# Predicción de Precios de Viviendas. Un Enfoque Práctico Utilizando Técnicas de Minería de Datos en el Conjunto de Datos  y Machine Learning

Marcela Abarca Delgado

Universidad Internacional San Isidro Labrador

Marcela Abarca Delgado

Departamento de ingeniería en Sistemas

Profesor

Dr. Samuel Saldaña Valenzuela.

# Contenido

# Resumen

En este proyecto se planea realizar el método CRISP-DM para preparar los datos del Data Set House Price Advanced. Donde se emplean las fases de conocimiento del negocio, conocimiento de los datos y preparación de los datos además, se observó y aplico las técnicas necesarias para la limpieza de los datos. Por otra parte se pudo comprobar que el data set corresponde a alguna empresa u organización dedicada a la venta de bienes raíces en el estado de Iowa. Se comprobó la existencia de datos nulos con variables categóricas y numéricas a las cuales se les aplico diferentes técnicas para realizar la limpieza de datos como análisis de subgrupos, imputación y eliminación logrando de esta manera la limpieza del data set. Se relevo la incidencia de algunas columnas en el precio de venta de una casa tales son el caso de LotArea, GrLivArea, BedroomAbvGr, TotRmsAbvGrd, LotFrontage, 1sFlrSF 2sFlrSF, FullBath y GarageArea.

**Palabras clave**: Método CRISP-DM, Bienes Raíces, Conocimiento del negocio, Conocimiento de los datos, Preparación de los datos, Técnicas de limpieza de datos, Modelado de datos

# Introducción

La minería de datos o Data Mining es un proceso donde se descubren información importante como patrones en grandes flujos de datos o Data Set. Según IBM “La minería de datos, también denominada descubrimiento de conocimiento en datos (KDD, por sus siglas en inglés), es el proceso de descubrir patrones y otra información valiosa en grandes conjuntos de datos.” (IBM, 1986)

El Data Mining en es una técnica importante en el correcto procesamiento para revelar información importante que será necesaria en la revelación de información necesaria para el desarrollo de la entidad u organización que consulte dicha información. Según (TOTVS SA, 2022) “La minería de datos tiene una enorme importancia estratégica para una empresa, ya que le permite comprender de una manera más contextualizada y precisa los comportamientos de los consumidores y los movimientos del mercado.

Es por esto que en este proyecto se da a la tarea de investigar y aplicar nuevas técnicas como técnicas vistas en clases necesarias para el desarrollo de este proyecto, además de aplicar el método CRISP-DM para un correcto desarrollo de la fase teórica de dicho proyecto y todas las fases necesarias para completarlo ya sean; comprensión del negocio, comprensión de los datos y preparación de los datos.

> Según (International Business Machines, 1986) CRISP-DM, que son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos.
>
> Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.

Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

Por otra parte está planeado investigar y profundizar en un previo conocimiento adquirido atreves de las clases recibidas en la especialización Data Science sobre los métodos PCA y KMeans ya que son técnicas utilizadas en el campo de Data Science.

Según (Conectando Ideas, 2023) El análisis de PCA es una técnica multivariante utilizada para reducir la dimensionalidad de un conjunto de datos. El objetivo principal es identificar las variables más importantes que explican la variabilidad observada en los datos, al mismo tiempo que se descartan las variables menos relevantes. En otras palabras, el análisis de PCA permite simplificar y resumir la información contenida en un conjunto de datos complejo.

Según (IEB school, 2008) El algoritmo k-means es un método de agrupamiento que divide un conjunto de datos en k grupos o clusters. Los datos se agrupan de tal manera que los puntos en el mismo clúster sean más similares entre sí que los puntos en otros clusters.

Por otra el modelado de datos consiste en la aplicación de adormidos orientados al machinelearning, con el objetivo de encontrar información relevante para encontrar una solución con el problema planteado, o simplemente encontrar información que fomente el crecimiento del negocio propietario o generadora de los datos con los que se trabaja en el dataset elegido en este caso con el dataset que se utiliza es el dataset House-Price-Advanced que se utilizó para llevar acabo las fases anteriores como lo eran reconocimiento del negocio, reconocimiento de los datos y preparación de los datos. Según, (Mikelnino, CRISP-DM: Fase de "Modelado" (Modeling), 2016)"Consiste en la ejecución del algoritmo de modelado seleccionado sobre el dataset preparado siguiendo el procedimiento diseñado".

Lo que la fase de modelado de datos representa una de las fases más importantes para cumplir con los objetivos que se plantea cumplir el negocio dueña de los datos con los que se trabaja por lo cual se considera que se debe abordar con extremo cuidado cuidando la integridad de los datos y realizar

conclusiones acertadas con los resultados que arrojan los adormidos que se emplean en esta fase.

Según, (ConectaPyme, 2023), el modelado de datos es una práctica fundamental para la gestión eficiente de la información en cualquier organización. Proporciona una base sólida para la estructuración y manipulación de los datos, minimizando la redundancia y la inconsistencia, y permitiendo una gestión óptima de la información en entornos cambiantes.

Por lo que en este proyecto se planea investigar y utilizar diferentes técnicas de adormidos de modelos no supervisados con el objetivo de encontrar información relevante con la que se darán recomendaciones constructivas al negocio dueño de los datos.

Como estudiante de la especialización de Data Science en de suma importancia realizar este proyecto e investigar conceptos desconocidos para los estudiantes y ponerlos en práctica no solo por la obtención de una calificación si no para alcanzar un mayor conocimiento además de poner en práctica lo aprendido en el desarrollo de este proyecto además de lo aprendido atreves de las clases recibidas.

## Objetivos Generales

- Aplicar un correcto desarrollo de Minería de Datos en el Data Set House Prices Advanced
- El objetivo de este proyecto es investigar y aplicar diversas técnicas de modelos no supervisados en el dataset House-Price-Advanced. A partir de los resultados obtenidos, se generarán recomendaciones constructivas para el negocio propietario de los datos, con el fin de mejorar la gestión eficiente de la información y optimizar la toma de decisiones en entornos cambiantes.

