A logo with a gear and text

Description automatically generated **UNIVERSIDAD DEL VALLE**

**DEPARTAMENTO DE INFORMÁTICA**

**CARRERA EN INGENIERÍA INFORMÁTICA**

**PREDICCIÓN DE PRECIOS DE VIVIENDAS UTILIZANDO MODELOS DE APRENDIZAJE AUTOMÁTICO**

**INTEGRANTES:**

* David Calvi Arce 33.333%
* Jaider Alejandro Pinto Ribera 33.333%
* Mario Antonio Lazarte Vega 33.333%

Cochabamba – Bolivia

2023

**ÍNDICE DE CONTENIDO**

[**CAPÍTULO 1. GENERALIDADES** 3](#_Toc152629002)

[**1.1** **ANÁLISIS DE LA PROBLEMÁTICA** 3](#_Toc152629003)

[**1.2** **OBJETIVOS** 5](#_Toc152629004)

[**1.2.1** **OBJETIVO GENERAL** 5](#_Toc152629005)

[**1.2.2 OBJETIVOS ESPECÍFICOS** 5](#_Toc152629006)

[**1.3** **JUSTIFICACIÓN** 6](#_Toc152629007)

[**1.3.1 JUSTIFICACIÓN TÉCNICA** 6](#_Toc152629008)

[**1.3.2 JUSTIFICACIÓN ECONÓMICA** 6](#_Toc152629009)

[**1.3.3 JUSTIFICACIÓN SOCIAL** 6](#_Toc152629010)

[**CAPÍTULO 2. MARCO TEÓRICO.** 6](#_Toc152629011)

[**2.1 MODELOS PREDICTIVOS** 6](#_Toc152629012)

[**2.2 MODELOS PREDICTIVOS COMÚNMENTE UTILIZADOS** 6](#_Toc152629013)

[**2.2.1** **MODELO DE REGRESIÓN LÍNEAL MÚLTIPLE** 6](#_Toc152629014)

[**2.2.2 ALGORITMO DE KNN** 7](#_Toc152629015)

[**2.2.3 ARBOLES DE DECISION** 7](#_Toc152629016)

[**2.2.4 GRADIENT BOOST** 8](#_Toc152629017)

[**2.3 EL MERCADO INMOBILIARIO** 9](#_Toc152629018)

[**2.4 ESTADO ECONÓMICO DEL MERCADO INMOBILIARIO EN SAO PAULO** 9](#_Toc152629019)

[**2.5 PREDICCIÓN DE PRECIOS DE VIVIENDAS BASADAS EN CARACTERÍSTICAS ESPECÍFICAS** 11](#_Toc152629020)

[**2.6 USO Y RESULTADOS DEL USO MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE PRECIOS DE VIVIENDAS** 11](#_Toc152629021)

[**2.6.1 USO DE MODELO DE REGRESION LINEAL MÚLTIPLE** 11](#_Toc152629022)

[**2.6.2 USO DE ALGORITMO KNN** 11](#_Toc152629023)

[**2.6.3 USO DE ARBOLES DE DECISION** 11](#_Toc152629024)

[**2.6.4 USO DE GRADIENT BOOST** 12](#_Toc152629025)

[**2.7 TÉCNICAS DE PREPROCESAMIENTO Y EVALUACION** 12](#_Toc152629026)

[**2.7.1 TÉCNICAS DE PREPROCESAMIENTO** 12](#_Toc152629027)

[**CAPÍTULO 3. DESARROLLO DEL PROYECTO** 12](#_Toc152629028)

[**3.1 DESCRIPCIÓN DEL DATASET** 12](#_Toc152629029)

[**3.2 PREPROCESAMIENTO DEL DATASET** 13](#_Toc152629030)

[**3.3 ENTRENAMIENTO DE LOS MODELOS** 14](#_Toc152629031)

[**3.3.1 ENTRENAMIENTO DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE** 15](#_Toc152629032)

[**3.3.2 ENTRENAMIENTO DEL MODELO DE RANDOM FOREST** 15](#_Toc152629033)

[**3.3.3 ENTRENAMIENTO DEL MODELO DE KNN** 16](#_Toc152629034)

[**3.3.4 ENTRENAMIENTO DEL MODELO DE GRADIENT BOOST** 16](#_Toc152629035)

[**3.4 OPTIMIZACIÓN Y VALIDACIÓN CRUZADA DE LOS MODELOS** 17](#_Toc152629036)

[**3.5 EVALUACIÓN DE LOS MODELOS** 19](#_Toc152629037)

[**3.5.1 EVALUACIÓN DEL MODELO DE REGRESIÓN LÍNEAL MÚLTIPLE** 20](#_Toc152629038)

[**3.5.2 EVALUACIÓN DEL MODELO DE RANDOM FOREST** 20](#_Toc152629039)

[**3.5.3 EVALUACIÓN DEL MODELO DE KNN** 21](#_Toc152629040)

[**3.5.4 EVALUACIÓN DEL MODELO DE GRADIENT BOOST** 22](#_Toc152629041)

[**CAPÍTULO 4. CONCLUSIONES Y RECOMENDACIONES** 22](#_Toc152629042)

