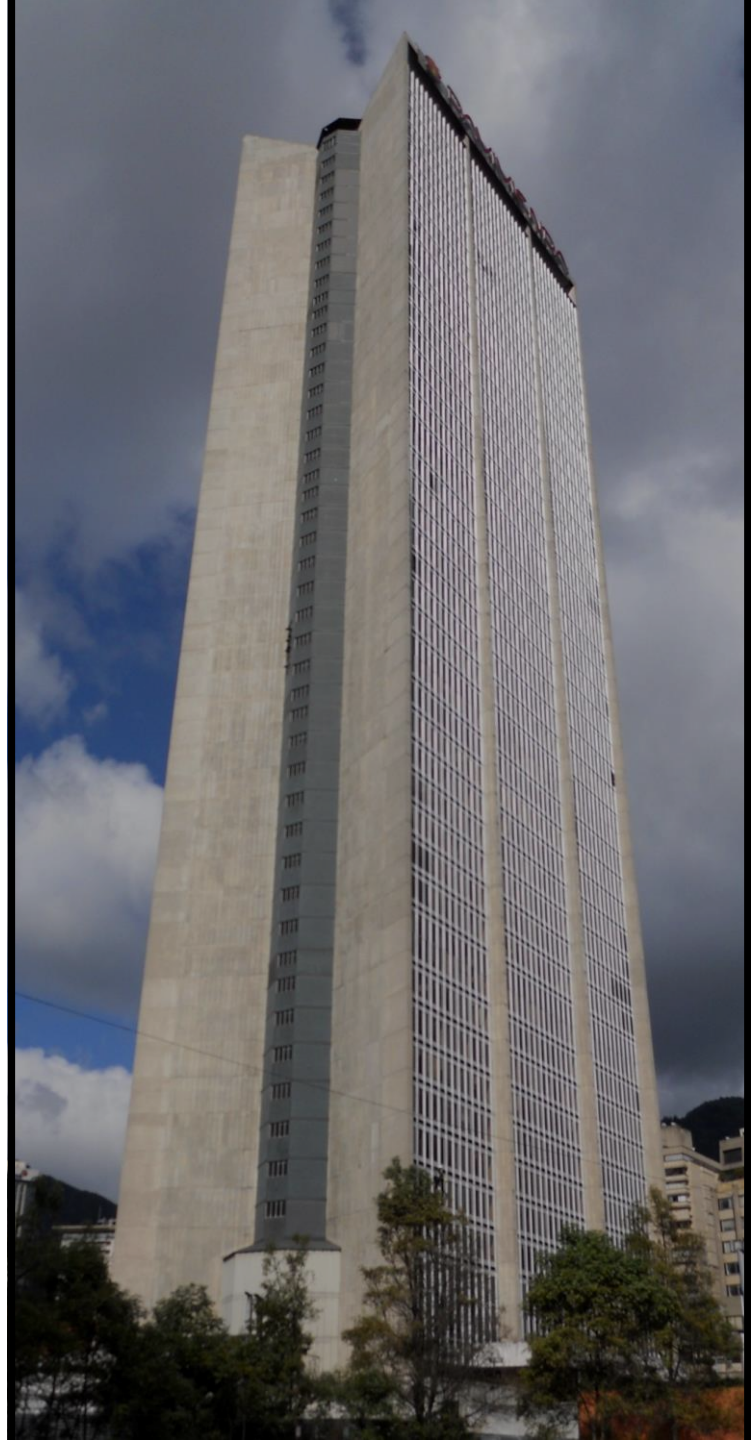
MODELO DE PREDICCIÓN DE PRECIOS DE VIVIENDA

PRESENTACIÓN: PRUEBA DE CONOCIMIENTOS – DATOS NO ESTRUCTURADOS

Nicolás Guerrero Caballero

Septiembre - 2022

Índice



1. Análisis y Detección de dificultades en la Base de Datos.

Al importar la base de datos `test_precios_vivienda.csv` mediante el comando `read_csv` se detectaron las siguientes características particulares de la base de datos “cruda”:

- Formato de separación decimal combinado: Uso de `.` a la par de `,` para indicar decimales.
- Comillas de texto combinadas: Uso de `"` y de `"""` para denotar texto.
- Valor Nulo combinado: Uso de `' '` y `0` como indicativo de valor faltante.
- Múltiples codificaciones: Se detecto el uso de codificación `UTF-8` y `Windows-1252.`
- Número de Columnas difiere entre registros.