## Objetivo Especificos

- Desarrollar correctamente el método CRISP-DM.
- Implementar técnicas necesarias en la minería de datos del Data Set House Prices Advanced.
- Realizar una estadística descriptiva y sumatoria del Data Set House Prices Advanced.
- Detectar algún tipo de características relacionadas al precio de las casa del Data Set
- House_Prices_Advanced
- Realizar los cambios necesarios relacionados con la preparación de los datos del dataset House-Price-Advanced para de esta manera mejorar la calidad de los datos y garantizar que los datos estén listos para la fase de modelado de datos.
- Evaluar diferentes algoritmos de modelos no supervisados en el dataset mencionado. El propósito es descubrir patrones ocultos y relaciones entre variables para obtener información relevante para el negocio dueño de los datos

- Generar recomendaciones concretas para el negocio propietario de los datos, basadas en los resultados del modelado de los datos del dataset House-Price-Advanced. Estas recomendaciones se centrarán en la optimización de decisiones y la gestión eficiente de la información.
- Analizar y determinar las características más valoradas por los compradores de viviendas en Iowa, evaluando variables como ubicación, tamaño, número de dormitorios y baños, y características adicionales. El objetivo es comprender los factores que influyen en las decisiones de compra para que agentes y promotores adapten sus ofertas, y contribuyan a una mejor planificación urbanística.

Método CRISP-DM

## Fase 1: Compresión del negocio.

Como se menciona anteriormente, el método CRISP-DM es un método utilizado para llevar a cabo proyectos en el campo de la Minería de Datos o Data Mining. Este método consta de varias fases. La primera se conoce como Compresión del negocio. En dicha fase se enfoca en comprender o entender los objetivos y mandatos que tiene la entidad u organización para con los proyectos de la misma. Por consiguiente dichos datos se convierten en información valiosa que ayudara a definir el problema que busca resolver la minería de datos, además de una visión previa de cómo se alcanzaran los objetivos.

Según, (Pedrosa, 2011) , Esta fase inicial se enfoca en la comprensión de los objetivos y exigencias del proyecto desde una perspectiva de negocio. Posteriormente convierte ese conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Es por esto que en este proyecto se tomará la iniciativa al indagar para encontrar datos valiosos que posteriormente serán importantes para el entendimiento de información encontrada en el Data Set House Prices Advanced.

Como se puede visualizar en el data set House Prices Advanced y Sale Price, se puede encontrar información detallada que se relaciona con múltiples características con las que puede contar una casa como Precio de venta, Área del lote, Vecindario, Calidad General, Año de construcción, Total de pies cuadrados del sótano entre otras. Con esta información encontrada y con otras características como Tipo de venta, Condición de venta y año en que

se vendió se puede afirmar que esta organización a la cual estudiaremos se dedique a la venta de bienes raíces específicamente a la venta de casas ubicadas en la ciudad de Ames en el estado de Iowa.

Teniendo esta información recaudada en cuenta, es posible que con un correcto proceso de data mining pueda revelarse información importante, por ejemplo los factores que causan que una casa baje o suba de precio o si el año de construcción de la casa afectaría en el precio a la que es vendida, entre otras.

Con una previa estrategia en mente de cómo se llevará a cabo este proyecto, es importante abordar como se ejecutará y que técnicas serán las mejores para llevar el proyecto acabo. En primera instancia se prevé realizar una adecuada limpieza de datos en ambos data set

(House Prices Advanced y Sale Prices) para asegurar que la información recopilada posteriormente sea verídica y no contenga errores en este caso se podrán aplicar técnicas de eliminación de datos duplicados o columnas no relevantes para la revelación de información importante, manejo de datos faltantes para asegurar que los datos que se constan estén lo más posible cerca de la realidad entre otras técnicas necesarias para una limpieza adecuada del Data

Set que deberán ser estudiadas a profundidad.

Por consiguiente se abordarán los pasos de una obtención estadística ya sea descriptiva o suma del Data Set resultado de la unión de House Prices Advanced y Sale Prices con la cual se visualizará aspectos importantes como promedios entre otros.

# Fase 2: Comprensión de los datos.

El nombre de la segunda fase del método CRISP-DM se le da el nombre Comprensión de los datos. En esta fase, se da una visualización inicial de los datos con el objetivo de analizar y familiarizarse con los datos respectivos del Data Set, en este caso los datos respectivos al Data Set House Price y Sale Price. Dicho análisis se realiza con el interés de evaluar la calidad de los datos y formular las primeras hipótesis que se validaran o desestimaran.

Según (Niño, 2016), En esta fase el objetivo principal es poder hacer una captura inicial de los datos a analizar para familiarizarse con ellos, identificar problemas de calidad en los mismos, detectar subconjuntos de los datos que pudieran ser interesantes para formular hipótesis específicas que validar posteriormente con el análisis, e incluso identificar las primeras claves del conocimiento que se puede extraer de los datos.

Con el concepto presente de la segunda fase del método CRISP-DM "Comprensión de los datos" se da la tarea de realizar tareas que esta fase conlleva como lo son; captación de datos iniciales, descripción de los datos y exploración de los mismos. Esto con el objetivo de realizar el método CRISP-DM y de adquirir conocimientos fundamentales para el desarrollo en el proyecto.

En esta segunda fase de Comprensión de los datos se realizan diferentes tareas las cuales son: captación de datos iniciales, descripción de los datos, exploración de los datos y verificación de los datos.

Captura de datos iniciales: Para este proyecto elaborado con base en el método de CRISPDM se contarán con los Data Set House Price Advanced y Sale Price ambos en el formato “cvs” ambos Data Set proporcionados exclusivamente para la realización del proyecto y de prácticas por el Dr. Samuel Saldaña Valenzuela ubicados en el repositorio de la plataforma GitHub del Dr. Samuel Saldaña Valenzuela.

La captura de ambos Data Set se lleva a cabo por medio de la descarga directa desde el respectivo repositorio de GitHub a la cual pertenece al Dr. Samuel Saldaña Valenzuela al cual se ingresó a través de un hipervínculo proporcionado en el documento .pdf donde se asigna el primer proyecto llamado Proyecto No. 1_Data Mining.pdf. Una vez que se descargan ambos

Data Set (House Price Advanced y Sale Price) son importados a la herramienta de análisis Júpiter Lab que se utilizará para llevar a cabo este proyecto. Con este proceso se asegura la integridad de los datos ya que son descargados de manera segura además de ser descargados de un sitio confiable y se garantiza la disponibilidad de los datos ya que ambos Data Set (House Price y Sale Price) ya se encuentran en una herramienta empleada para el análisis y exploración de los Data Set. De esta manera se está más cerca de concluir de manera exitosa la segunda fase del método CRISP-DM.

Descripción de los datos: los datos obtenidos con el objetivo de ser examinados en este proyecto No.1 se presentan en los Data Set (House Price y Sale Price) ambos Data Set proporcionados por el Dr. Samuel Saldaña Valenzuela a través de su repositorio de GitHub

en formato .csv, proporcionados exclusivamente para el desarrollo de proyecto y prácticas de clase. Características encontradas en los Data Set House Price y Sale Price Exploración de los datos.

