienda. Por último, se observa que el coeficiente de variación de los árboles de decisión es el menor, seguido del coeficiente del random forest, esto demuestra que dichos modelos presentan una menor dispersión en los resultados.

11.2 Resultado Dataset Estrato 3.

A continuación, se presenta el resultado de las medidas de desempeño generadas por cada uno de los modelos con el conjunto de datos del estrato 3.

	Estrato 3			
	Regresion Lineal	**Arboles de Decision**	**Random Forest**	**Red Neuronal Profunda**
R2	0.72	0.97	0.93	0.58
RMSE	117067859	41563151	57480731	264607274
CV	0.34	0.12	0.16	0.78

Tabla 4, Resultados obtenidos con el Dataset de Estrato 3 *Fuente: Propia*

Para el análisis de la tabla 4, los modelos que presentan un mejor coeficiente de determinación son los árboles de decisión y seguido de este el random forest. Continuando con el resultado obtenido con el RMSE, se puede deducir que los modelos que presentan un menor error son los árboles de decisión, seguido del random forest, esto da a entender que el valor de la predicción del precio de vivienda, es más cercano respecto al valor real y conocido de la vivienda para estos. Por último, como se observó en los resultados del estrato 2, el coeficiente de variación de los árboles de decisión sigue siendo el menor, seguido igualmente del coeficiente del random forest, esto indica que dichos modelos presentan una menor dispersión en los resultados respecto a los demás modelos.

11.3 Resultado Dataset Estrato 4.

A continuación, se presenta el resultado de las medidas de desempeño generadas por cada uno de los modelos con el conjunto de datos del estrato 4.

	Estrato 4			
	Regresion Lineal	**Arboles de Decision**	**Random Forest**	**Red Neuronal Profunda**
R2	0.62	0.75	0.90	0.32
RMSE	1257383853	119475369	74187084	249438110
CV	0.67	0.25	0.15	0.53

Tabla 5, Resultados obtenidos con el Dataset de Estrato 4 *Fuente: Propia*

Ahora una vez obtenidos los resultados en la tabla 5, a diferencia de las tablas anteriores, los modelos que presentan un mejor coeficiente de determinación es el random forest, seguido por los árboles de decisión, esto demuestra que a diferencia del estrato 2 y 3, según el coeficiente de determinación, el modelo de random forest obtuvo un mejor ajuste a los precios de vivienda que los árboles de decisión, es decir, el modelo de random forest se ajusta mejor a la naturaleza de los datos con las viviendas de estrato 4. De la misma manera, al observar el valor del RMSE generado por los modelos, se puede deducir que los modelos que presentan un menor error, son el modelo random forest, seguido de los árboles de decisión, esto da a entender que el modelo de random forest presentó una mejor aproximación a los valores reales.

Por último, como se observó en los resultados del estrato 2, el coeficiente de variación de los árboles de decisión continúo siendo el menor, seguido igualmente del coeficiente del random forest, esto da a entender que dichos modelos presentan una menor dispersión en los resultados respecto a los demás modelos. Es importante mencionar que los modelos de regresión lineal y la red neuronal profunda, siguen siendo los de menor desempeño según los resultados obtenidos desde el estrato 2. De esto se puede decir que, el mejor modelo de regresión lineal destaca más que la red neuronal, pues este modelo presenta un bajo ajuste a los precios de vivienda del Dataset.

11.4 Resultado Dataset Estrato 5.

A continuación, se presenta el resultado de las medidas de desempeño generadas por cada uno de los modelos con el conjunto de datos del estrato 5. Al observar que sus resultados se asemejan a los resultados obtenidos con el estrato 6, se realizó un análisis de ambos resultados, destacando como el modelo de regresión lineal obtiene un mejor resultado en sus medidas de desempeño en estos últimos estratos.

	Estrato 5			
	Regresion Lineal	**Arboles de Decision**	**Random Forest**	**Red Neuronal Profunda**
R2	0.90	0.88	0.90	0.002
RMSE	127343848	142181241	129605780	634640990
CV	0.16	0.18	0.16	0.83

Tabla 6, Resultados obtenidos con el Dataset de Estrato 5 Fuente: Propia

11.5 Resultado Dataset Estrato 6.

A continuación, se presenta el resultado de las medidas de desempeño generadas por cada uno de los modelos con el conjunto de datos del estrato 6.

	Estrato 6			
	Regresion Lineal	**Arboles de Decision**	**Random Forest**	**Red Neuronal Profunda**
R2	0.95	0.58	0.92	0.11
RMSE	247346943	712662302	304152345	1620352905
CV	0.18	0.52	0.22	0.19

Tabla 7, Resultados obtenidos con el Dataset de Estrato 6 Fuente: Propia

Después de obtener los resultados en los estratos 5 y 6, así como se puede evidenciar en las tablas 6 y 7 respectivamente, se dedujo que los modelos de regresión lineal lograron una mejoría significativa en estos últimos estratos, obteniendo un coeficiente de determinación igual y mejor que el modelo de random forest en los estratos 5 y 6 respectivamente.

En cuanto a el valor del RMSE, se puede observar que los estratos 5 y 6 tuvieron un pequeño incremento, esto se deduce debido al incremento del costo que tienen las viviendas de estos últimos estratos, de la misma manera, el modelo de regresión lineal, presentó una mejoría en dichos indicadores, pues presentó un menor error en esta medida de desempeño.

Por último, el coeficiente de variación fue menor en el modelo de regresión lineal, estos resultados dan a entender que este modelo presentó una menor dispersión de resultados respecto a los demás modelos. Es importante mencionar que el modelo de la red neuronal fue el modelo que menos se ajustó en cada uno de los

estratos, obteniendo en el R^2 de cada estrato resultados inferiores al 0.60. De los resultados obtenidos, los diagramas de dispersión más relevantes de cada modelo en los diferentes estratos fueron los siguientes.