[**4.1 CONCLUSIONES** 22](#_Toc152629043)

[**4.1.1 CONCLUSIÓN GENERAL** 22](#_Toc152629044)

[**4.1.2** **CONCLUSIÓN DE OBJETIVOS ESPECÍFICOS** 22](#_Toc152629045)

[**4.2 RECOMENDACIONES** 23](#_Toc152629046)

[**BIBLIOGRAFÍA** 23](#_Toc152629047)

# **CAPÍTULO 1. GENERALIDADES**

## **ANÁLISIS DE LA PROBLEMÁTICA**

Desde el inicio de la era tecnología y el *boom* inmobiliario en el mundo, se ha visto un gran crecimiento y movimiento del sector inmobiliario, surgiendo la necesidad de poder crear una herramienta tecnológica la cual pueda facilitarnos el arduo trabajo que es calcular el precio de un inmueble, ya sea dándonos predicciones de precios referentes dependiendo distintos factores los cuales puedan ser modificados dependiendo distintos factores.

Debido a que los precios de los inmuebles van variando alrededor de los años, con más fuerza aun en Latinoamérica ya que la gran mayoría de los países que se encuentran están en constante desarrollo y crecimiento, un estudio elaborado por la Universidad Torcuato Di Tella y Zonaprop analizó el sector inmobiliario de 9 países de Latinoamérica estimando el precio promedio del metro cuadrado. En la capital uruguaya se estima un precio de 3.166 dólares estadounidenses, según datos de septiembre de 2023. Esto la posiciona como la ciudad más cara para tener un techo propio entre algunas de las mayores urbes de la región. En Segundo Lugar, se encuentra la Ciudad de México, donde los interesados deberían pagar, en promedio, 2.948 dólares por metro cuadrado. Para las también mexicanas Monterrey y Guadalajara se estima un valor de aproximadamente 2.600 y 2.500 dólares, respectivamente, lo que las convierte en la tercera y cuarta ciudades más caras, de las incluidas en el estudio, para convertirse en dueño de un apartamento. (RIAL Di Tella; Zonaprop, 2023)

Ilustración 1, ¿Cuánto cuesta el metro cuadrado en América Latina?



Fuente: (RIAL Di Tella; Zonaprop, 2023)

El informe revela que el precio del metro cuadrado en las ciudades analizadas subió un 5,2% entre marzo y septiembre de este año. Si se tienen en cuenta sólo las ciudades capitales, el valor del metro cuadrado experimentó un aumento del 4,7%. Con el objetivo de conseguir una muestra homogénea y comparable, el informe tiene en cuenta los precios pedidos por los vendedores en anuncios clasificados de apartamentos de uno o dos dormitorios con una superficie de entre 20 y 100 metros cuadrados y cuyo precio total se sitúa entre los 10.000 y 300.000 dólares estadounidenses. (RIAL Di Tella; Zonaprop, 2023)

Con esta problemática que surge podemos formularnos la siguiente pregunta:

***¿Puede un modelo predictivo ser usado para predecir de manera precisa y confiable los precios de las viviendas?***

Ilustración 2, Diagrama de Ishikawa

Diagrama

Descripción generada automáticamente

Fuente: Elaboración Propia

## **OBJETIVOS**

### **OBJETIVO GENERAL**

Desarrollar un modelo predictivo el cual nos proporcione precios de viviendas en el Estado de São Paulo, Brasil tomando en cuenta diferentes características de las viviendas.

### **1.2.2 OBJETIVOS ESPECÍFICOS**

* Recopilar y preparar un dataset con información relevante sobre viviendas en el Estado de São Paulo, que incluya características importantes como la ubicación, el tamaño y número de habitaciones; dataset que se nos será proporcionado por medio del sitio web de Kaggle.
* Realizar un análisis exploratorio para encontrar posibles correlaciones que nos ayudarán con la limpieza del dataset.
* Entrenar una serie de modelos que incluye regresiones lineales múltiple, KNN, Random Forest y Gradient Boost para permitirnos determinar la variabilidad de los precios de las viviendas. Además, se buscará ajustar los modelos para abordar posibles problemas como predicciones incorrectas.
* Evaluar la precisión de los modelos mediante el uso de diferentes métricas, incluyendo el Error Cuadrático Medio (MSE), el Mean Absolute Error y el Root Mean Square Error, con el fin de establecer un rango aceptable de predicción.
* Aplicar la validación cruzada K-Folds en los modelos para asegurar un ajuste adecuado a los datos de entrenamiento y un rendimiento óptimo con nuevos datos, aumentando la confiabilidad de las predicciones.

## **JUSTIFICACIÓN**

### **1.3.1 JUSTIFICACIÓN TÉCNICA**

Existe una disponibilidad en aumento constante de tecnologías, herramientas y métodos avanzados en el campo del aprendizaje automático que permiten abordar la problemática de la predicción de precios de viviendas a nivel global. Estas tecnologías ofrecen la capacidad de analizar grandes conjuntos de datos y generar modelos precisos.