De estas características se derivan cuatro problemas principales:

- Formato de columna erróneo (Ej. Columna `float` tomaba como `object` a causa del separador decimal).
- Número de registros sobreestimado en algunas variables a causa del uso de dos tipos de valores nulos.
- Imposibilidad de clasificar a priori el tipo de variables (Numérica / Dummy / Categórica / etc.)
- Datos en columnas incorrectas (Ej. Registro de número de habitaciones en la variable Piscina).

2. Limpieza y Manipulación de Datos

Con el fin de corregir las particularidades de la BD encontradas en el punto 1 se implementaron las siguientes acciones en orden:

Datos en Columnas Incorrectas

La alta cantidad registros con más o menos columnas que el estándar no permitía generar una solución general para todos los registros en un rango de tiempo considerable por lo que se optó por detectarlos y eliminarlos de la muestra.

¿Cómo?:
Se detectó cuáles variables eran fundamentalmente numéricas (Ej, área, precio) y se revisó si contenían respuestas tipo string, Si era el caso, el registro se suprimía de la muestra.

¿Por qué se tomó como referencia la variable numérica?
Debido a que los demás tipos de variables pueden contener respuesta en texto sin que esto sea necesariamente un error.

Clasificación de variables

1. Dummy: Se verificó si la columna registraba respuestas afirmativas y negativas conjuntamente. Además se verificó que no hubieran más de tres tipos de respuesta (Si/No/0).
2. Texto: Se tomaron las variables cuyo nombre contenían las palabras (descripción/observación/etc.)
3. Numéricas: Se tomaron las respuestas de la columna y se aplico el formato float, si se permitía, se denotaba a la variable como numérica.
4. Categórica: se revisó que la columna no tuviera más de 40 respuestas pues un mayor número respuestas en una variable categórica es poco común.

Formato de Columna

1. Dummy:
Las respuestas afirmativas se tomaron como 1, las negativas como 0 y los ceros se transformaron a NaN.

2. Categóricas:
Se transformaron a dummy como indicador de la categoría (Ej. Bogotá: 1 si pertenece / 0 si no) y los ceros se tomaron como NaN.

3. Texto:
Se decodificó según el tipo: UTF-8 o Windows-1252.

4. Numéricas:
Se normalizaron los separadores decimales. A falta de información no se pudo tomar el valor cero como nulo o dato faltante.

Nota: Al dar un formato a los valores Nulos, se soluciona el problema sobre la sobreestimación de los registros por variable.

2. Limpieza y Manipulación de Datos

Una vez se obtuvo la base de datos con formato correcto y los datos bien ubicados, se preparara la base para el paso al análisis exploratorio descartando algunas variables mediante los siguientes criterios:

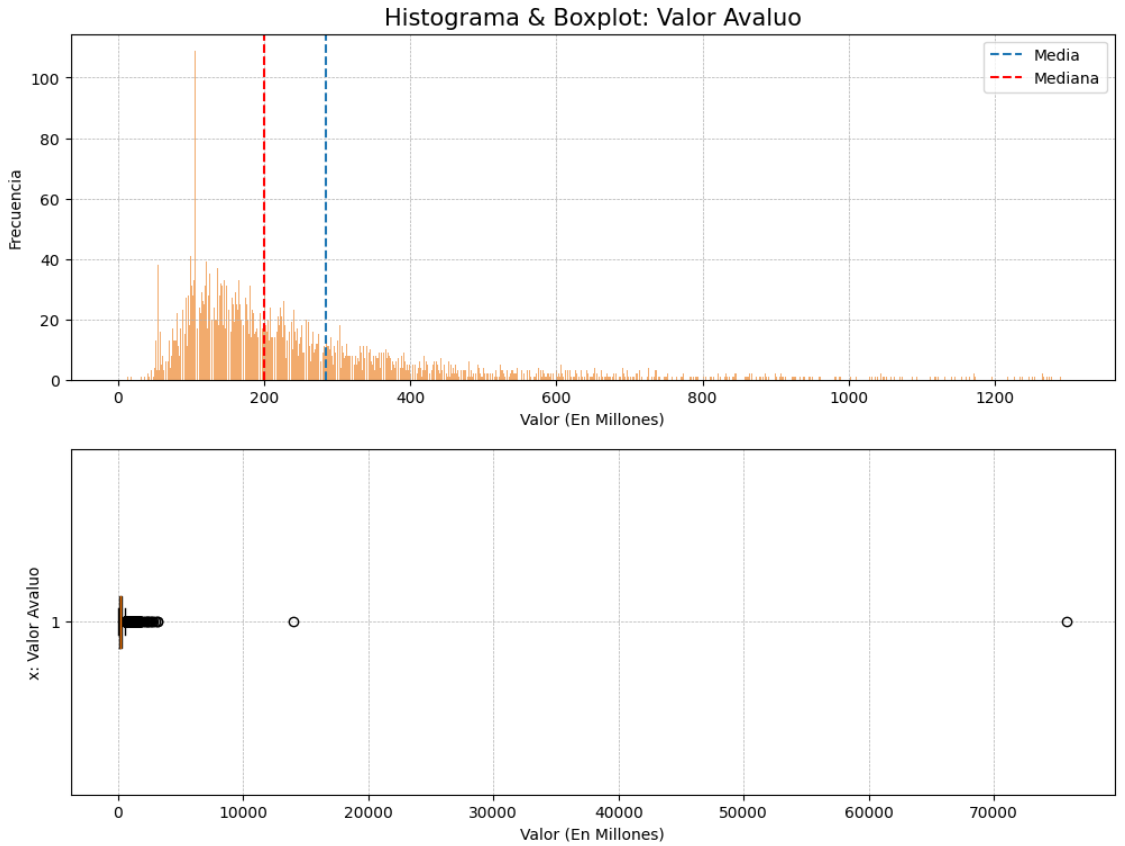
Tipo de Variables	Concepto/Definición	Razón
Variables Enteramente Aleatorias	Para el caso de estudio se definen las variables aleatorias como aquellas que están ligadas a un determinado registro por obra del azar y que teóricamente no guardan ninguna relación causal con la variable de interés. Bajo esta definición se suponen como variables enteramente aleatorias las matriculas inmobiliarias, número de dirección, el número de garaje, número de parqueadero (no confundir con la cantidad de garajes, parqueaderos, etc.)	Al no tener una relación teórica, el estudio de estas variables no afecta positivamente el comportamiento del modelo de predicción. Contrariamente, pueden detectarse relaciones espurias que introduzcan sesgos el modelo.
Variables con varios Registros Nulos	Se separan de la muestra aquellas variables cuya cantidad de valores nulos sea mayor a 70%.	Estas variables se eliminan puesto que modelos como el Random Forest o el de Regresión Lineal no admiten valores nulos, lo que se traduce en que se toman sólo los registros completos, por lo que mantener las variables implica perder el 70% o más de la información al ejecutar estos modelos.

3. Análisis Descriptivo y Exploratorio

3.1. Análisis de la Variable Independiente

Del análisis gráfico y de los estadísticos descriptivos obtenemos las siguientes inferencias respecto a la variable dependiente:

- Los datos no se distribuyen de manera normal, por lo que se debe optar por el uso de estadísticos no paramétricos para analizar el comportamiento de la variable.
- La variable deberá ser normalizada con el fin de implementar algún modelo probabilístico.
- La desviación es bastante alta lo que implica una gran variabilidad en los datos.
- Existe outliers bastante alejados de la muestra por lo que se pueden llegar a incluir sesgos en algunos modelos si estos datos no son tratados adecuadamente.