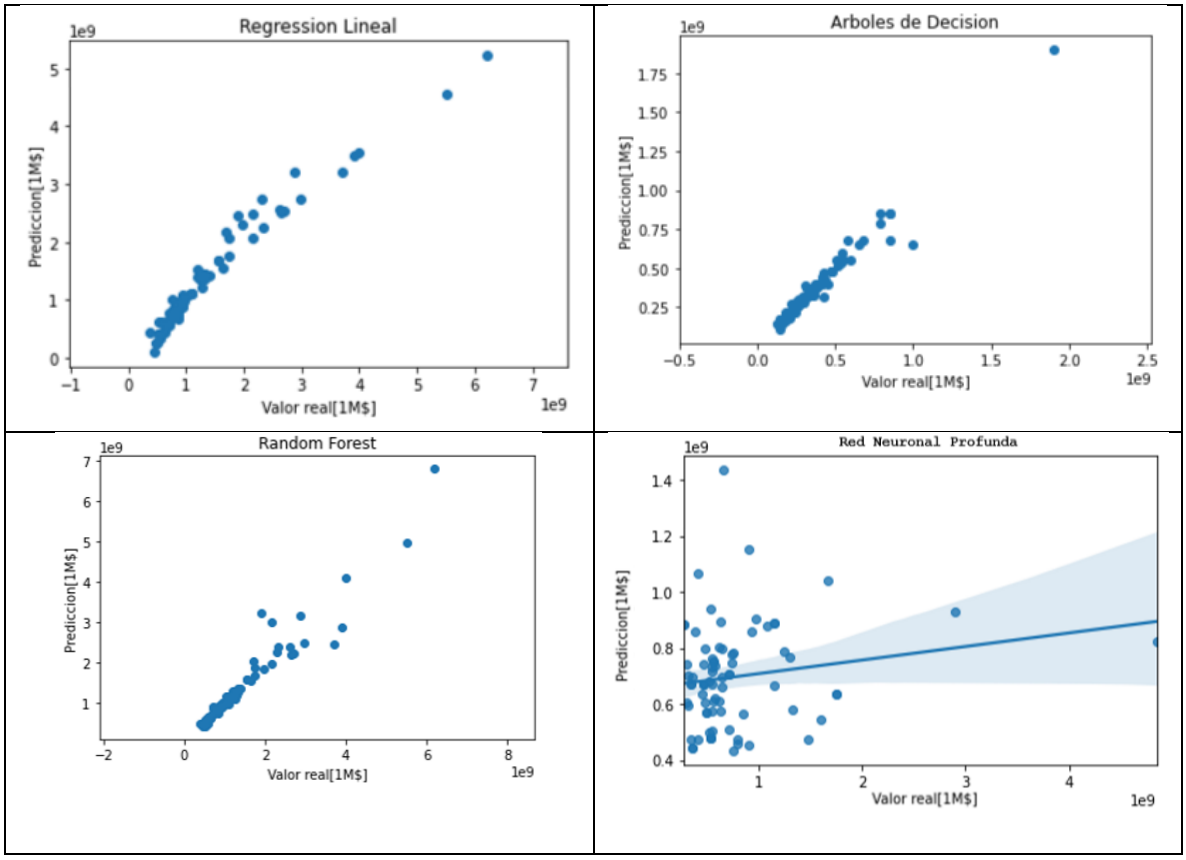

Figura 28, Diagramas de dispersión de los modelos empleados Fuente: Propia

De los diagramas de dispersión anteriores se puede observar el mejor diagrama obtenido para cada modelo. Se observa primeramente el diagrama que presenta un mejor ajuste lineal en el conjunto de datos del estrato 6 con el modelo de regresión lineal, seguido de este, se puede observar el modelo de árboles de decisión el cual presentó un mejor diagrama de dispersión en el conjunto de datos del estrato 3, posterior a este, en el estrato 4 se encontró que el mejor diagrama de dispersión lo presenta el modelo de random forest. Por último, se observó que, de la red neuronal profunda, el diagrama de dispersión con el cual se presentaron mejores predicciones, es en el estrato 5. A simple vista se puede evidenciar que, los modelos que más se ajustan linealmente a la predicción de las viviendas en el conjunto de pruebas son los modelos de regresión lineal, random forest y árboles de decisión, esto se debe a cómo se ajusta cada modelo según la naturaleza del conjunto de datos.

Teniendo presente que cada uno de los modelos trabajo con la misma partición de datos de entrenamiento y prueba. Los modelos de regresión lineal, árboles de decisión y Random forest logran presentar un ajuste lineal en alguno de los estratos según el diagrama de dispersión. Después de obtener dichos resultados, se puede observar que el modelo que presenta el mejor desempeño fue el Random forest y los árboles de decisión según lo observado en los resultados de los Dataset. Ahora bien, en el siguiente capítulo se procedió con el análisis y comparación de resultados respecto a los de Yuri Grajales[71] en su proyecto de predicción de precios de vivienda en Medellín.

[71] GRAJALES, Yuri, Op, Cit, p.91

12 RESULTADOS Y ANÁLISIS DE RESULTADOS

Con respecto a la obtención y raspado de los datos con ayuda de técnicas de web scraping y la indagación de los diferentes portales web que ofrecen dichos servicios, con sus diferentes utilidades y características se eligió el portal web de Dexi.io por su facilidad de entendimiento y baja complejidad para la elaboración del robot, también cabe mencionar que un factor clave para la elección de que portal de raspado web fue la calidad que este tiene ante sus clientes por su soporte técnico y servicio al cliente, con rápido tiempo de respuesta a las peticiones que pudieron surgir durante la elaboración del robot.

A pesar del resultado satisfactorio por el desempeño del robot y su eficacia en la extracción de datos, se tiene presente que habían diferentes alternativas u opciones para la extracción de los datos, por ejemplo se podría haber realizado un web scraping mediante la programación en entornos de Python y las librerías que este ofrece para el raspado web, como se presentó en un trabajo de investigación desarrollado por Yuri Vanesa Grajales Álzate[72] en el cual se desarrolló un web scraping mediante el entorno Python con las librerías "Beautiful Soup (Richardson, 2004) y Selenium (Huggins, 2004)", también hay que tener en cuenta que la extracción se realizó simplemente en la página web de *Finca Raíz* dado que al probar con las demás páginas web como lo pueden ser metro cuadrado, estas no mostraban la información al navegar con el robot de DEXI, se podría haber expandido a más de una página con la elaboración de un web scraping en Python que no limitará la visibilidad de las mismas, sin mencionar que se podría haber ahorrado el costo de la suscripción a Dexi y de esta manera quizás lograr un volumen de datos mayor.

Sin embargo, con el proceso de extracción realizado se pudo obtener la cantidad de 3174 datos del robot que extrajo mayor información de cada vivienda, en contraste con el robot que solo extraía la información general de las viviendas, en donde se obtuvieron un total de 4000 datos. Con estos conjuntos de datos obtenidos, se optó por realizar la limpieza y transformación del Dataset con 3174, en donde a pesar de obtener menos registros de viviendas, se obtuvo una cantidad mayor de características a evaluar. Sin embargo, muchas de esas características implican un procesamiento de lenguaje natural para poder aprovechar el potencial de la información que traía, (las columnas que traían la información de la descripción, características interiores, características exteriores y características del sector), así que se decidió eliminarlas.