Formato: ambos conjuntos tanto House Price como Sale Price se encuentran en un formato.csv el cual es un formato flexible permitiendo el intercambio de datos entre programas diferentes evitando la posibilidad de que se dañen los Data Sets al ser descargados del repositorio de donde fueron proporcionados a los estudiantes del curso.

Según (Excel Dashboards, 2021), Formato CSV de Excel significa valores separados por comas, y es un formato de archivo comúnmente utilizado para intercambiar datos entre diferentes aplicaciones de software. Es un archivo de texto plano que contiene datos separados por comas, lo que facilita la importación y exportación de datos dentro y fuera de los programas de hoja de cálculo como Microsoft Excel.

Cantidad de registros y campos.

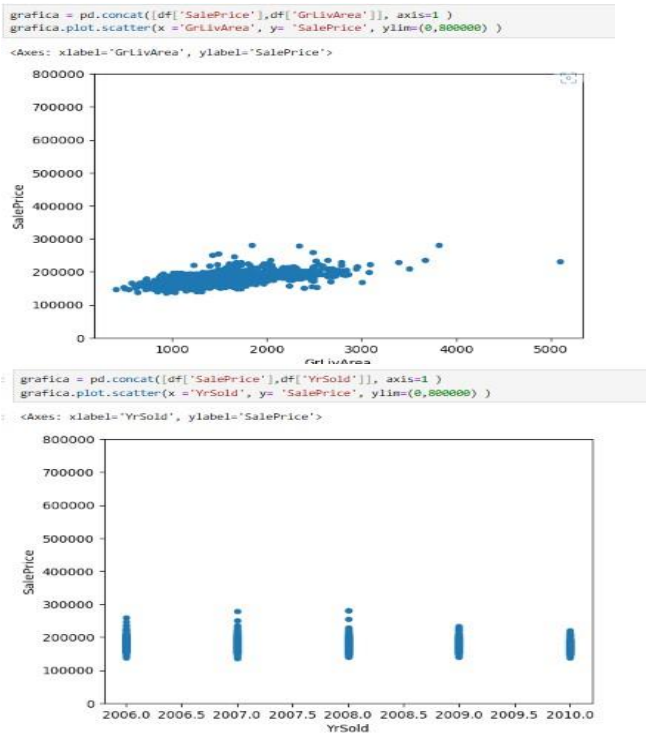
- Data Set House Price: este Data Set cuenta con 1459 registros y 80 campos.
- Data Set Sale Price: este Data Set cuenta con 1459 y registros y 2 campos.
- Otras características: en el Data Set House Price contiene las características con la que cuenta cada una de las casas, cada descripción dada en los campos es de carácter técnico relacionada en el campo en el que se aplica, el cual es bienes raíces o venta de casa. También con una

previa visualización de los datos se puede destacar que hay una gran cantidad de datos nulos así como de columnas que serán filtradas para realizar el proyecto solo con la información necesaria y no tener una cantidad de datos tan voluminosa de igual manera se tratarán de rellenar los datos nulos lo mejor posible ya que la información recaudada este lo más cerca de la realidad.

Selección de herramientas de análisis: para este proyecto que se llevara a cabo en la plataforma Júpiter Lab a través del lenguaje de programación Python utilizando bibliotecas empleadas en la limpieza, visualización y operaciones como; pandas, usada en la manipulación y análisis de datos, matplotlib, utilizada mayormente para la creación de gráficas y diagrama de los datos, numpy, también utilizada en el análisis de datos además de ser utilizado en realizar operaciones matemáticas complejas capaz de hacer matrices y arreglos de gran tamaño Esto para realizar operaciones matemáticas si es necesario y mostrarlo de una forma gráfica para mayor entendimiento.

Con estos primeros resultados se puede determinar que hay características que tienen más valor que otras a la hora de vender una casa como es el caso tan contrastante de la comparación entre el precio de la casa con la medida de área vivible con el año en el que se vendió, donde solo en un caso el área vivible de la casa era muy extenso y su precio oscilaba el promedio mientras que el caso del año en el que se vendió la casa no sigue una forma lineal para con la columna precio de venta sino que las casas vendidas a mayor precio fue entre los años 2007 y 2008, esto es probable que se deba a la crisis llamada la gran Recesión que se dio entre los años 2007 y 2008, según, (Lab, 2021) se dio por “Los préstamos abusivos

dirigidos a compradores de viviendas de bajos ingresos, la asunción de riesgos excesivos por parte de las instituciones financieras mundiales y el estallido de la burbuja inmobiliaria de los Estados Unidos". (Ver anexo 1)




Con esto podemos comprobar que más adelante se deberá elegir las características adecuadas para la revelación de información deseada ya que cada característica es importante según el escenario en el que se trabaje así que se deberá proceder con gran cuidado y detalle para encontrar la información correcta y más cercana a la realidad.

Verificación de la calidad de errores.

A la hora en que se evalúa con más detenimiento los Data Set House Price y Sale Price se encontraron datos nulos. Esto es preocupante ya que son alrededor de 7878 datos nulos encontrados en una búsqueda preliminar de los mismos, esto dificulta que los resultados obtenidos en un futuro sean completamente apegados a la realidad. En una

examinación preliminar se puede observar que hay algunas coincidencias entre los datos nulos de

BsmtExposure y BsmtQual ya que cuentan con la misma cantidad de datos nulos, es importante mencionar que no es en único caso pero si uno de los más significantes ya que es una de las coincidencias con más datos nulos.

Verificación de los datos.

Como se hizo alusión anteriormente en la exploración de los datos fue posible notar que los data set más específicamente House Price existe una cantidad considerable de datos nulos y valores en cero que deben ser limpiados. Aunque no todos los campos son obligatorios ya que las características de una casa son variadas existen casos como Sale Type que es probable que deba ser obligatorio ya que las casas que se encuentran en este Data Set fueron vendidas y necesitan cumplir con esta característica. De esta manera se puede afirmar que los datos de los data Set no están completos en su totalidad.