### **1.3.2 JUSTIFICACIÓN ECONÓMICA**

La implementación de un modelo predictivo puede ser más económica a largo plazo que depender exclusivamente de métodos tradicionales de valoración de viviendas, como las tasaciones manuales. Además, una mayor precisión en la estimación de precios puede llevar a decisiones financieras más acertadas, lo que puede resultar en ahorros significativos para los compradores y vendedores de viviendas.

### **1.3.3 JUSTIFICACIÓN SOCIAL**

Un modelo predictivo preciso puede contribuir a la estabilidad y transparencia del mercado inmobiliario, lo que a su vez puede mejorar el acceso a la vivienda y reducir la especulación descontrolada. Esto beneficia a la sociedad al hacer que la compra y venta de viviendas sea más justa y accesible.

# **CAPÍTULO 2. MARCO TEÓRICO.**

## **2.1 MODELOS PREDICTIVOS**

El modelado predictivo es una técnica estadística utilizada para predecir el resultado de eventos futuros basados en datos históricos. Implica construir un modelo matemático que tome variables de entrada relevantes y genere una variable de salida predicha. Los algoritmos de aprendizaje automático se utilizan para entrenar y mejorar estos modelos para ayudarlo a tomar mejores decisiones. El modelado predictivo se utiliza en muchas industrias y aplicaciones y puede resolver una amplia gama de problemas, como la detección de fraudes, la segmentación de clientes, el diagnóstico de enfermedades y la predicción del precio de las acciones. (Qlik, 2022)

Los dos métodos de modelado predictivo más comúnmente empleados son la regresión y las redes neuronales. La precisión de análisis y cada modelo predictivos depende de varios factores, incluida la calidad de sus datos, su elección de variables y las suposiciones de su modelo. (Qlik, 2022)

## **2.2 MODELOS PREDICTIVOS COMÚNMENTE UTILIZADOS**

### **MODELO DE REGRESIÓN LÍNEAL MÚLTIPLE**

(Granados, 2016) La regresión lineal múltiple es un algoritmo de aprendizaje automático(supervisado) ampliamente utilizado que implica ajustar modelos lineales entre una variable dependiente y más de una variable independiente. Su fórmula es:

Donde:

### **2.2.2 ALGORITMO DE KNN**

(Larrañaga, Inza, & Moujahid, 2017) El algoritmo KNN es un algoritmo de agrupamiento y predicción, según el caso se puede aplicar para predecir datos o eventos y para clasificar elementos de un conjunto en grupos o clases frecuentes a la que pertenecen sus K vecinos más cercanos. (Seco, 2004)

Puede extenderse utilizando no uno, sino un conjunto de datos más cercanos para predecir el valor de los nuevos datos conocidos como k-vecinos más cercanos. Al considerar más de un vecino se brinda inmunidad ante el ruido y se suaviza la curva de estimación. (Puentes, 2018)

(Seco, 2004) Su funcionamiento para la predicción o clasificación está basado en los siguientes pasos:

* Se recibe como entrada el elemento Xi a ser clasificado.
* Se elige el valor de k para determinar cuántos vecinos más cercanos se van a tener en cuenta para comparar con el nuevo elemento.
* Normalizar los datos si es necesario.
* Realizar la siguiente iteración:

Para cada Ki:

* Hallar la distancia entre el nuevo elemento y el Ki vecino más cercano.
* Clasificar o agrupar el elemento nuevo al grupo donde se encuentra los k vecinos más cercanos.
* Termina cuando haya comparado el nuevo elemento con los k vecinos y el nuevo elemento se haya clasificado.
* Fin del Algoritmo.

(Seco, 2004) El algoritmo tiene como ventaja lo sencillo de aprender e implementarlo, pero contra que utiliza todo el dataset para entrenar “cada punto” y por eso requiere de uso de mucha memoria y recursos de procesamiento (CPU). Por estas razones KNN tiende a funcionar mejor en datasets pequeños y sin una cantidad enorme de features (las columnas).

### **2.2.3 ARBOLES DE DECISION**

(Rokach, 2014) Es un modelo predictivo el cual se define por ser un modelo jerárquico de decisiones y sus consecuencias, dependiendo de su uso puede ser denominado árbol de clasificación cuando se utiliza para tareas de clasificación y árbol de regresión para tareas de regresión. Estos árboles son aplicados en distintas áreas como medicina, ingeniería y negocios.

Un árbol de decisiones puede ser llamado un clasificador el cual consta de varios nodos formando una raíz la cual no posee bordes de entrada y el resto de los nodos poseen un borde entrante, esto quiere decir que a los nodos con bordes salientes se los denominan internos y al resto se los denomina hojas o terminales. (Rokach, 2014)

(Amaya, 2010) Esta técnica nos permite analizar decisiones basadas en resultados y probabilidades asociados, es empleada en situaciones las cuales se necesita de la toma de decisiones permitiéndonos tomar la decisión más acertada, son muy similares a los árboles de decisiones, los árboles de clasificación son excelentes predictores debido a que también pueden dar resultados más específicos o filtrados llamados reglas de clasificación como “condición” → “resultado”.

### **2.2.4 GRADIENT BOOST**

(Friedman, 2001) Gradient Boost es una técnica de Machine Learning la cual está conformada por un conjunto de árboles de decisión individuales, realizando una estimación de funciones desde la perspectiva de la optimización numérica, mostrando la conexión entre los Árboles de Decisión y la minimización de una función de costo ya sea Regresión o Clasificación.