Adicionalmente, con la primera revisión del Dataset y la columna de "Usado", se lograron identificar registros que no cumplían con las características para entrar en

[72] GRAJALES Yuri, op. Cit p.92

el análisis, ya que eran consideradas como viviendas nuevas, además que se identificaron filas que no tenían registros consistentes, como números aislados en las filas o el nombre de las columnas en cada fila. Realizando esta limpieza se obtuvo un total de 2350 datos, donde la mayoría de registros eliminados correspondían a viviendas nuevas. El Dataset seleccionado a trabajar fue aquel que contaba con 1549 registros y 10 columnas, comparado con el trabajo de Grajales Yuri[73], en donde se obtuvo un Dataset combinado de las distintas páginas con 3033 registros y 18 columnas pero que el 50% de quedaron nulos o duplicados, se trabajó con una cantidad de datos similar, teniendo en cuenta que ambos estudios son realizados para ciudades de Colombia y obteniendo información de páginas web públicas como *Finca Raíz* en donde cualquier persona sube información, a diferencia de trabajos como el presentado por Yong Piao y compañía[74] en donde se trabaja con un Dataset de 171555 registros, se trabajó con una mayor cantidad de datos, teniendo en cuenta que el contexto social y tecnológico de donde se obtuvo esta información es el de un país como China en donde el acceso y uso a distintas tecnologías de registro de transacciones y obtención de datos es superior al de Colombia.

El diseño del método automático fue realizado con base al estudio realizado por todos los demás artículos científicos e investigaciones que se relacionan con este proyecto. Empezando por la variedad de algoritmos con los que se puede dar solución al problema, se seleccionaron los modelos de regresión lineal, árboles de decisión, random forest, y la red neuronal profunda, todos estos modelos conforman el método automático propuesto a implementar, estos modelos se seleccionaron por su buena interpretación y su facilidad de codificación, además de que algunos de estos modelos como las redes neuronales y el random forest son de los más empleados en los trabajos de investigación similares, así como lo presentó Junchi Bin[75] o Yuri Grajales[76] en sus proyectos de predicción de precios de vivienda. Por otra parte, después de terminar el diseño y desarrollo de los presentes modelos, se estudió que se podrían haber implementado más modelos como las máquinas de soporte vectorial y del gradiente mejorado y de esta manera comparar su desempeño con los demás algoritmos.

En cuanto a las medidas de desempeño presentadas en el diseño del método, se utilizaron las medidas de dispersión más relevantes en los modelos de regresión lineal y tomando como apoyo las métricas que son más utilizadas en los trabajos de investigación similares presentados en el estado del arte, es decir, del gran conjunto de medidas que se pueden utilizar para modelos de regresión, se eligieron las

[73] GRAJALES Yuri, op. Cit p.93
[74] Yong Piao, Ansheng Chen, Zhendong Shang, Op, Cit. p.93
[75] Bin, J.a, Gardiner, B.b, Li, E.c, Liu, Z.a, op. cit p.93
[76] GRAJALES Yuri, op. Cit p.93

medidas de R2 score, el coeficiente de variación y la raíz del error cuadrático medio para la evaluación de los modelos por su facilidad de interpretación y el gran aporte que brindan para los modelos de regresión lineal. Por otro lado, se tiene presente que se podrían haber incluido otras métricas como se muestran en la figura 26, pero el hecho de utilizar solo las mencionadas brindo un mejor análisis de los 4 modelos implementados.

Finalmente, para el análisis y selección de los Dataset se eligió la técnica de los mapas de calor, gracias a su fácil lectura por cómo se representan los datos y cómo diferenciar las diferentes correlaciones entre las variables, de los análisis realizados con los mapas de calor, se optó por dejar todas las variables del Dataset, exceptuando el estrato, esto debido a que, al realizar la última segmentación por estratos, ya se tenía presente que cada conjunto pertenece a un estrato específico. A pesar de haber realizado todo este análisis con los mapas de calor, se tiene presente que existía la posibilidad de realizar este análisis con otras formas de estudiar las correlaciones, cómo presentó Yuri Grajales[77] en su proyecto con diagramas de dispersión y de barras para observar cómo se correlacionan las variables.

El desarrollo del método automático se realizó por medio del entorno Spyder por la facilidad que ofrece al programador con la codificación de los algoritmos, al ser amigable y tener un espacio visual para las gráficas y variables del algoritmo, así como se explicó anteriormente en el capítulo de desarrollo. Aspectos como estos convirtieron Spyder como el entorno de programación Python elegido para la construcción del método automático y la codificación de los modelos, en este mismo se generaron todas las gráficas, mapas de calor y resultados que fueron de utilidad durante el desarrollo y validación del método.

A pesar de tener una buena experiencia con el entorno Spyder, es importante mencionar que no es el único entorno de programación en Python, existen más entornos para la programación, por ejemplo, Jupyter Notebook y Pycharm, entornos con los cuales Yuri Grajales[78] en su proyecto, llevó a cabo el proceso de extracción de la información y el desarrollo de sus modelos de predicción de viviendas en el municipio de Rio negro, Medellín. Es por esto que, a pesar de los resultados obtenidos, se tiene en cuenta que el desarrollo se pudo haber llevado a cabo con Jupyter Notebook y de esta manera lograr los mismos resultados, a fin de cuentas, el desarrollo del método se realizó con el entorno Spyder y gracias a las pocas complejidades de codificación durante la implementación se logró el desarrollo del método automático.

[77] GRAJALES Yuri, op. Cit p.94
[78] GRAJALES Yuri, op. Cit p.94

Posteriormente, se puede observar que los modelos que tuvieron un mejor promedio en las medidas de desempeño, según el resultado de cada uno de los estratos mostrado previamente en la evaluación, fue el modelo de random forest, pues, como se observa en la tabla 8, este modelo obtuvo el mayor coeficiente de determinación igual a 0,93. Esto nos demuestra estadísticamente, que este modelo presenta un mejor ajuste al precio de las viviendas del Dataset. En cuanto al valor del RMSE, dicho modelo obtuvo un valor de $125'768.700, lo que representa una mejor aproximación del valor de vivienda predicho al real. Por último, el coeficiente de variación demostró que el modelo de random forest presenta una menor dispersión de los datos respecto a la media de los precios y el valor del RMSE.

	Regresion Lineal	Arboles de Decision	Random Forest	Red Neuronal
Promedio R^2	0,77	0,82	0,93	0,43
Promedio RMSE	376142332	213913139	125768700	594806075
Promedio CV	0,33	0,24	0,17	0,66

Tabla 8, Promedio de los resultados obtenidos por los modelos *Fuente: Propia*

Una vez observados los resultados obtenidos por cada uno de los modelos y determinado que el mejor modelo, según las medidas de desempeño empleadas, fue el modelo de random forest. Se decidió analizar y comparar los resultados obtenidos con los resultados de los modelos obtenidos por Yuri Grajales[79], esto se debe por las semejanzas que se presentan. Por ejemplo, la extracción de datos se realizó solo con *Finca Raíz*, las viviendas seleccionadas fueron casas y apartamentos en oferta de la ciudad de Bogotá, Colombia. Mientras que el otro caso de estudio realizó la extracción de datos en páginas web de *Finca Raíz* y *MercadoLibre,* también extrajo fincas, casas y apartamentos y en el municipio de Rio Negro, Medellín, Colombia. Por otro lado, se comparten modelos de Machine Learning similares como lo son los de regresión lineal, árboles de decisión y Random Forest. Por último, las medidas de desempeño utilizadas por ella son R^2 y el RMSE, similar a las presentadas en este proyecto. Los resultados obtenidos por Yuri Grajales se pueden evidenciar en la tabla 9, donde se muestran los modelos empleados y los resultados del coeficiente de correlación y de RMSE.