Verificando la frecuencia con los que ocurren o aparecen datos nulos, se puede observar que existe casos donde la cantidad de datos nulos coincide con otras columnas, como es el caso mencionado anteriormente de las columnas BsmtExposure y BsmtQual, esto se puede ver como algo normal ya que es probable que no todas las casas cuenten con estas características como tener un sótano o exposición al mismo, además que las dos columnas apuntan a un lugar específico el cual es el sótano, lo que reafirma que es normal que existan datos nulos ya que como se dijo anteriormente, las características de una casa son variadas. Al igual que existen los casos en el que una casa no cuente con un sótano también es posible que una casa no cuente con una zona de garaje o área de piscina, esto es normal ya que

existen casos en los que las casas no cuentan con un garaje o piscina, lo que se puede deducir en el caso a la abundancia de datos nulos en los campos de la columna Pool Area y PoolQC ya que el clima de Iowa es un lugar frio y húmedo, según (Kwebeman, 2023), “El estado de Iowa tiene un climas continentales. Es cálido en verano y frío en invierno. La temperatura media anual en Iowa es 14° y la precipitación media anual es 553 mm.”. esto podría explicar la razón de por la cual las casillas relacionadas a piscinas o características relacionadas a casa de verano tengan datos nulos en abundancia.

Aunque esto sea algo normal no es posible trabajar con una cantidad tan abrumadora de datos nulos, es por esto que es necesario emplear técnicas de limpieza de datos mencionadas anteriormente: Eliminación de duplicados, Manejo de datos faltantes y Manejo de valores atípicos, necesarios para revelar información concisa y verídica lo más cercano a la realidad.

Método CRISP-DM

## Fase 3: Preparación de los datos.

La tercera fase del modelo CRISP-DM se le da el nombre de Preparación de los datos. En esta fase se da lugar a varias practicas o técnicas empleadas a la limpieza de los datos tales como; Selección de datos, limpieza de los datos, Construcción de datos, Integración de datos y Dar formato a datos, necesarias para realizar una correcta limpieza y preparación de los datos, con ello desarrollar adecuadamente el método CRISP-DM.

Según, (Niño, CRISP-DM: Fase de “Preparación de los datos”, 2016), El objetivo principal de esta fase es la construcción, a partir de los datos “en crudo”, del Data Set final

a utilizar como datos de entrada para las herramientas de modelado. Las tareas englobadas en esta fase (centradas en la limpieza y transformación de los datos) son susceptibles de realizarse repetidas veces y en un orden que dependerá del caso concreto.

La tercera fase es una de las más necesarias he importantes ya que se preparan limpian y se les da un correcto formato con los datos que se trabajaran. Es por esto que se le debe dar importancia a cada detalle realizado en esta fase al igual que al que se le da a cualquier otro detalle.

Es por esto que a continuación se detalla el proceso de preparación de los datos en el data

Set House-Prices-Advanced resultado de la unión de los data Set Sale Price y House Price.

Primer paso: Carga de los Data Sets.

El primer paso que se realiza en la tercera fase va de la mano con los pasos anteriores el cual fue cargar los data Sets Sale Price y House Price a los cuales se les cambio el nombre para que fuera más fácil su manipulación (House-Prices-Advanced por House Price) por otra parte el data Set Sale Price se decide no cambiarle el nombre puesto que es un nombre de fácil manipulación. Para la carga de estos archivos se decide usar Jupyter Lab por su fácil manejo y popularidad de fácil uso a la hora de cargar los archivos esto por su interactividad con el usuario. Después de cargar los data Set se importa la biblioteca pandas utilizada en el manejo de Data Set en el lenguaje de programación en Python con el comando “import pandas as pd” se utiliza la palabra “pd” como una forma de nombrar a pandas de una manera