En este modelo se presentan algoritmos específicos para mínimos cuadrados, desviación mínima absoluta y funciones de pérdida de Huber-M para regresión y probabilidad logística multiclase para clasificación.

También este modelo proporciona mejoras y ventajas debido a que los componentes principales son los árboles de regresión y estos nos presentan herramientas tales como los modelos “TreeBoost”. (Friedman, 2001)

Ilustración 3, Fórmula del Gradient Boost

A math equations and formulas

Description automatically generated with medium confidence

Fuente: (Friedman, 2001)

(Puentes, 2018) La Figura, muestra el primer algoritmo genérico de boosting basado en árboles

desarrollado por Friedman donde:

1. Inicializa F(x), en base al argmin de la función de costo L en un punto fijo.

2. Inicia el bloque FOR.

3. Computa la gradiente negativa.

4. Ajusta el modelo con el fin de encontrar los parámetros a que pertenecen al árbol.

5. Elije un tamaño de paso.

6. Actualiza la estimación de F(x).

7. Fin del bloque FOR

## **2.3 EL MERCADO INMOBILIARIO**

El Mercado Inmobiliario es el entorno en el que se realizan transacciones de compra, venta y alquiler de propiedades, que incluyen tierras y edificaciones. Este mercado se divide típicamente en tres categorías: Residencial, Comercial e Industrial, basándose en el uso dado a la tierra. (Alves, 2021)

La parte Residencial del Mercado Inmobiliario implica el entorno empresarial donde compradores y vendedores, a menudo asistidos por agentes o corredores, realizan transacciones de propiedades destinadas a vivienda. (Alves, 2021)

Existen varios factores que influyen en el mercado inmobiliario:

**OFERTA Y DEMANDA**

La oferta de viviendas en venta en relación con la demanda de compra determina si el mercado favorece a los compradores o a los vendedores. (Science Direct, 2018)

**PRECIOS MEDIANOS DE LAS VIVIENDAS**

Los precios de las viviendas pueden afectar la accesibilidad y la demanda en el mercado. (Science Direct, 2018)

**CONSTRUCCIÓN NUEVA**

La construcción de nuevas viviendas o edificios comerciales puede aumentar la oferta y afectar los precios. (Science Direct, 2018)

**TASAS DE HÍPOTECA**

Las tasas de interés de las hipotecas pueden influir en la capacidad de los compradores para adquirir propiedades. (Science Direct, 2018)

**ECONOMÍA**

La salud económica general puede afectar la confianza y la capacidad de los compradores y vendedores en el mercado. (Science Direct, 2018)

## **2.4 ESTADO ECONÓMICO DEL MERCADO INMOBILIARIO EN SAO PAULO**

Los bienes raíces brasileños han tenido un buen año, con una fuerte demanda en todos los principales mercados del país en 2021. Sin embargo, a una industria le fue extraordinariamente bien en 2021, mucho más allá de las proyecciones. Tanto en Sao Paulo como en Río de Janeiro, las ventas y lanzamientos de propiedades de lujo aumentaron en dos dígitos. (Mordor Intelligence, 2023)

Los dos principales mercados de lujo del país, São Paulo y Río de Janeiro, han experimentado un gran aumento en la demanda de propiedades por valor de más de R$ 2 millones (USD 392.102,66). (Mordor Intelligence, 2023)

Los apartamentos de lujo han sido la preferencia entre los compradores adinerados en las dos ciudades más grandes de Brasil. En São Paulo, las ventas de apartamentos con precios superiores a R$ 2 millones (USD 392.102,66) aumentaron un 3,2% en el año hasta junio de 2021. En Río de Janeiro, el aumento fue aún mayor, con un aumento de las transacciones del 10%. (Mordor Intelligence, 2023)

En el escalón de precios más bajo (más de R$ 1,5 millones (USD 291828,75)), las cifras fueron aún más impresionantes. La Asociación Brasileña de Desarrolladores (ABRAINC) informó que las ventas de este tipo de propiedades aumentaron un 43% en Sao Paulo entre enero y septiembre de 2021. (Mordor Intelligence, 2023)

El segmento de lujo y súper lujo en Sao Paulo, de enero a septiembre de 2021, totalizó 4.217 unidades, lo que representó un aumento del 129%. Junto con las sólidas ventas, fue el gran salto en el número de lanzamientos de propiedades de lujo en Brasil. Solo en Sao Paulo, aumentaron un 232% en los primeros nueve meses de 2021. En Río de Janeiro, el aumento en la región fue del 35%. (Mordor Intelligence, 2023)

Gráfico, Gráfico de barras

Descripción generada automáticamenteIlustración 4, Número de nuevas viviendas en venta, del primer trimestre del 2020 al primer trimestre del 2021

FUENTE: (Mordor Intelligence, 2023)

## **2.5 PREDICCIÓN DE PRECIOS DE VIVIENDAS BASADAS EN CARACTERÍSTICAS ESPECÍFICAS**

La predicción y cálculo de precios basadas en características de viviendas o inmuebles depende e influye de diferentes factores:

**CARACTERÍSTICAS FÍSICAS Y ESTRUCTURALES**

* TAMAÑO

La superficie total en metros o pies cuadrados.