[79] GRAJALES Yuri, op, cit, p.95

	R2	RMSE
Regresión Lineal	0.530519	9.658456e+07
Decision Tree	0.551575	9.439382e+07
Random Forest	0.702875	7.437956e+07
Gradient Bosting Machine	0.711298	7.573973e+07
SVM	0.034413	1.385142e+08

***Tabla 9, Resultados generados por los modelos de Yuri Grajales** Fuente: GRAJALES Yuri, op, cit p.81*

Como se puede observar en la tabla 8, los resultados obtenidos superan los resultados de los modelos de Yuri Grajales, incluso con el hecho que el modelo de Random forest sea el mejor de ambos proyectos. Esto se debe a que el proyecto presentado por Grajales[80] presenta semejanzas, pero de la misma manera se diferencia de este caso de estudio

Es por esto que se presenta una diferencia entre ambos casos de estudio, pues los modelos presentados están entrenados y probados por conjuntos de datos diferentes. Es decir, aunque sean los mismos modelos, unos están entrenados con un Dataset de viviendas de casas y apartamentos en venta en la página de *Finca Raíz* y, por otro lado, un Dataset de viviendas de fincas, casas y apartamentos en la página de *Finca Raíz* y de MercadoLibre en el municipio de Rio Negro. Por otro lado, con el Dataset de viviendas de Bogotá, se realizaron más segmentaciones de datos para separar los datos por estrato para manejar una mejor correlación entre variables, a diferencia del Dataset de Rio Negro, donde Grajales[81] no realiza esta segmentación, en lugar de realizar un tratamiento con los datos atípicos, pues al ser eliminados pueden presentar una mejoría en los modelos.

A pesar de no haber implementado los modelos de Gradient Boosting Machine y las máquinas de soporte vectorial (SVM) se pudo demostrar que la segmentación por estratos del Dataset principal fue de gran utilidad para mejorar el desempeño de los modelos y obtener mejores resultados con las medidas de desempeño utilizadas para evaluar el desempeño de los modelos como se muestra en la tabla 10. A diferencia de cómo lo desarrollo Yuri Grajales en su proyecto con un solo conjunto

[80] GRAJALES Yuri, op, cit p.96
[81] GRAJALES Yuri, op, cit p.96

de datos con todos los estratos y teniendo presente que Río Negro tiene una población de 101.046[82] habitantes aproximadamente.

	Arboles de Decision			Random Forest			
	R^2	RMSE	CV	R^2	RMSE	CV	# DATOS
Estrato 2	0.94	5,368E+09	0.13	0.92	6,342E+09	0.15	144
Estrato 3	0.97	4,156E+09	0.12	0.93	5,742E+09	0.16	418
Estrato 4	0.75	1,195E+10	0.25	0.90	7,419E+09	0.15	467
Estrato 5	0.88	1,422E+10	0.18	0.90	1,296E+10	0.16	255
Estrato 6	0.58	712662302	0.52	0.92	3,042E+10	0.22	253
	Total Datos						1537

Tabla 10, Resultados generados por los modelos Fuente: Propia

Finalmente, como podemos observar en la tabla 10, se evidencian los modelos más relevantes del método automático, cada estrato finalizó con un número de datos distinto, es decir, el estrato 2 trabajo con un total de 144 datos, el estrato 3 con un total de 418 datos, el estrato 4 con un total de 467 datos, el estrato 5 con un total de 255 datos y el estrato 6 con un total de 253 datos, para un total de 1537 datos. Es importante recordar que el estrato 1 no se empleó para la ejecución de estos modelos, al ser un estrato que no aportaba valor por la poca información de datos recopilada.

Para concluir el método automático que se obtuvo en el desarrollo del presente proyecto es el siguiente:

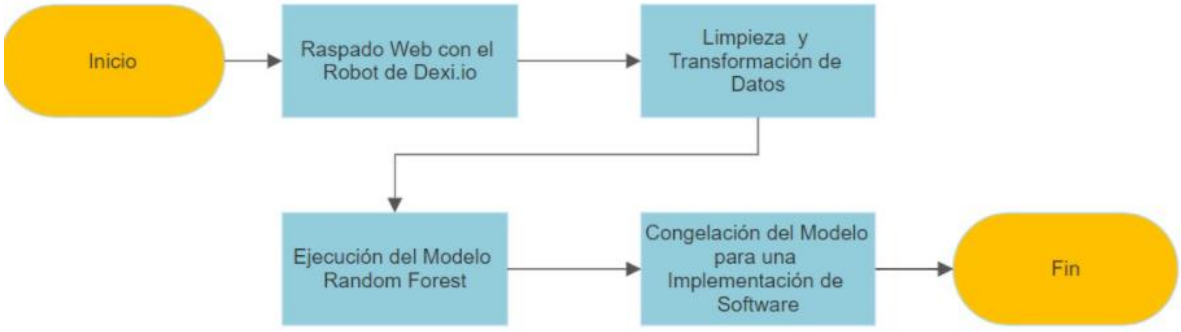

Figura 29, Diagrama de Desarrollo del Método Automático Fuente: Propia

[82] Municipios de Colombia, EL MUNICIPIO DE RIONEGRO, {En Línea} {13 de mayo de 2021} Disponible en (https://www.municipio.com.co/municipio-rionegro-ant.html) p.97

En donde la obtención de los datos se realizaría con una frecuencia de 1 vez por mes, para obtener los datos actualizados respecto al cambio de ofertas que se puedan presentar en el mercado. En la transformación de los datos, es importante destacar la limpieza de los datos en Python y la separación de los datos obtenidos por estratos, para asegurar la eficacia en el método, según los resultados obtenidos. Por último, se procede a la ejecución del modelo de mejor desempeño que según los datos obtenidos es Random forest, para cada estrato de Bogotá, para así congelar los modelos y brindarlos para su implementación ya sea en una aplicación de escritorio o una aplicación web.

Con este método automático que apoye a la predicción, se verían beneficiados tanto las personas que buscan vender sus inmuebles al poder publicar el valor de venta y saber con un nivel de certeza como esta su inmueble respecto a lo demás del sector, así como a los mismos peritos avaluadores, al ayudar a validar que el proceso que están realizando para obtener el valor de los inmuebles es acorde con el sector, poder dar un precio rápido de una vivienda a quien se lo pregunten sin tener que desplazarse hasta el inmueble para poder emitir un juicio. Por último, beneficia a la industria inmobiliaria, al estar un paso más cerca de la implementación de tecnologías de la industria 4.0 en la realización de sus procesos.