Gráfica 1. Histograma y Box-Plot de la variable dependiente con outliers.

Mean	Median	Median Abs. Desv.	Skew	Kurtosis
284.87	200.38	81.58	69.94	5309.09

Tabla 1. Estadísticas no paramétricas (excepto por la media) de la variable dependiente

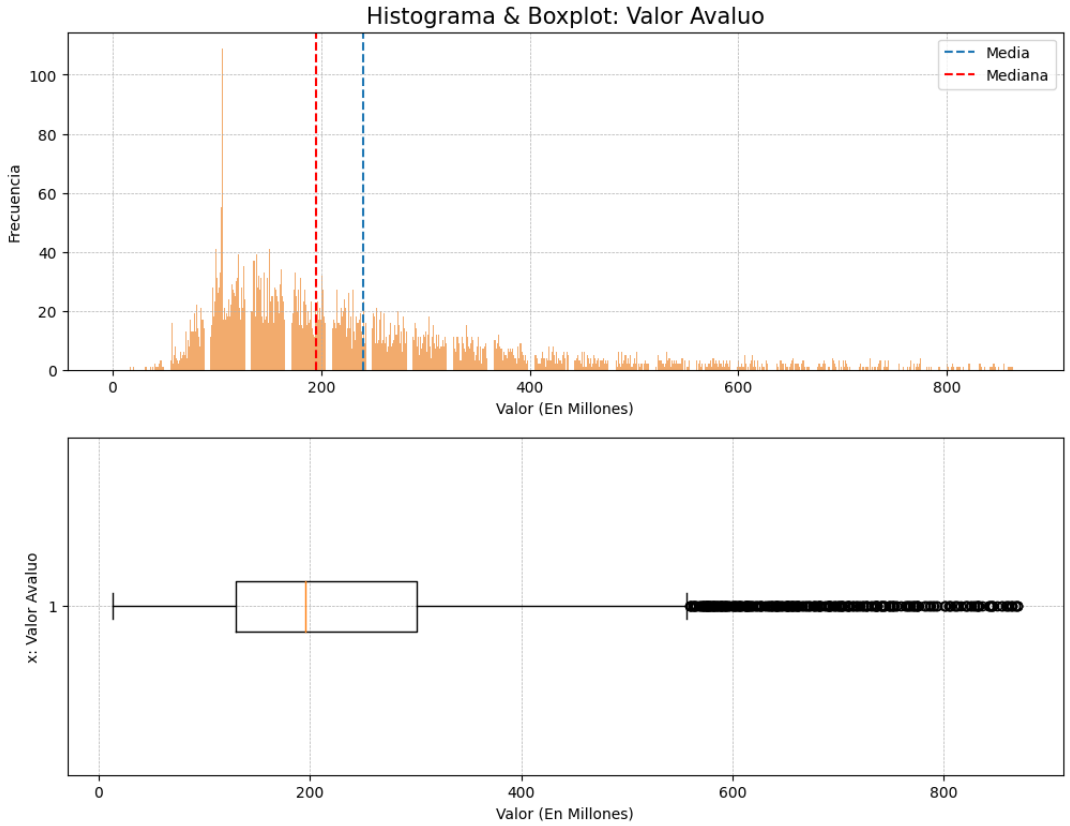
3. Análisis Descriptivo y Exploratorio

3.2. Análisis de Outliers

Del Box-Plot revisado anteriormente se detectó que había al menos dos valores (visibles) extremos. Si bien pueden ser inmuebles con características especiales, el estudio del mismo queda fuera del alcance de este desarrollo por lo que se opta por eliminar los valores detectados bajo la metodología Box-Plot:

$$x_i^{Extremo} \begin{cases} si\ x_i > Q_1 + IQR * 3 \\ si\ x_i < Q_1 - IQR * 3 \end{cases}$$

Es así como resulta la gráfica 2.
De aquí se puede constatar que los precios de los inmuebles son mucho más consistentes, eliminando posibles sesgos que se incluían con los outliers.