* NÚMERO DE HABITACIONES Y BAÑOS

Dependiendo de la cantidad de habitaciones y baños incrementa el precio.

* CONDICIÓN Y AÑO DE CONTRUCCIÓN

Las viviendas para estrenar o construidas más recientemente tienen precios más altos.

**UBICACIÓN Y ENTORNO**

* LOCALIZACIÓN GEOGRÁFICA

La cercanía a centros urbanos o lugares específicos como escuelas, trabajos y atracciones incrementan el valor de una propiedad.

Otros factores como disponibilidad de servicios básicos, tasa de criminalidad, disponibilidad de transporte, entre otros pueden influir en el valor de una propiedad. (Mir, 2020)

## **2.6 USO Y RESULTADOS DEL USO MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE PRECIOS DE VIVIENDAS**

### **2.6.1 USO DE MODELO DE REGRESION LINEAL MÚLTIPLE**

Este modelo se utiliza en el ámbito inmobiliario para predecir el mejor momento para vender una vivienda.

Considera factores que pueden afectar el precio de venta, como ser la antigüedad de la vivienda, el precio de otras viviendas en el vecindario, el desempeño del sistema de escuelas públicas y la cantidad de parques cercanos. Es a partir de estas variables independientes que se construye un modelo predictivo para predecir el precio estimado de venta de una vivienda. (Historia de la Empresa, 2013)

### **2.6.2 USO DE ALGORITMO KNN**

En el contexto de la predicción de precios de viviendas, el KNN se utiliza para entrenar con datos que incluyen el tamaño de la casa, el año de construcción y el precio. Luego, se usa para predecir el precio de otra casa. Por ejemplo, si se elige k = 5, el algoritmo encuentra los cinco vecinos más cercanos del nuevo punto de datos (la casa que nos interesa) y calcula un valor medio de estos para predecir el precio.​ (Dobilas, 2020)

### **2.6.3 USO DE ARBOLES DE DECISION**

Los árboles de decisión son parte de un método de aprendizaje automático supervisado. Crean muchos árboles de decisión, conocidos como bosque, que se utilizan en la predicción. Cada árbol genera su propia predicción y se utiliza como parte de un esquema de votación para hacer predicciones finales. Este enfoque ayuda a evitar el exceso de ajuste y es útil para predecir el valor de la vivienda basándose en variables como el número de habitaciones, la distancia hasta las escuelas, la proximidad a carreteras importantes, el nivel medio de ingresos y los recuentos de delitos​​. (ESRI, 2021)

### **2.6.4 USO DE GRADIENT BOOST**

Es una técnica de aprendizaje automático que mejora la precisión de los modelos predictivos mediante la combinación de varios modelos simples para crear uno más complejo y preciso. En la predicción de precios de viviendas, el Gradient Boosting utiliza árboles de decisión secuenciales, donde cada árbol se enfoca en corregir los errores del modelo anterior. Así, se van creando árboles que se combinan para formar un modelo predictivo más preciso​​. (Remolinator, 2020)

## **2.7 TÉCNICAS DE PREPROCESAMIENTO Y EVALUACION**

### **2.7.1 TÉCNICAS DE PREPROCESAMIENTO**

El preprocesamiento de los datos es una parte crucial a la hora de entrenar nuestros modelos. Ningún dataset es perfecto, algunos de ellos pueden contener desde inconsistencia en los datos, hasta errores. Estos problemas pueden afectar negativamente en el rendimiento de nuestros modelos si no son tratados adecuadamente, es por ello por lo que el preprocesamiento es una técnica crucial para mejorar la calidad de nuestros datos de entrada. (Apure Digital , 2023)

Algunos de los pasos que se suele seguir para garantizar que los datos sean los más limpios posibles incluyen:

* **DEFINICIÓN DEL PROBLEMA**

Es esencial definir claramente el problema que buscamos resolver. Por ejemplo, el tema que estamos tratando en esta investigación, “crear un modelo de aprendizaje automático para predecir el precio de las viviendas”.

* **RECOLECCIÓN DE LOS DATOS**

Los datos pueden provenir de diversas fuentes, como bases de datos públicas o API’s de plataformas inmobiliarias.

* **LIMPIEZA Y EXPLORACIÓN DE LOS DATOS**

Este paso implica verificar y tomar decisiones sobre valores nulos y errores en las entradas. También incluye visualización de datos para comprender de mejor manera las relaciones y las distribuciones.

# **CAPÍTULO 3. DESARROLLO DEL PROYECTO**

## **3.1 DESCRIPCIÓN DEL DATASET**

El dataset nos lo proporciona el sitio web de Kaggle y consta de 10,033 filas y 5 características identificadas como 'Street', 'City', 'Area', 'Rooms', 'Bathrooms', 'Parking Space' y 'Price'. Estas características proporcionan información detallada sobre las propiedades inmobiliarias en la región de São Paulo, incluyendo la ubicación, tamaño del área, número de habitaciones, baños y espacios de estacionamiento, así como el precio de venta.