13 CONCLUSIONES

En este proyecto se logró el diseño, desarrollo e implementación de un método automático, a partir de la utilización de una técnica de Web Scraping, técnica proporcionada por la herramienta web DEXI.io para la extracción de la información. Esta información fue de gran utilidad para el entrenamiento, prueba y validación del método automático compuesto de 4 modelos a partir de técnicas de Machine Learning Random Forest, Árboles de decisión y regresión lineal) y Deep Learning (Redes neuronales profundas). Donde se eligió el mejor método a partir de las métricas de desempeño de R^2, RMSE y CV, se obtuvo que el mejor método para la predicción del avalúo comercial de un inmueble de vivienda en la ciudad de Bogotá es el de random forest, al presentar un promedio en su coeficiente de determinación superior a 0.9, su valor de RMSE en promedio fue igual al $125’768.700, y un coeficiente de variación de 0.17. Estas medidas demuestran cómo el modelo de random forest presenta un mayor ajuste a los precios de vivienda del conjunto de datos, una mayor cercanía entre los datos predichos con los observados y finalmente una menor dispersión.

A través del raspado web de datos con Dexi y de una transformación de los mismos utilizando Python, se escogió un Dataset con un total de 1549 registros y 10 columnas, de los cuales destacó el estrato, factor determinante para segmentar una vez más el Dataset en cada uno de estos (descartando el estrato 1 por la poca información aportada al proyecto), obteniendo pequeños segmentos de Dataset diferenciados por el estrato, acción que facilitó el cálculo y mejora de las medidas de desempeño generadas por los modelos.

Comparando con otro estudio realizado, los modelos generaron mejores medidas de desempeño gracias a la segmentación de los datos por estrato, esto logró una mejor correlación de variables y de la misma manera, un mejor desempeño en la predicción de precios de vivienda.

14 TRABAJOS FUTUROS Y RECOMENDACIONES

- Probar con otras herramientas de Web Scraping como son las librerías que ofrece Python para realizar este proceso, para así no limitar la cantidad de datos obtenidos a una suscripción pagada.
- Explorar otras páginas web para la obtención de información, como lo son *Metro Cuadrado, Properati* o *MercadoLibre*, debido a que la herramienta de extracción de datos fue DEXI, esta no permitía la visualización de dichas páginas porque se bloqueaba, el objetivo con esta propuesta es comparar la calidad de la información obtenida en distintas páginas, evaluar si se tienen similitudes en la información que presenta y un tratamiento de datos más sencillo que el que se presenta con *Finca Raíz.*
- Ampliar la cantidad de variables e información utilizada, teniendo en cuenta la información que se descartó en el trabajo como son las características interiores, exteriores y del sector que se pueden obtener de *Finca Raíz*, así como la descripción proporcionada por el vendedor dentro de la página, utilizando procesamiento del lenguaje natural para obtener variables cualitativas de cada vivienda para mejorar la calidad de la información y de la predicción.
- Experimentar con otros modelos de Machine Learning y Deep Learning, en donde se pueda comparar con los resultados obtenidos en este trabajo, ya que, de los métodos existentes, hubo varios que no se realizaron en el mismo, como lo puede ser el Gradient Boosting Machine (GBM) o máquinas de soporte vectorial (SVM), dentro del campo de Machine Learning, y dentro del campo del Deep Learning se pueden explorar diferentes redes neuronales, como las recurrentes o las convolucionales.
- A partir de los pasos seguidos para realizar el método automático, realizar un desarrollo de software, en donde se monte una aplicación en línea, que extraiga y actualice los datos periódicamente, segmentados por estratos e implementando el mejor modelo obtenido que en este caso fue Random Forest, para que cualquier persona pueda digitar las características de su vivienda y obtener el precio de la misma de manera rápida y gratuita.

15 BIBLIOGRAFÍA

Bin, J.a, Gardiner, B.b, Li, E.c, Liu, Z.a, 2019, Multi-source urban data fusion for property value assessment: A case study in Philadelphia, ScienceDirect, p.16

Rotimi Boluwatife Abidoye, Albert P.C. Chan and Funmilayo Adenike Abidoye, Olalekan Shamsideen Oshodi | 2018 | Predicting property price index using artificial intelligence techniques Evidence from Hong Kong | International Journal of Housing Markets and Analysis. P.16

Bin, J.a, Gardiner, B.b, Li, E.c, Liu, Z.a, Op.Cit, p.16

Rotimi Boluwatife Abidoye, Albert P.C. Chan and Funmilayo Adenike Abidoye, Olalekan Shamsideen Oshodi, Op.Cit. p.16

Junchi Bin, Bryan Gardiner, Zheng Liu | 2019| Attention-based multi-modal fusion for improved real estate appraisal: a case study in Los Angeles. P.16

Acuerdo 265 del 2018, Consejo superior de la Universidad Católica de Colombia, “Por el cual se aprueban los lineamientos y las opciones de grado para los programas de la Facultad de ingeniería de la Universidad Católica de Colombia, 12 de diciembre del 2018, Consejo superior de la Universidad Católica de Colombia. P.17

MINISTERIO DE LAS TECNOLOGÍAS DE LA INFORMACIÓN Y LAS COMUNICACIONES, Aspectos básicos de la industria 4.0, República de Colombia, OFICINA ASESORA DE PLANEACIÓN Y ESTUDIOS SECTORIALES, 2019. P.18

DIRECCIÓN DE MINERÍA EMPRESARIAL, Análisis del comportamiento del PIB minero primer trimestre de 2019, Bogotá D.C, 2019. P.18

PROPERATI, Disponible en: (www.properati.com). p.18

INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI SEDE CENTRAL, RESOLUCIÓN 620 DE 2008, Bogotá, Secretaría Jurídica Distrital, 2008. P.19

Ley 388 de 1997. Por la cual se modifica la Ley 9 de 1989, y la Ley 2 de 1991 y se dictan otras disposiciones. 18 de julio de 1997. D.O. No. 43091. P.19

SECRETARIA DISTRITAL DE PLANEACIÓN, Decreto Distrital 551 de 2019, Bogotá, Secretaria Distrital de planeación, 2019. P.19

Julián Pérez Porto y Ana Gardey. Definicion.de: Definición de inmueble, {En línea}. {23 de octubre del 2020} disponible en: (https://definicion.de/inmueble/). P.23

Rebajatuscuentas.com, ¿Qué es un inmueble y qué tipo de inmuebles existen?, {En línea}, {8 de mayo del 2021} disponible en: (https://rebajatuscuentas.com/pe/blog/que-es-un-inmueble) p.23

Julián Pérez Porto y Ana Gardey Definicion.de: Definición de vivienda, {En línea}. {23 de octubre del 2020} disponible en: (https://definicion.de/vivienda/) p.23

Diccionarioactual.com, ¿Qué es vivienda?, {En línea}, {8 de mayo del 2021}, disponible en: (https://diccionarioactual.com/vivienda/) p.24