Gráfica 2. Histograma y Box-Plot de la variable dependiente luego de retirar outliers.

3. Análisis Descriptivo y Exploratorio

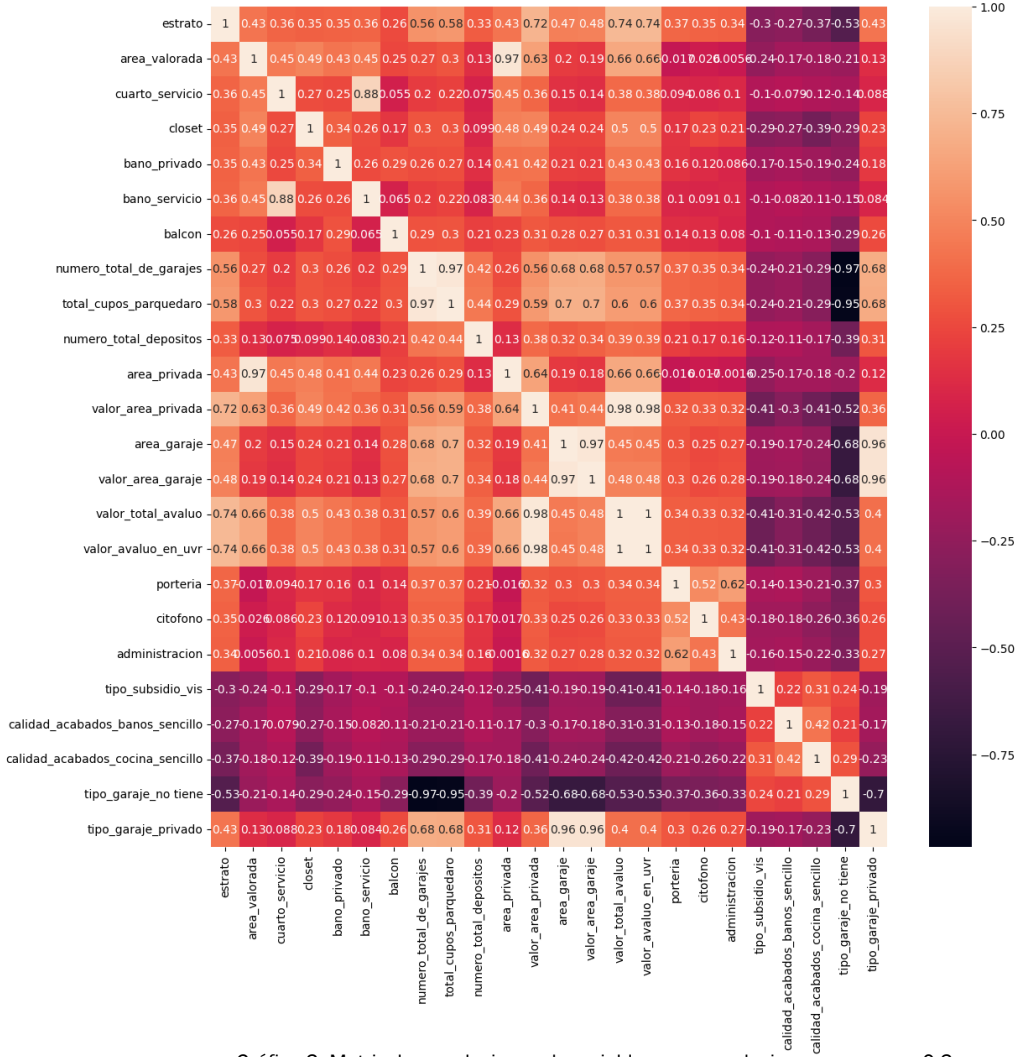
3.3. Análisis de Correlación de las Variables Independientes

Debido a la dimensión de las variables independientes, no es posible analizarlas una por una, por lo que, para detectar las variables más relevantes para predecir el precio de un inmueble, se usó el análisis de correlación de las variables y se extrajeron aquellas variables cuyo valor en términos absolutos fuera mayor a 30%.