Ilustración 5, Descripción del dataset

Pantalla de computadora

Descripción generada automáticamente con confianza media

Fuente: Elaboración Propia

## **3.2 PREPROCESAMIENTO DEL DATASET**

El preprocesamiento se inicia con la evaluación de la composición y tipos de datos del dataset utilizando la librería de ***pandas*** para ejecutar ***df.dtypes***.

Ilustración 6, Tipos de datos

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

Se identificaron columnas con valores numéricos representados de manera no estándar, incluyendo intervalos. Para resolver esto, se promediaron los valores de los intervalos y se redondearon a un entero más cercano.

Ilustración 7, Restructuración de los datos

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

Posteriormente, se formatearon las columnas 'Street' y 'City' para abordar problemas que podrían afectar el entrenamiento de modelos. En 'Street', se eliminaron los números y en 'City', se eliminó información redundante como la repetición de 'São Paulo'. Además, se verificó la existencia de valores nulos y se aplicaron métodos de imputación adecuados para tratarlos.

Ilustración 8, Formateo de Columnas

Captura de pantalla de un celular

Descripción generada automáticamente

Fuente: Elaboración Propia

Se aplicó ***ColumnTransformer*** para estandarizar las variables numéricas con ***StandardScaler y*** transformar las categóricas con ***OneHotEncoder***, facilitando así su uso en el modelado.

## **3.3 ENTRENAMIENTO DE LOS MODELOS**

Esta sección se centra en el entrenamiento específico para cada uno de los modelos de aprendizaje automático seleccionados. Se implementó y ajustó los modelos utilizando la biblioteca Sklearn para optimizar su rendimiento en la predicción de precios.

### **3.3.1 ENTRENAMIENTO DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE**

**IMPLEMENTACIÓN:**

Se procedió a utilizar la función de regresión lineal en sklearn para implementar este modelo. A menudo, la regresión lineal ayuda a establecer una relación lineal entre las variables independientes y la variable dependiente (precio).

Ilustración 9, Entrenamiento del modelo de Regresión Lineal Múltiple

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

### **3.3.2 ENTRENAMIENTO DEL MODELO DE RANDOM FOREST**

**IMPLEMENTACIÓN:**

Se utilizó la función de “**RandomForestRegressor**” para este modelo. Como mencionamos anteriormente, Random Forest es un método conjunto que utiliza múltiples árboles de decisión para mejorar la precisión y controlar el sobreajuste.

**ENTRENAMIENTO Y OPTIMIZACIÓN:**

Los parámetros constan de “**n\_estimators**” y “**random\_state**”, que se pueden ajustar para el optimizador del modelo. Se realizó capacitación para capturar relaciones lineales y no lineales en los datos.

Ilustración 10, Entrenamiento del modelo de Random Forest

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

### **3.3.3 ENTRENAMIENTO DEL MODELO DE KNN**

**IMPLEMENTACIÓN**:

Utilizando la función de “**KNeighborsRegressor**”, implementamos un modelo KNN que predice el valor de una nueva observación en función de la similitud con sus "**K**" vecinos más cercanos.

**ENTRENAMIENTO**:

Se experimentó con diferentes valores de "**K**" para encontrar un equilibrio entre el sesgo y la varianza para asegurarse de que el modelo se generalice bien a datos inéditos.

Ilustración 11, Entrenamiento del modelo de KNN

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

### **3.3.4 ENTRENAMIENTO DEL MODELO DE GRADIENT BOOST**

**IMPLEMENTACIÓN**:   
Se procedió a implementar la función “GradientBoostingRegressor”. Como ya es de conocimiento, este modelo construye árboles de decisión de forma secuencial, cada uno corrigiendo los errores del anterior, lo que le permite lidiar eficientemente con la complejidad de los datos.

**ENTRENAMIENTO Y AJUSTE:**

El modelo se entrena y ajusta para mejorar su capacidad de realizar predicciones precisas y procesar características de datos complejas.

Ilustración 12, Entrenamiento del modelo de Gradient Boost

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

## **3.4 OPTIMIZACIÓN Y VALIDACIÓN CRUZADA DE LOS MODELOS**

La optimización y la validación cruzada es de vital importancia para el éxito de los modelos predictivos. Se utilizó la técnica de K-Fold para mejorar la solidez y eficiencia de los modelos en comparación con datos nuevos. Estos métodos son necesarios para adaptar los parámetros del modelo y garantizar su precisión y generalización.

**3.4.1 OPTIMIZACIÓN Y VALIDACIÓN CRUZADA DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE**

Se realizó una validación cruzada de K-Fold para garantizar la solidez del modelo de regresión lineal múltiple. Esta validación lo que hace es dividir el registro de datos en cinco subcantidades, entrena el modelo para cuatro de ellas y valida el resto, donde este proceso se repite cinco veces para asegurar una evaluación completa.

Ilustración 13, Validación cruzada K-Folds en una Regresión Lineal Múltiple

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

**3.4.2 OPTIMIZACIÓN Y VALIDACIÓN CRUZADA DEL MODELO DE RANDOM FOREST**

Se realizó una validación cruzada de “K” veces con el modelo de Random Forest, que se configuró con cinco divisiones y aleatorización para garantizar la variabilidad de los resultados. Esta técnica permite una estimación confiable del desempeño de un modelo al evaluar varias subcantidades de datos.