SUÁREZ, Alexandra, Antes de vender o comprar, piensa en hacer un avalúo. {En línea}, {27 de octubre 2020}, disponible en: (https://www.metrocuadrado.com/noticias/guia-de-compra/antes-de-vender-o-comprar-piensa-en-hacer-un-avaluo-2863) p.24

GRAJALES Yuri, Modelo predicción precios viviendas proyecto Medellín, Medellín, 2020, 74 Págs, Modelo predicción precios viviendas proyecto Medellín, Facultad de Ingeniería, Departamento en Tecnologías++ de la Información y Comunicación. p.25

REALIA. ¿Qué es el mercado inmobiliario?, {En línea}. {3 de mayo 2021} Disponible en (https://www.realia.es/que-es-mercado-inmobiliario) p.25

Equipo de redacción de OIKOS, Perspectivas del sector inmobiliario para este 2021, {En Línea}.{3 de mayo 2021} Disponible en (https://www.oikos.com.co/inmobiliaria/noticias-inmobiliaria/como-va-el-sector-inmobiliario) p.25

BALAGUERO Thais, ¿Qué son los Dataset y los dataframe en el Big Data?, {En Línea} {13 de mayo de 2021}, Disponible en (https://www.deustoformacion.com/blog/programacion-tic/que-son-datasets-dataframes-big-data). P.25

GARCÍA, Carlos, ¿Qué es el Deep Learning y para qué sirve? {en línea}, {23 de octubre de 2020}, Disponible en (https://www.indracompany.com/es/blogneo/deep-Learning-sirve) p.26

GARCIA Carlos, ¿Qué es el Machine Learning?Ibíd. P.26

ZAMBRANO, Juan, ¿Aprendizaje supervisado o no supervisado? Conoce sus diferencias dentro del Machine Learning y la automatización inteligente, {En línea}, {23 de octubre de 2020}, Disponible en (https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b) p.27

SANTOS Paloma, Tipos de aprendizaje en Machine Learning: supervisado y no supervisado, {En Línea}, {10 de mayo de 2021}, Disponible en https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/) p.27

SANTOS Paloma, Op, Cit, p.27

GARCÍA, Carlos. Op. Cit. P.28

GARCÍA, Oscar, Redes Neuronales artificiales: Qué son y cómo se entrenan, {En línea} {23 de octubre del 2020}, Disponible en (https://www.xeridia.com/blog/redes-neuronales-artificiales-que-son-y-como-se-entrenan-parte-i) p.28

GARCÍA, Oscar. Op. Cit. P.29

Julián Pérez Porto y María Merino. Definicion.de: Definición de página web, {En línea}. {23 de octubre del 2020} disponible en: (https://definicion.de/pagina-web/) p.29

LAFUENTE, Ainhoa, Que es el web Scraping, {En Línea}, {23 de octubre 2020} Disponible en (https://aukera.es/blog/web-scraping/) p.30

LAFUENTE, Ainhoa, Bases de datos relacionales vs. no relacionales: ¿qué es mejor?, {En Línea},{23 de octubre 2020 }Disponible en (https://aukera.es/blog/bases-de-datos-relacionales-vs-no-relacionales/) p.30

Ayudaleyprotecciondatos.es, Base de datos no relacional. ¿Qué es? Características y ejemplos, {En línea}, {9 de mayo. de 2021}, Disponible en (https://ayudaleyprotecciondatos.es/bases-de-datos/no-relacional/) p.31

FERNÁNDEZ, Rubén, MongoDB: qué es, cómo funciona y cuándo podemos usarlo (o no), {En Línea},{23 de octubre 2020 }Disponible en (https://www.genbeta.com/desarrollo/mongodb-que-es-como-funciona-y-cuando-podemos-usarlo-o-no) p.31

ROBLEDANO Ángel, Qué es MongoDB, {En línea}, {9 de mayo de 2021} Disponible en (https://openwebinars.net/blog/que-es-mongodb/) p.31

Base de Datos de Scopus. Disponible en: (https://www-scopus-com.ucatolica.basesdedatosezproxy.com/search/form.uri?display=basic#basic) p. 32

Chao Ma, Zhenbing Liu, Zhiguang Cao, Wen Song, Jie Zhang, Weiliang Zeng, Cost-sensitive Deep Forest for Price Prediction, Pattern Recognition (2020), doi: https://doi.org/10.1016/j.patcog.2020.107499. P.32

Bin, J.a, Gardiner, B.b, Li, E.c, Liu, Z.a, Op. Cit. P.33

Boyko, A.N. Email Author, Gladyshev, A.G. Email Author, Kadyrova, G.M. Email Author, Barmenkova, N.A. Email Author, Zybenko, A.V. 2020 Predicting the cost of housing using neural networks. P.33

A Bojanowska1 and J Lipski | 2019 | The use of data by smart systems for price forecasting in the context of building customer relationships on the Lublin real estate market | IOP. P.33

Yong Piao, Ansheng Chen, Zhendong Shang | 2019 | Housing Price Prediction Based on CNN | ICIST. P.33

Yao Sun, | 2019 | Real Estate Evaluation Model Based on Genetic Algorithm Optimized Neural Network | Data Science Journal. P.34

Junchi Bin, Shiyuan Tang, Yihao Liu, Gang Wang, Bryan Gardiner, Zheng Liu, Eric Li | 2017 | Regression model for appraisal of real estate using recurrent neural network and boosting tree | IEEE Intternational Conference. P.34

Poursaeed, O. Matera, T. Belongie, S.| 2018 | Vision-based real estate price estimation | Machine Vision and Aplications. P.35

Rotimi Boluwatife Abidoye, Albert P.C. Chan and Funmilayo Adenike Abidoye, Olalekan Shamsideen Oshodi. Op. cit. P.35

GRAJALES, Yuri, Modelo predicción precios viviendas proyecto Medellín, Op, Cit, p.36

FINCA RAÍZ.COM, ¿Quiénes Somos?, {en línea}, {25 de octubre 2020}, disponible en (https://www.fincaraiz.com.co/). P.37

DEXI .IO SOLUTIONS {En Línea} {29 de octubre 2020} Disponible en (https://www.dexi.io/solutions/). P.38

DEXI DATA Solutions - ETL Engine 2015. P.38

MARTINEZ, José, “15 Librerías de Python para Machine Learning” {En Línea} {29 de octubre 2020} Disponible en (https://www.iartificial.net/librerias-de-python-para-Machine-Learning/) p.39

Ibíd. P.40

MARTINEZ, José. Op, Cit. P.41

Python Data Analysis Library – pandas: Python Data Analysis Library». pandas. {En Línea} {29 de octubre 2020} disponible en (https://pandas.pydata.org/pandas-docs/stable/index.html). P.41