más rápida y sencilla lo que agiliza el proceso de la preparación de los datos. Luego de importar la primera biblioteca que se usara se procede a leer ambos data Set con el método ".read_csv" para ambos data Set que en el note book se llamaran "df pasa House Price" y "dfs  pada sale Price", además de utilizar comandos especiales para que la herramienta Júpiter notebook sea capaz de mostrar los resultados en su totalidad y no los comprimaEsto con la finalidad de poder reanalizar los datos de ambos data Set con mayor facilidad desde el notebook creado para la preparación de los datos y poder tomar decisiones con base a lo encontrado en ambos data Sets. (Ver Anexo 2).

```
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

Segundo paso: Análisis preliminar de los datos

Después de la creación del notebook, carga de ambos data Sets, la aplicación de comandos para mejorar la visualización d los datos y la importación de la librería pandas. Se decide hacer un análisis más exhaustivo y profundo complementario al que se realizó en la segunda fase del método CRISP-DM Comprensión de los datos el cual brindara un escenario más completo sobre la calidad de los datos de los data Sets esto se realiza de la mano de los comandos:

- .head(): utilizado para ver la cabecera de los data Set y los primeros datos aunque se puede regular la cantidad de filas indicando el número de filas deseadas entre los paréntesis.
- .describe(): muestra valores importantes de todas las columnas tipo numéricas (float, int64) como el valor de los cuartiles la desviación

estándar la cantidad de valores, el promedio entre otras aunque se puede elegir una o varias columnas en específico con ayuda de otros acrónimos.

- .info(): muestra información relevante de cada una de las columnas de los data Sets como la cantidad de datos no nulos que tiene cada columna así como el tipo de dato(object, float,int64).
- Isnull().sum: ambos comandos combinados muestran la cantidad de datos nulos por columna lo que ayuda a ver el verdadero escenario sobre la calidad de los data Sets con los que trabajamos.

Luego de aplicar esta serie de comandos se reafirma que existe siete mil ochocientos setenta ocho datos nulos en el data Set House Price, por otra parte el data Set Sale Price se encuentra encentó de datos nulos como de datos en cero.

Tercer Paso: Reconocimiento de los conceptos

Aunque ya se llevó un reconocimiento del negocio y los datos, es importante llevar a cabo un análisis más exhaustivo y profundo sobre el tema de venta de bienes raíces, es por esto que como tercer paso se dio a la tarea de investigar cada concepto ya sea de un campo o de alguna columna del cual se desconocía el concepto de esta, por lo cual se considera realizar un documento tipo diccionario donde se anota cada concepto de la columna y el significado de los posibles contenidos que se podría encontrar en los campos de cada una de las columnas.

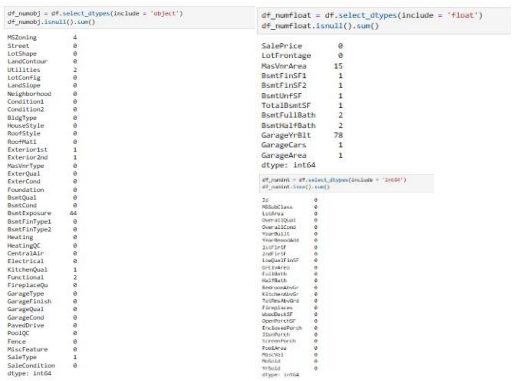
Esto con el fin de entender que tan importante o representativa fuese cada columna para el data Set o que tan relacionada pudiese estar con el data Set Sale Price, además de

permitir la oportunidad de entender cada concepto y su significado en el campo de bienes raíces necesarias para abordar el tema de la preparación de los datos relacionadas con este proyecto. Esto resulta en un proceso importante ya que para un adecuado manejo de los datos es importante conocer los conceptos de los datos con los que se trabaja para realizar un trabajo lo más concientizado posible hacia el entendimiento he importancia que los datos pueden brindar para encontrar información relevante.

Cuarto paso: Seccionamiento del data Set House Price

Después de estudiar el documento con los conceptos investigados se optó por seccionar las columnas momentáneamente por sus tipos de variables (object, float, int64) esto con el fin de un manejo y visualización de los datos más sencillo y eficiente, esto con el objetivo de tomar decisiones con respecto a los datos nulos que estas columnas contenían.

Este seccionamiento del data Set House Price fue posible gracias a la implementación del método (select_dtypes()) al cual se le puede dar los parámetros para almacenar distintos tipos de variables ya sea object, float, int64. Esto se realiza con el comando “select_dtypes”(Ver anexo 3)



Por lo que se procede a almacenar en variables los datos seccionados por sus tipos de variables se debe tomar en cuenta que para este caso se usaron variables fáciles de escribir y relacionadas con lo que almacenan. A continuación se muestran el nombre de las variables y que almacenan:

- df_numobj: en esta variable se almacenan las columnas del data Set tipo objeto o texto
- df_numfloat: en esta variable se almacenan las columnas del data Set tipo flotantes o números que contengan decimales
- df_numint: en esta variable se almacenan las columnas del data Set tipo números enteros

Quinto paso: Investigación de técnicas de limpieza de datos

Luego de seccionar el data Set se procede a investigar diferentes técnicas que se pueden emplear para el manejo de los datos nulos del data Set.

Como es de esperar los datos nulos ubicados en variables tipo objetos y tipo numéricos no es posible tratarlos de igual manera, es por esto que a continuación se investiga diferentes técnicas de limpieza de datos que se aplicaran a este proyecto.

Eliminar los valores nulos: esta técnica consiste en eliminar los datos nulos de la columna y con el dato toda la fila en donde se encontraba dicho dato. Esta técnica es viable cuando una fila contiene gran cantidad de datos nulos, por lo que se debe usar solo en casos especiales para no reducir la cantidad de los datos con los que se trabajan. Según, (Jiménez,

2024) “Sin embargo, esta técnica puede llevar a la pérdida de información importante y reducir el tamaño de la muestra, lo que podría afectar la validez de los resultados.

Imputación de valores: esta técnica consiste en completar los datos nulos de las columnas con base a predicciones como la moda y el promedio y también se puede realizar una imputación relacionando columnas por como GarageCars y GarageArea ya que se puede hacer predicciones con ambas columnas.

Según, (Jiménez, 2024), La imputación de valores consiste en reemplazar los valores nulos por estimaciones o predicciones. Esto se puede hacer de diferentes maneras, como sustituir los valores nulos por la media, la mediana o el valor más frecuente del conjunto de datos.

Análisis por subgrupos: esta técnica se puede relacionar con las anteriores. Consiste en seccionar el data Set en sub grupos que guarde similitudes o compartan características entre sí. Según, (Jiménez, 2024),” Esto puede ser útil cuando existen diferencias significativas entre los subgrupos y los valores nulos están relacionados con alguna característica específica.”

Sexto paso: aplicación de técnicas de limpieza de datos al data Set House Price

En este sexto paso se aplicará las técnicas investigadas en el paso anterior por lo que a continuación se presentaran las decisiones y técnicas empleadas en cada uno de los sub grupos mencionados anteriormente.

Decisiones tomadas con las columnas tipo objetos

Para este sub grupo se optó por utilizar técnicas de imputación y eliminación datos esto con la visión de mantener o recuperar la mayor cantidad de datos y perder la menor cantidad de los datos con los que cuenta el Data Set.

En el caso de la columna Alley realizó una técnica de eliminación la columna Alley con ayuda del comando “drop” esto por la razón de la existencia de una gran cantidad de datos nulos además que la columna Streep podría sustituirla fácilmente. Por otra parte a la existencia de estas dos columnas en el data Set podría causar redundancia en los datos.

En el caso de las columnas Neighborhood, MasVnrType, BsmtQual, BsmtCond, BsmtFinType1, BsmtFinType2, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PoolQC y Fence se opta por completar los datos nulos pon el mensaje Desconocido esto porque resulta imposible predecir los valores o por que se deduce que simplemente la casa no cuenta con cierta parte. Esta técnica se realizó con el comando “fillna”.