De este primer análisis se concluye que existen al menos dos variables cuya implementación se puede no ser efectiva al incluirse dentro de los modelos de predicción por tener una correlación muy alta:

- Valor Área Privada
- Valor Avaluó en UVR

Desde el punto de vista estadístico, en un modelo de regresión lineal, incluir estas variables puede generar problemas de multicolinealidad. En la práctica, el uso de estas variables de valor es inviable puesto que para la predicción de precios de inmuebles no se puede esperar tener de antemano el valor del inmueble en UVR o el valor de las distintas áreas de modo que quedan excluidas de los modelos.



Gráfica 3. Matriz de correlaciones de variables con correlaciones mayores a 0,3.

4. Modelos de Predicción de Precios

4.1. Definición de los modelos

Para predecir el valor total del avalúo se optó por utilizar tres metodologías distintas según el conjunto de datos de entrenamiento que se utilice como se muestra a continuación.

- Decision Tree
- Random Forest
- Regresión Lineal Múltiple

Para los dos primeros modelos se utilizará la muestra de datos completa mientras que para el modelo de regresión lineal se hará uso de las variables más significativas analizadas en el punto 3.3

4. Modelos de Predicción de Precios

4.2. Resultado y Comparación de los modelos

Metric	Decision Tree	Random Forest	RLM
MAPE	0.260	0.153	0.297
RMSE	78,712,926.2	53,483,814.5	92,594,827.0
Cross Validation Score	0.739	0.877	0.677
R2	0.750	0.884	0.654

Tabla 2. Métricas de los Modelos Entrenados.

Al revisar los resultados de predicción de los modelos, se observa que el modelo Random Forest tuvo mejor comportamiento que los demás modelos en todas sus métricas. Específicamente, utilizando la métrica MAPE (que mide porcentualmente la media del error del modelo) vemos que en promedio la diferencia entre un precio predicho y su valor real es de 15,3% valor que dista del modelo Decision Tree y RLM cuyos errores tienden a ser alrededor de un tercio del valor real.

Adicionalmente, se puede detectar que la aproximación mediante regresión lineal tiene un comportamiento similar al modelo Decision Tree con respecto a la medida MAPE, por lo que las variables seleccionadas mediante el análisis de correlación fueron acertadas en cierta medida para predecir el valor total de avaluó.