Ilustración 14, Validación cruzada K-Folds en un modelo de Random Forest

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

**3.4.3 OPTIMIZACIÓN Y VALIDACIÓN CRUZADA DEL MODELO DE KNN**

En el proceso de optimización, se utilizó tecnología de validación cruzada K-Fold para evaluar la generalización del modelo KNN. El ajuste K se realizó con cinco divisiones, lo que significa que el modelo se puede entrenar y validar en diferentes partes del registro de datos. Los resultados de la medición de la determinación (R²) de cada pliegue y su promedio se calcularon para evaluar la capacidad predictiva del modelo.

Ilustración 15, Validación cruzada K-Folds en un algoritmo KNN

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

**3.4.4 OPTIMIZACIÓN Y VALIDACIÓN CRUZADA DEL MODELO DE GRADIENT BOOST**

El modelo de Gradient Boosting se validó de forma cruzada utilizando nuevamente el método K-fold para garantizar una evaluación confiable del resultado del modelo. Implementamos una validación cruzada de cinco etapas, que nos permitió evaluar el modelo en función de múltiples registros de datos y garantizar que el modelo no se viera perjudicado.

Ilustración 16, Validación cruzada en un modelo de Gradient Boosting

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

## **3.5 EVALUACIÓN DE LOS MODELOS**

Después de la optimización y validación cruzada, cada modelo se clasificó en detalle. Utilizamos métricas de rendimiento de la librería SkLearn para evaluar la capacidad de cada modelo para realizar predicciones precisas.

Cada evaluación se complementó con resultados que se lograron mediante validación cruzada K-Fold. Esto permite una comprensión más profunda del rendimiento del modelo en diferentes partes del registro de datos y garantiza que los modelos no solo coincidan con los datos de entrenamiento, sino que también mantengan un rendimiento óptimo para nuevos datos.

### **3.5.1 EVALUACIÓN DEL MODELO DE REGRESIÓN LÍNEAL MÚLTIPLE**

**EVALUACIÓN Y MEDICIÓN:**

Evaluamos el modelo usando el coeficiente de determinación R² y métricas de error como RMSE y MAE para medir la precisión. Estas métricas nos permitieron determinar cuán bien el modelo explica la variabilidad de los precios de las viviendas.

Ilustración 17, Evaluación del modelo de Regresión Lineal Múltiple

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

### **3.5.2 EVALUACIÓN DEL MODELO DE RANDOM FOREST**

**EVALUACIÓN Y MEDICIÓN:**

Medimos la precisión del modelo KNN a través del RMSE y MAE, junto con el R², para entender la precisión y el desempeño del modelo.

Ilustración 18, Evaluación del modelo de Random Forest

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

### **3.5.3 EVALUACIÓN DEL MODELO DE KNN**

**EVALUACIÓN Y MEDICIÓN:**

Medimos la precisión del modelo KNN a través del RMSE y MAE, junto con el R², para entender la precisión y el desempeño del modelo.

Ilustración 19, Evaluación del modelo de KNN

Texto

Descripción generada automáticamente

Fuente: Elaboración Propia

### **3.5.4 EVALUACIÓN DEL MODELO DE GRADIENT BOOST**

EVALUACIÓN EN PROFUNDIDAD:

Realizamos un análisis detallado del modelo de Gradient Boosting, ajustando los parámetros y utilizando las métricas de rendimiento para evaluar su precisión en la predicción de precios.

Ilustración 20, Evaluación del modelo de Gradient BoostTexto

Descripción generada automáticamente

Fuente: Elaboración Propia

# **CAPÍTULO 4. CONCLUSIONES Y RECOMENDACIONES**

## **4.1 CONCLUSIONES**

### **4.1.1 CONCLUSIÓN GENERAL**

El desarrollo de los modelos predictivos para estimar precios de viviendas en São Paulo se ha completado con éxito, incorporando una variedad de características relevantes de las viviendas. Este modelo demuestra ser una herramienta valiosa para comprender mejor el mercado inmobiliario en la región.

### **CONCLUSIÓN DE OBJETIVOS ESPECÍFICOS**

* **RECOPILACIÓN Y PREPARACIÓN DEL DATASET**

Se recopiló y preparó con éxito un dataset completo de Kaggle, que incluyó datos relevantes como la ubicación, tamaño y número de habitaciones de las viviendas en São Paulo. Este dataset fue crucial para el entrenamiento y validación de todos nuestros modelos predictivos.

* **ANÁLISIS EXPLORATORIO**

El análisis exploratorio reveló correlaciones significativas entre diversas características de las viviendas y sus precios, facilitando una limpieza eficaz del dataset y mejorando la calidad del modelo predictivo.

* **ENTRENAMIENTO DE LOS MODELOS**

Se entrenaron satisfactoriamente modelos de regresión lineal múltiple, KNN, Random Forest y Gradient Boost. Estos modelos demostraron ser capaces de captar la variabilidad en los precios de las viviendas, aunque el Random Forest sobresalió por su precisión y robustez.