Por otra parte en el caso de las columnas Neighborhood, MSZoning, Utilities, Exterior1st, Exterior2nd, KitchenQual, Functional y SaleType se emplean técnicas de imputación de datos nulos para completar los datos faltantes con la moda, según la moda de cada columna a la que pertenecían los datos nulos. Esta técnica se realizó con el comando “fillna y mode”

Decisiones tomadas con las columnas tipo decimales

Para este sub grupo de columnas se eligió usar técnicas de imputación de datos con la idea de reducir la cantidad de datos perdidos, de esta manera se podrán realizar la fase de

modelado de los datos con una cantidad de datos suficientes para encontrar información relevante.

En el caso de las columnas LotFrontage, MasVnrArea se completaron los datos nulos con el promedio de las columnas a la que pertenece cada dato faltante. Esta técnica es aplicada para completar los datos de manera que los resultados estén lo más cerca de la realidad. Esta técnica se realizó con los comando “fillna y mean”

Por otra parte se opta completar los datos de las columnas BsmtFinSF1, GarageYrBlt, BsmtFinSF2, BsmtUnfSF, BsmtFullBath, GarageCars y GarageArea con el valor “0.0 ” esto gracias a que se realizó un estudio y comparaciones que revelo que los datos faltantes de estas columnas eran nulos porque no contaban con esa área de la casa por lo que se toma la decisión de completarlos con valores que reflejen el faltante de este por lo que se opta por el valor 0.0 además que al realizar la sumatoria de estos valores no afectara en resultado. Esta técnica se realizó con el comando “fillna”

En consecuencia del estudio de estas variables las columnas GarageCars, GarageYrBlt y YrSold se decide cambiar su tipo de variable por tipo entero esto para mejor visualización y coherencia de los valores ya que parece imposible la existencia de números decimales en estas columnas, también la columna LotFrontage solo se le permite contener un decimal esto para mejorar su visualización y manejo de los datos en estas columnas. Esto se realizó con los comandos “astype, pop y insert ”

Números enteros

Al realizar las visualizaciones y comparaciones necesarias para el estudio de este subgrupo de números enteros, se descubre la inexistencia de datos nulos en el subgrupo de los datos nulos y la normalización de los valores en cero que se pudieran encontrar en las columnas de estos grupos esto se debe a que es posible que refleje la inexistencia de una parte de la casa.

Ordenamiento de los datos

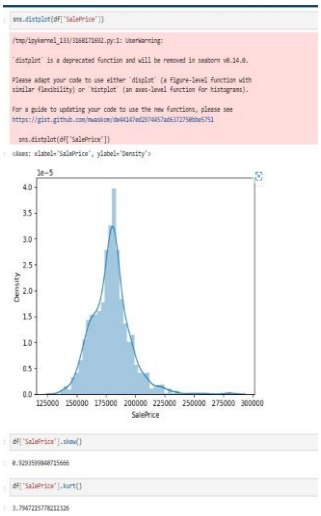
Para una mejor visualización de los datos se inserta la columna Sale Price a la columna

House Price creando un nuevo data Set House Price Advansed con el cual se realizará la fase de modelado con el comando “concat”, además se reacomoda la columna Sale Price en segunda posición esto para una mejor visualización de los datos.

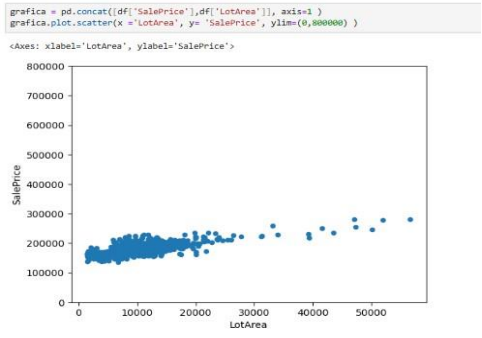
Visualización de los datos

Después de la limpieza de los datos nulos se crean graficas donde se investiga las columnas que inciden en el precio de las casas del Data Set House_Prices_Advanced. Por lo que a continuación se comenta los resultados que más se destacan.

Se crea una gráfica con la biblioteca “seaborn” y los comando “skew y kurt” a la cual se le aplica a la columna Sale Price y se detecta que hay un sesgo en dicha columna por lo que reafirma que algunas columnas inciden en el precio de la casa.(Ver anexo 4)

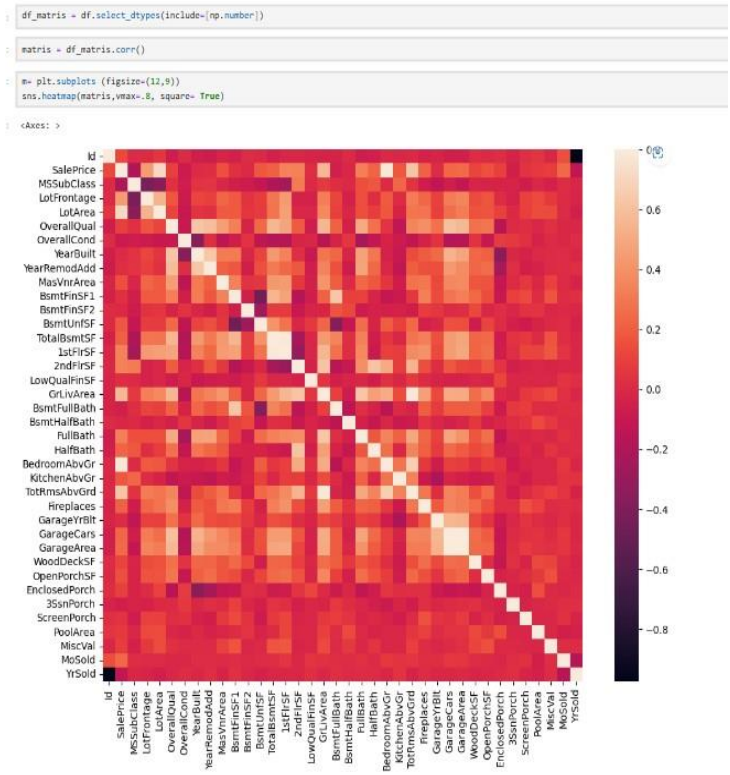



Luego se decide crear grafica “plot.scatter y boxplot” de la biblioteca “matplotlib y seaborn” donde se evalúa la relación de ciertas columnas con la columna Sale Price como son el caso de las columnas LotArea y OverallQual, lo que resulta en la visualización de una clara incidencia por parte de la columna LotArea hacia Sale Price pues a mayor cantidad de pies cuadrados de una propiedad mayor es el precio de la casa. También se decide aplicar el mismo proceso a la columna BedroomAbvGr la cual sigue una trayectoria parecida a LotArea lo que significa que a mayor cantidad de cuartos el precio de la casa subirá. (ver Anexo 5)




Luego de esto se realiza una matriz que a través de colores nos mostrara las columnas más relacionadas con la columna Sale Price. Esto con los comandos:

“select_dtypes, corr, subplots, heatmap” y las bibliotecas, numpy y matplotlib donde se revela la siguiente información: (Ver Anexo 6).




Se está en lo correcto al mencionar que las columnas BedroomAbvGr y LotArea incidan en el precio de las casas del Data Set puesto que ambas muestran una correlación de BedroomAbvGr 0.72 y LotArea 0.79 siendo el 1.00 el valor más cercano el cual es el mismo Sale Price relacionado con sí mismo.

Además de esta columnas se encuentran otras columnas que inciden en el Precio las cuales son TotRmsAbvGrd(Número total de habitaciones (excluyendo baños)) con un valor de 0.63, GrLivArea (Área habitable sobre el nivel del suelo en pies cuadrados) con el valor de 0.57 y LotFrontage(longitud de la frontera de la propiedad con la calle en pies.) con un

valor de 0.45. Existen otras columnas como 1sFlrSF 2sFlrSF, FullBath y GarageArea los cuales también se relacionan con la columna SalePrice

Se puede considerar que las características a tomar en cuenta al comprar una casa en el estado de Iowa son LotArea, GrLivArea, BedroomAbvGr, TotRmsAbvGrd, LotFrontage, 1sFlrSF 2sFlrSF, FullBath y GarageArea.

Método CRISP-DM

## Fase 3.2: Reajustes de la Preparación de los Datos

En esta fase del método CRIS-DM, Preparación de los datos, uno de los objetivos principales es aplicar las técnicas necesarias para pasar de un dataset crudo a un dataset listo para la fase de modelado de datos, donde se eliminan e imputan filas y columnas para dar como resultado un data set libre de datos nulos y datos dummies además de seleccionar las columnas necesarias para la próxima fase la cual es la fase de modelado de datos.

Según, (Mikelnino, El blog de Mikel Niño, 2016) , El objetivo principal de esta fase es la construcción, a partir de los datos "en crudo", del dataset final a utilizar como datos de entrada para las herramientas de modelado. Las tareas englobadas en esta fase (centradas en la limpieza y transformación de los datos) son susceptibles de realizarse repetidas veces y en un orden que dependerá del caso concreto.

En el anterior proyecto se trabajó con esta fase, Preparación de los datos donde se aplicaron diferentes técnicas para que fura aplicada con éxito pero para

la fase de modelado de datos es necesario realizar algunos cambios en la aplicación de las técnicas de eliminación e imputación de los datos con el objetivo de que los adormidos que se utilizaran en la fase de modelado de datos revele información correcta que será necesaria para realizar recomendaciones acertadas.

Por lo que a continuación se documenta los cambios realizados en las técnicas de preparación de los datos del dataset House-Price-Advanced .

Carga de los data set

En primera instancia, se toma la decisión de reiniciar la preparación de los datos desde cero, esto implica que se debe volver a cargar los data ser House-Price-Advanced y Sale Price tal y como fueron descargados originalmente del repositorio de Git-Hub del Dr. Samuel Saldaña Valenzuela a la plataforma Jupyter Lab esto con el objetivo principal de enfocarse  corregir los errores que se cometieron en la anterior preparación de los datos. Al cargar ambos dataset nuevamente en la plataforma Jupyter Lab se asegura la solides de las bases para un correcto análisis y construcción de los datos que será utilizados para la aplicación de la siguiente fase, modelado de datos de machine learning. Lo cual es fundamental para obtener resultados precisos y confiables para el proyecto además de desarrollar las técnicas de limpieza de los datos con mayor comodidad.(Ver Anexo 7)