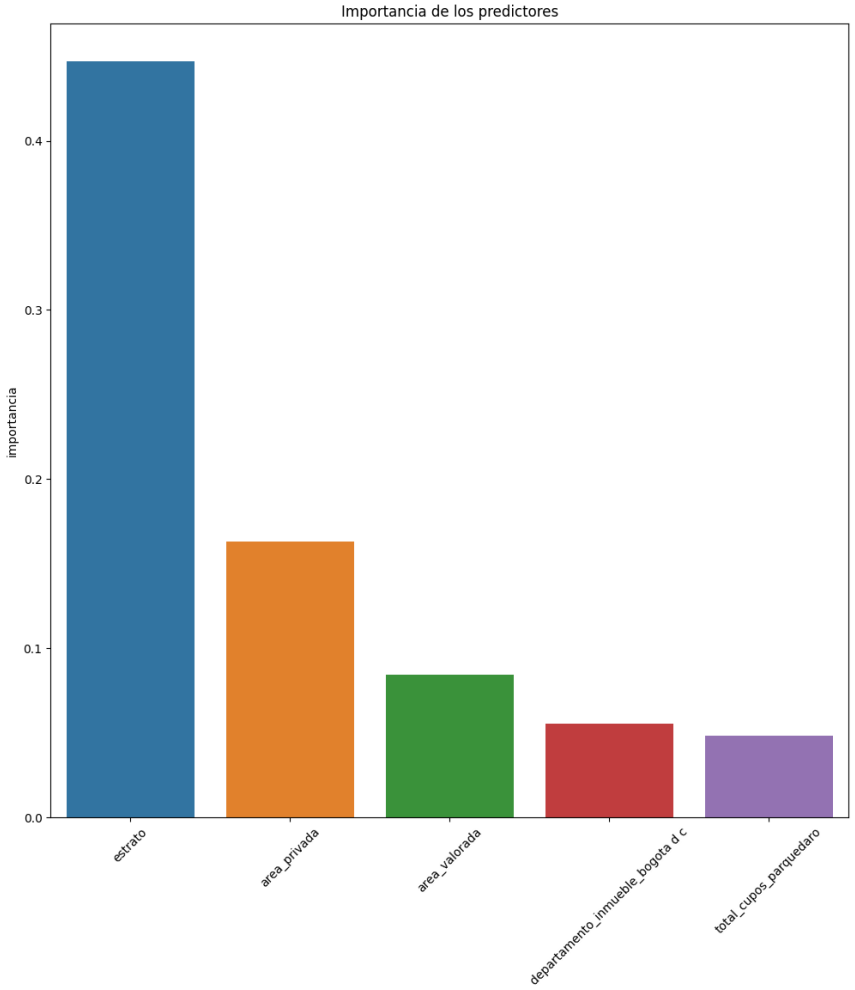
5. Conclusiones

- 5.1. Técnicas:

A lo largo del estudio se detectó que, en efecto, el modelo Random Forest tuvo el mejor comportamiento para predecir los valores de avalúos total dada la base de datos proporcionada. Esto tiene sentido ya que el modelo Decision Tree puede verse afectado por sesgos en las variables que no se pudieron visualizar. Por otro lado, el RLM al ser un modelo puramente lineal pudo omitir relaciones no lineales entre sus variables, generando una predicción sesgada.

- 5.2. Sobre Resultados:

Con respecto a los resultados del modelo seleccionado, se observa que la variable más importante para describir el valor de avaluó de un inmueble es su estrato. Si bien es natural pensar que el área privada es más importante pues a mayor área mayor es la suma de valor por mt2, lo cierto es que este último no vale lo mismo en todas las ubicaciones, y esto se ve corroborado con la variable dummy que indica si un inmueble está en Bogotá y que es una de las más importantes. Finalmente, como hallazgo adicional, es curioso que ninguna de las características del inmueble en sí tengan una importancia significativa para el valor del mismo, pero esto puede llegar a ser explicado asumiendo que estas características se ven reflejadas en el área privada.



Gráfica 4. Importancia de los predictores en el modelo Random Forest.

6. Anexos

- 6.1. Dificultades Encontradas (Extra):
  - Bases de Datos con múltiples formatos: Generó problemas a la hora de importar el archivo pues los parámetros de *read_csv* no permitía incluir algunos arreglos como, por ejemplo, poner como comilla de texto, las comillas """.
  - Sin definición de la naturaleza de los datos: Para lo cual se diseñaron algoritmos que permitían definir el tipo de variables.
  - Número de variables: Naturalmente, entre mayor sea la base de datos, más complicado es analizarla en su totalidad, para ello se aplicaron modelos de Machine Learning para predicción y, para el uso de modelos probabilísticos, se usó la reducción de dimensionalidad a través del análisis de correlaciones.

- 6.2. Cómo poner a disposición los Resultados / el Modelo:

  El modelo se puede poner a disposición del cliente mediante una API cuyos parámetros de entrada sean:
  - Tipología del inmueble.
  - Ubicación (barrio, departamento, etc).
  - Estrato.
  - Área privada (se puede esperar que sea un aproximado), etc.
  - Parqueaderos.

  Se puede diseñar una función que, con estos parámetros, encuentre inmuebles similares (o comparables) en la base de datos que se posee y retorne una media de los precios predichos por el modelo. De este modo, con un input relativamente pequeño se puede obtener un resultado del modelo fácilmente.