LANARO, Gabriele, Python high performance: build robust application by implementing concurrent and distributed processing techniques (Second edition edition). ISBN 978-1-78728-243-8. OCLC 990086907. {En Línea} {29 de octubre 2020} Disponible en (https://www.worldcat.org/title/python-high-performance-build-robust-application-by-implementing-concurrent-and-distributed-processing-techniques/oclc/990086907). P.42

MARTÍNEZ, José. Op. Cit. P.42

BUHIGAS, Javier, Todo lo que necesitas saber sobre TensorFlow, la plataforma para Inteligencia Artificial de Google 2018, {En línea}, {29 de octubre 2020} Disponible en: (https://puentesdigitales.com/2018/02/14/todo-lo-que-necesitas-saber-sobre-tensorflow-la-plataforma-para-inteligencia-artificial-de-google/). P.42

LANARO, Gabriele, Python high performance: build robust application by implementing concurrent and distributed processing techniques (Second edition edition). ISBN 978-1-78728-243-8. OCLC 990086907. {En Línea} {29 de octubre 2020} Disponible en (https://www.worldcat.org/title/python-high-performance-build-robust-application-by-implementing-concurrent-and-distributed-processing-techniques/oclc/990086907). P.42

MARTÍNEZ, José. Op. Cit. P.42

BUHIGAS, Javier, Todo lo que necesitas saber sobre TensorFlow, la plataforma para Inteligencia Artificial de Google 2018, {En línea}, {29 de octubre 2020} Disponible en: (https://puentesdigitales.com/2018/02/14/todo-lo-que-necesitas-saber-sobre-tensorflow-la-plataforma-para-inteligencia-artificial-de-google/). P.42

MARTÍNEZ, José. Op. Cit. P.43

ORTIZ Moisés, Que es Excel y para qué sirve, {En Línea], {10 de mayo de 2021} {Disponible en: https://exceltotal.com/que-es-excel/} p.46

GUJARATI Damodar, PORTER Dawn. Econometría: Mínimos cuadrados ordinarios. Quinta Edición. México: Mc Graw Hill, 2010. p.47

Damodar N. Gujarati. Econometría, Op, cit. P.48

GRAJALES Yuri, Modelo predicción precios viviendas proyecto Medellín, Op, Cit, p.50

CAMARA DE COMERCIO DE BOGOTÁ, Como ubicar tu empresa en Bogotá desde el punto de vista administrativo, {En línea} {9 de mayo 2021}, disponible en (http://recursos.ccb.org.co/ccb/pot/PC/files/ley388.html#:~:text=En%20el%20a%C3%B1o%201997%20el,respectivos%20Planes%20de%20Ordenamiento%20Territorial.) p.51

Ley 388 de 1997.Op. Cit . P.51

ALCADÍA DE BOGOTÁ. Documentos para PLAN DE ORDENAMIENTO TERRITORIAL :: Estatutos Orgánicos. {En línea}. {10 de mayo 2021}. Disponible en: (https://www.alcaldiabogota.gov.co/sisjur/listados/tematica2.jsp?subtema=21178#:~:text=El%20Decreto%20Distrital%20190%20de,es%20decir%2C%20no%20ha%20sido) p.52

ALCADÍA DE BOGOTÁ. Documentos para POT UNIDADES DE PLANEAMIENTO ZONAL -UPZ- :: Reglamentación. {En línea}. {10 de mayo 2021}. Disponible en: (https://www.alcaldiabogota.gov.co/sisjur/listados/tematica2.jsp?subtema=21291&cadena=p#:~:text=Decreto%20159%20de%202004%20Alcald%C3%ADa,Unidades%20de%20Planeamiento%20Zonal%20%E2%80%93UPZ.&text=Relaci%C3%B3n%20de%20las%20Unidades%20de,uno%20de%20los%20decretos%20respectivos.) p.53

Departamento Administrativo Nacional de Estadística – DANE. Estratificación socioeconómica. {En línea}, {9 de mayo 2021}. Disponible en (https://www.dane.gov.co/index.php/69-espanol/geoestadistica/estratificacion/468-estratificacion-socioeconomica). P.53

SECRETARIA DISTRITAL DE PLANEACIÓN, Decreto Distrital 551 de 2019, Op. Cit. P.54

INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI SEDE CENTRAL. Op. Cit, p.55

MARTINES H José, Random Forest (Bosque Aleatorio): combinando árboles, {En Línea} {11 de mayo de 2021} {Disponible en: https://www.iartificial.net/random-forest-bosque-aleatorio/#Diferencia_intuitiva_entre_un_arbol_de_decision_y_un_random_forest}. P.74

MARCO S Francisco Javier. Coeficiente de variación {En Línea} {11 de mayo de 2021} {Disponible en: https://economipedia.com/definiciones/coeficiente-de-variacion.html} p.79

GRAJALES Yuri, op. Cit p.93

Yong Piao, Ansheng Chen, Zhendong Shang, Op, Cit. p.93

Bin, J.a, Gardiner, B.b, Li, E.c, Liu, Z.a, op. cit p.93

GRAJALES Yuri, op. Cit p.93

GRAJALES Yuri, op. Cit p.94

GRAJALES Yuri, op. Cit p.94

GRAJALES Yuri, op, cit, p.95

GRAJALES Yuri, op, cit p.96

GRAJALES Yuri, op, cit p.96

Municipios de Colombia, EL MUNICIPIO DE RIONEGRO, {En Línea} {13 de mayo de 2021} Disponible en (https://www.municipio.com.co/municipio-rionegro-ant.html) p.97

16 ANEXO 1: CÓDIGO UTILIZADO PARA LA TRANSFORMACIÓN DE LOS DATOS.

En el siguiente enlace se podrá encontrar el código utilizado para procesar y organizar la información obtenida, se utilizaron las librerías de pandas y string para poder trabajar los datos, el código se encuentra con comentarios de los procesos realizados para el procesamiento de la información.

Link:
https://colab.research.google.com/drive/14iYlJBcsVelNyQUtqcnIbcEhn0lHhag_?usp=sharing

Nota: Para poder visualizar el código es necesario una cuenta de Gmail diferente a la proporcionada por la Universidad Católica de Colombia, preferiblemente una cuenta personal de Gmail, dado a los permisos que tienen esta cuenta con respecto a Google Colaboratory.

17 ANEXO 2. DATASET GENERADOS AL FINAL DE PROCESAMIENTO

Como se explicó durante la transformación de los datos se generaron 3 Dataset, los cuales se pueden encontrar en el siguiente enlace de drive.