```
# carga de las bibliotecas
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

```
# carga de los Datasets
```

```
df_ha =pd.read_csv('House-Prices-Advanced.csv')
```

```
df_sp =pd.read_csv('SalePrices.csv')
```

```
# comandos para ver todas las columnas
```

```
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

```
# union de los data set y previa visualizacion de del dataset
```

```
df_u = pd.concat([df_sp['SalePrice'],df_ha], axis = 1)
df_ha = df_u

df_ha.to_csv('House-Prices-Advanced.csv', index=False)
df_ha.head()
```

Visualización de los datos profunda.

Luego de realizar la carga de los dataset ser House-Price-Advanced y Sale Price se decidió emplear diferentes comandos para su visualización de aspectos importantes como columnas, cantidad de filas, tipos de variables, cantidad de datos nulos y números en cero por columna, se realizó con los comandos:

- .head(): utilizado para ver la cabecera de los data Set y los primeros datos aunque se puede regular la cantidad de filas indicando el número de filas deseadas entre los paréntesis.
- .describe(): muestra valores importantes de todas las columnas tipo numéricas (float, int64) como el valor de los cuartiles la desviación estándar la cantidad de valores, el promedio entre otras aunque se puede elegir una o varias columnas en específico con ayuda de otros acrónimos.

- .info(): muestra información relevante de cada una de las columnas de los data Sets como la cantidad de datos no nulos que tiene cada columna así como el tipo de dato(object, float,int64).
- .Isnull().sum: ambos comandos combinados muestran la cantidad de datos nulos por columna lo que ayuda a ver el verdadero escenario sobre la calidad de los data Sets con los que trabajamos.
- .apply(lambda col: (col == 0).sum()) : esta combinacion de comandos muestran la cantidad de datos en cero que tiene cada columna del data set

Esto con el objetivo de realizar un análisis más profundo de los datos además de tomar decisiones respecto a los datos nulos y datos en cero del dataset.

También con el comado .concat() se une la columna SalePrice del data set Sale Price con el data set House-Price-Advanced. Esta decisión de unir la columna Sale Price con el data set House-Price-Advanced desde el inicio de la fase de preparación de los datos considerando la probabilidad de que algunas filas del dataset House-Price-Advanced sean eliminadas durante el proceso de limpieza del data set House-Price-Advanced.

El comando .concat() facilita la concatenación de los DataFrames, permitiendo que la nueva columna "SalePrice" se integre correctamente con las demás columnas del dataset
"House-Price-Advanced".

Seccionamiento del data set House-Price-Advanced

Luego de visualizar los aspectos mencionados anteriormente, se toma la decisión seccionar el dataset House-Price-Advanced de manera temporal agrupando las columnas seccionadas por sus diferentes tipos de variables esta estrategia o técnicas realiza con objetivo facilitar la aplicación de técnicas de limpieza de datos en los subgrupos para crear códigos más optimizados. A continuación se documenta el seccionamiento realizado al data set House-Price-Advanced:

- Variables categóricas: en este subgrupo se encuentran las columnas que acumulan en su registro variables categóricas o tipo texto (Ver Anexo 8).

```
df_numob = df_ha.select_dtypes(include = 'object')
df_numob.isnull().sum()
```

- Variables tipo números enteros: en este subgrupo se encuentran las columnas que acumulan en su registro variables tipo numéricas en específico números enteros.(Ver Anexo 9)

```
df_numint = df_ha.select_dtypes(include = 'int64')
df_numint.isnull().sum()
```

- Variables tipo números decimales: en este subgrupo se encuentran las columnas que acumulan en su registro variables tipo numéricas en específico números decimales(Ver Anexo 10)