* **EVALUACIÓN DE LA PRECISIÓN DE LOS MODELOS**

La evaluación de los modelos a través de métricas como MSE, MAE y RMSE confirmó la precisión de nuestras predicciones. El modelo de Random Forest mostró un coeficiente de determinación (R²) de 0.8, lo que indica su eficacia en explicar aproximadamente el 80% de la variabilidad en los precios de las viviendas.

* **IMPLEMENTACIÓN DE VALIDACIÓN CRUZADA**

La implementación de K-Folds ha sido esencial para validar y optimizar los modelos predictivos. Esta técnica permitió evaluar sus rendimientos y reducir el riesgo de sobreajuste, mejorando notablemente la precisión y fiabilidad de los modelos.

## **4.2 RECOMENDACIONES**

Para futuras investigaciones y aplicaciones prácticas, se recomienda la exploración de modelos mucho más complejos de aprendizaje profundo, además de exploración de hiperparámetros para sustentar la posibilidad de que algún otro modelo tenga un mejor desempeño que podría ayudar a descubrir patrones más sutiles en el mercado inmobiliario. Además, sería beneficioso integrar datos en tiempo real utilizando técnicas de web scraping a sitios de bienes raíces, así como también aumentar la cantidad de datos, lo que potencialmente podría tener un impacto positivo en la calidad de las predicciones. Finalmente, para maximizar el impacto de este estudio, sería interesante desarrollar un aplicativo que permita a diversos comprender y utilizar estos modelos en sus decisiones inmobiliarias.

# **BIBLIOGRAFÍA**

Alarcon, J. E. (15 de Noviembre de 2020). Obtenido de https://rstudio-pubs-static.s3.amazonaws.com/691346\_b9faa382cd1f4b66ae38a528c2c9dff7.html

Alves, P. (2021). *Real Estate Agent*. Obtenido de https://www.realestateagent.com/real-estate-glossary/real-estate/real-estate-market.html#:~:text=First%2C%20the%20Real%20Estate%20Industry,brokers%20%E2%80%93%20deal%20with

Amaya, J. A. (2010). *Toma de Decisiones Gerenciales.* ECOE Ediciones.

Apure Digital . (12 de Febrero de 2023). *Apure Digital*. Obtenido de https://apuredigital.net/que-es-el-preprocesamiento-de-datos-data-preprocessing/

Baptista, R. R. (2023). *Kaggle*. Obtenido de https://www.kaggle.com/datasets/ex0ticone/house-prices-of-sao-paulo-city

Dobilas, S. (2020). *Ichi*. Obtenido de https://ichi.pro/es/k-vecinos-mas-cercanos-knn-como-hacer-predicciones-de-calidad-con-aprendizaje-supervisado-235101999252418

ESRI. (2021). *ESRI*. Obtenido de https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/how-forest-works.htm

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Euclid*, 44.

Granados, R. M. (2016). *Universidad de Granada.* Obtenido de https://www.ugr.es/~montero/matematicas/regresion\_lineal.pdf

Historia de la Empresa. (2013). *Historia de la Empresa*. Obtenido de https://historiadelaempresa.com/regresion-multiple

Larrañaga, P., Inza, I., & Moujahid, A. (2017). *Reasearch Gate.* Obtenido de Redes Neuronales: https://www.researchgate.net/profile/Pedro-Larranaga/publication/268291232\_Tema\_8\_Redes\_Neuronales/links/55b7b5c408ae9289a08c0c68/Tema-8-Redes-Neuronales.pdf

Mir, M. (2020). *Towards Data Science*. Obtenido de https://towardsdatascience.com/house-prices-prediction-using-deep-learning-dea265cc3154#:~:text=The%20following%20features%20have%20been,levels%29%20in%20house

Mordor Intelligence. (2023). *Mordor Intelligence*. Obtenido de https://www.mordorintelligence.com/es/industry-reports/residential-real-estate-market-in-brazil

Puentes, J. F. (13 de Agosto de 2018). *LinkedIn*. Obtenido de https://es.linkedin.com/pulse/c%C3%B3mo-predecir-el-valor-de-un-inmueble-con-big-data-y-puentes-n%C3%BA%C3%B1ez

Qlik. (2022). *Qlik*. Obtenido de https://www.qlik.com/us/predictive-analytics/predictive-modeling

Remolinator. (2020). *Remolinator*. Obtenido de https://remolinator.com/que-es-el-gradient-boosting/

RIAL Di Tella; Zonaprop. (15 de Noviembre de 2023). *Statista*. (M. M. Roa, Editor) Recuperado el 2 de Diciembre de 2023, de Statista: https://es.statista.com/grafico/18134/precio-de-venta-por-metro-cuadrado-de-un-apartamento-en-america-latina/

Rokach, L. &. (2014). *DATA MINING WITH DECISION TREES.* Northeastern University.

Science Direct. (2018). *Science Direct*. Obtenido de https://www.sciencedirect.com/topics/economics-econometrics-and-finance/real-estate-market#:~:text=International%20Real%20Estate,of%20a%20parcel%20of%20land

Seco, F. M. (2004). Clasificadores eficaces basados en algoritmos rápidos de búsqueda del vecino más cercano. *Dialnet*.