Enlace:

https://drive.google.com/drive/folders/1tbOjdOWJxWuXJwD8mFwrG6xw-LyjYXOK?usp=sharing

18 ANEXO 3: DESARROLLO DEL MÉTODO AUTOMÁTICO

Para el desarrollo del método automático se implementaron un total de 4 modelos de Machine Learning, como primer modelo consta de una red neuronal artificial, para esto se empezó el proyecto trabajando en el entorno Python proporcionado por Spyder, el cual a su vez es proporcionado por Anaconda; de esta manera en una nueva página en blanco se comenzó importando las librerías básicas que son de utilidad para el desarrollo de estos modelos de Machine y Deep Learning, empezando por la librería TensorFlow con el código “import TensorFlow, y trabajándolo en el código como “tf”, TensorFlow es una de las librerías principales que vuelven el desarrollo de los métodos más sencillos por el conjunto de herramientas que contiene junto con Keras, continuamente se importó la librería de pandas (import pandas as pd) y aplicándola en el código con su abreviatura de pd, esta librería es la especializada en el análisis y manejo de estructuras de datos, la cual fue la ideal para la lectura de los datos de bienes raíces extraídos accediendo a ellos por sus filas y columnas, seguida de esta se importó la librería de Numpy presentada en el código con su abreviatura de np (import Numpy as np) Numpy es una potente librería muy necesaria para el manejo de grandes estructuras de datos que proporciona la implementación de matrices o vectores, después se implementó la librería Seaborn y se trabajó en el código con su abreviatura de sns (import Seaborn as sns) la cual es una librería que se utilizó para la visualización de las gráficas y la personalización de las mismas, también se implementó la librería de matplotlib.pyplot, presentada e invocada en el código con su abreviatura de plata (import matplotlib.pyplot as plt) la cual fue de utilizada para la generación de gráficas a partir de los datos que son almacenados en vectores o Arrays.

De este modo, una vez con las librerías ya declaradas, se empezó con el análisis y visualización de los datos en su archivo csv, esto es posible gracias a la librería de pandas, el cual con su función “read_csv” y de esta manera se leyó el archivo csv, y se almacenó en un DataFrame “df”, este procedimiento y los que se nombran a continuación se realizaron igualmente con los 5 Dataset según cada estrato.

Después se creó una matriz llamada “X” en donde con un DataFrame se crea un marco de datos con todas las variables que se relacionan con el precio, es decir las variables con las que se predice el precio de vivienda, esto se realiza de la siguiente manera con la línea de código “ X = df.drop[‘Precio’] ” esto por lado de las variables independientes, y ahora en cuanto a la variable dependiente “Y” se encuentra el target o el precio, el cual fue el valor a predecir a partir de las variables independientes, y para asignar el precio a la variable “Y” siendo el precio que depende de las características seleccionadas previamente se asigna de la siguiente manera “ y= df[‘Precio’] ” y de esta manera ya se tienen las variables dependientes e independientes almacenadas en las variables **y X** respectivamente.

Ahora bien, con los datos una vez escalados, se separaron el marco de datos en dos conjuntos uno para el entrenamiento conformado por el 70% de los datos y otro para prueba conformado por el 30% restante, esto se asignan en variables como X_train, X_test, y_train, y_test y gracias a la función “train_test_split” de la librería sklearn.model_selection y se le asignaron las variables del marco de datos que contiene las selected features y los precios, pero escalados.

Ahora se procedió a definir el modelo, este modelo consta de una red que se construyó de manera secuencial el cual permite agregar las capas en manera de secuencia, estas capas se crean gracias a la API de Keras de la librería de TensorFlow, cada capa cuenta con un total de 100 neuronas exceptuando la capa de salida el cual es una sola unidad y la primera capa siendo la capa de entrada cuenta con tamaño de entrada de 9 siendo este el número de selecta features que se procesarán dentro de la red neuronal. Todas las capas tienen como función de activación la función Relu, esta función es la más utilizada puesto que permite el aprendizaje muy rápido en las redes neuronales, y la función lineal para la capa de salida y de esta manera la red neuronal generará un valor único, en este caso el valor de la vivienda predicho.

Ahora para optimizar la red neuronal se modificaron los híper parámetros, siendo estos los parámetros que se puede modificar y de esta manera ajustar la velocidad con la que la red neuronal está aprendiendo, siendo un primer híper parámetro el optimizador, el optimizador utilizado es el optimizador Adam el cual combina dos buenos optimizadores como lo son el RMSprop y el Adagrad siendo uno de los mejores optimizadores para el Deep Learning, otro híper parámetro que se modificó es la función de pérdida, aplicando la función del error cuadrático medio.

Después se procedió al entrenamiento del modelo, el entrenamiento del modelo se realiza con el parámetro “.fit”, posteriormente, se ajustan los argumentos con las variables que se han manejado, comenzando por los datos de entrada y los datos de destino, está el batch size de 50, el cual equivale al número de muestras o el lote asignado que entrenara durante cada época, después se asignaron el número de épocas trabajando con un total de 80 épocas.

Ya para finalizar el entrenamiento se realizó una gráfica para visualizar el progreso de la red durante el entrenamiento, en el cual se observó como la pérdida disminuye a razón del número de épocas, dicha grafica se puede visualizar a continuación.

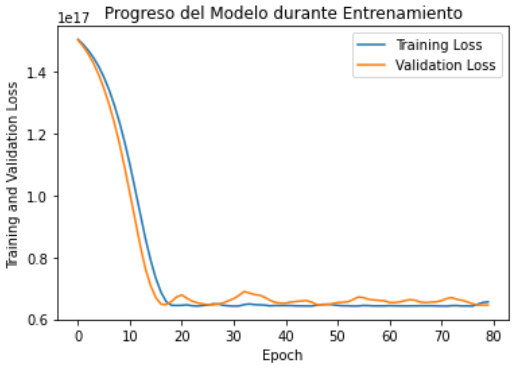

Figura del anexo 1 No 1, Progreso de la Red durante el Entrenamiento Fuente: Propia

Para finalizar se realizó la predicción del conjunto de datos de prueba y visualizando el resultado en una nueva gráfica.

El desarrollo de los modelos restantes, se empezó de igual manera con la importación de las librerías necesarias.

Posteriormente antes de proceder con el modelamiento de los datos, se realizó una partición de los datos previamente almacenados en las variables “X” y “y” previamente explicadas con el 70% de los datos para entrenamiento, y el 30% restante para pruebas del mismo modo como se menciona anteriormente con la red neuronal.

Una vez definidos los modelos de Machine Learning se procedió a la codificación y ejecución de cada uno de estos para su posterior evaluación de calidad en función a las predicciones realizadas por los modelos y los datos reales de las viviendas, la selección de estos métodos fue su relación con lo simple para la interpretación de los resultados, es por esto que se desarrollaron en su orden los siguientes modelos:

- Regresión lineal simple.
- Árboles de decisión.
- Random Forest.

Es importante mencionar que gracias a la librería de sklearn y las funciones que esta proporciona, convierte el desarrollo de estos últimos modelos de una manera más sencilla.

En el siguiente enlace se encuentra el repositorio de GitHub donde se pueden encontrar los Dataset y el código utilizado para la ejecución de los modelos en el entorno Spyder:

https://github.com/DNMSnicolas999/Metodo-Automatico-para-la-Prediccion-de-Precios-de-Vivienda-.git