```
df_numfloat = df_ha.select_dtypes(include = 'float')
df_numfloat.isnull().sum()
```

La documentación y el desarrollo de esta técnica de seccionamiento de columnas resulta ser esencial para el desarrollo e implementación de las siguientes técnicas de eliminación e imputación de los datos que realizaran en el futuro. Esta técnica organiza y simplifica el proceso de limpieza de datos y además que también proporciona una base estructurada sobre la cual se pueden construir análisis más complejos.

Investigación de técnicas de limpieza de datos

En el campo de la ciencia de datos, la limpieza de los datos es uno de los procesos más importantes que se relaciona directamente con la calidad y precisión de los resultados que se obtiene en etapas posteriores como el modelado de datos. Para este segundo proyecto de los se toma la decisión de utilizar las mismas técnicas de limpieza de datos que en el en el proyecto anterior, por la razón de que son las que mejor se adaptan al desarrollo de este proyecto.

Por lo que a continuación se documenta el funcionamiento de cada una de las técnicas que se utilizaran

Imputación de datos nulos

La técnica de imputación de datos nulos es una técnica que busca rellenar o sustituir los valores faltantes o nulos a partir de predicciones con base a ciertos aspectos como los valores encontrados en los registros de la columna a la que se le aplique dicha técnica o bien columnas relacionadas a la misma. La imputación de datos es una técnica que busca mantener la integridad de los datos y evitar que estos se borren.

Según, (Estadisticool, 2023) , La imputación de valores faltantes es un método estadístico utilizado para estimar valores perdidos o incompletos. Este método asigna valores a los datos faltantes en base a la información disponible de los datos existentes. La imputación de valores faltantes se puede utilizar cuando hay datos perdidos o faltantes en un conjunto de datos y se desea estimar estos valores en base a la información disponible. El método de imputación de valores faltantes es útil cuando se trata de datos perdidos o incompletos, ya que permite que los datos sean utilizables para el análisis. Sin embargo, este método no es perfecto y puede introducir sesgos en los resultados del análisis.

Existen diferentes métodos para utilizar esta técnica así que se debe analizar de manera profunda cada columna para decidir que método se adapta mejor a las necesidades del dataset en este caso al dataset House-Price-Advanced. Algunos de estos métodos son:

- Método de mediana: este método consiste en realizar la imputación de los valores faltantes de la columna con la que se trabaja con el valor de la mediana de dicha columna. Esto con el objetivo de no desbalancear los el resto de los datos que alberga dicha coluna.
- Método del valor más probable: este método consiste en imputar los datos faltantes de una columna con su respectiva moda o valor más repetitivo.

- Método de la regresión: este método consiste en desarrollar un código para predecir cual será el valor nulo o faltante esto con base a los existentes.
- Método de la interpolación este método busca imputar los datos de una columna con base a los valores existentes que se encuentran cerca de cada valor perdido o faltante.
- Método de la extracción: este método utiliza una técnica de imputación que consiste en realizar un muestreo aleatorio donde se selecciona valores ya existentes que se utilizan para estimar los valores faltantes de la columna donde se aplica esta técnica.
- Método de imputación aleatoria: dicho método de imputación consiste en asignar a los valores nulos un valor de manera aleatoria utilizando rangos de numero ya sea mediana cuartiles, mediana o bien rangos entre esos valores.

Eliminación de datos nulos

Esta técnica de limpieza de datos por el contrario de la imputación de datos que busca imputar o rellenar los datos, esta técnica busca eliminar los datos nulos ya sean filas o columnas. Esta técnica resulta viable en estos escenarios.

Según, (Schoools, 2022) , Los valores faltantes en un conjunto de datos pueden ser problemáticos para el análisis. Una técnica común es eliminar las filas que contienen valores faltantes. Sin embargo, esta técnica debe usarse con precaución, ya que puede llevar a la pérdida de información relevante.

Altos porcentajes de datos nulos: en este tipo de escenarios resulta más eficiente eliminar las columnas que no resultan ser precisamente necesarias para el modelado de datos cuando el porcentaje de datos es alto.

Filas poco significativas en este escenario plantea que las filas poco relevantes o para o que representen una minoría en el dataset para análisis de los datos pueden ser eliminados.

Es importante recordar que ambas técnicas, tanto imputación de datos como eliminación de datos, deben ser aplicadas con especial cuidado para evitar problemás significativos en etapas posteriores, tal y como son los casos relacionados a perdidas excesivas, es esencial prevenir la perdida excesiva de los datos para realizar análisis en etapas posteriores. De mismo modo la imputación de los datos debe ser mejorados de manera adecuada y buscar el método adecuado para evitar problemás de dimensionalidad, sesgos o inexactitudes de los datos. por lo que para el desarrollo de la fase de preparación de los datos se intentara proceder con ambas técnicas de la manera más consciente posible para evitar problemás con el desarrollo de etapas posteriores.

Aplicación de las técnicas de limpieza de datos

Después de realizar una exhaustiva investigación sobre cada uno de los métodos y escenarios en los que son viable utilizar las técnicas de eliminación e

imputación de datos, se procede a documentar el uso que se le dio a ambas técnicas y los resultados obtenidos.

Imputación de los datos

Después de segmentar los datos por sus tipos de variables y analizar cada subgrupo, se decidió imputar los datos nulos y aquellos valor fueran cero. Este proceso tiene como objetivo de no tener problemás cuando aplicaran técnicas de regresión algebraica.

Por lo que para las columnas con variables enteras, los valores cero fueron remplazados por el numero uno (1) mientras que para las columnas con variables numéricas decimales se utilizó el digito uno punto cero (1.0). Esta técnica de imputación se utilizaron listas y comando .remplace() esto para para realizar la imputación de manera más ordenada y optimizada (Ver Anexo 11).

```
col_r= ['2ndFlrSF','LowQualFinSF','HalfBath','Fireplaces','WoodDeckSF','OpenPorchSF','EnclosedPorch','3SsnPorch','ScreenPorch','PoolArea','MiscVal']
```

```
valor_r = 1
```

```
#aplicacion de tecnicas de imputacion
```

```
#impu
df_ha[col_r]=df_ha[col_r].replace(0,valor_r)
df_ha.to_csv('House-Prices-Advanced.csv', index=False)
df_ha[col_r].head(5)
```

| | 2ndFlrSF | LowQualFinSF | HalfBath | Fireplaces | WoodDeckSF | OpenPorchSF | EnclosedPorch | 3SsnPorch | ScreenPorch | PoolArea | MiscVal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 | 140 | 1 | 1 | 1 | 120 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 393 | 36 | 1 | 1 | 1 | 1 | 12500 |
| 2 | 701 | 1 | 1 | 1 | 212 | 34 | 1 | 1 | 1 | 1 | 1 |
| 3 | 678 | 1 | 1 | 1 | 360 | 36 | 1 | 1 | 1 | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 | 1 | 82 | 1 | 1 | 144 | 1 | 1 |

Por otra parte, la imputación de los valores nulos se realizó con el método de imputación aleatoria. Para llevar a cabo este método acabo se utilizar funciones listas diccionarios (en el caso de las variables categóricas) y comando o acrónimos como “rd.choice”, “.apply” y “lambda” utilizando. Para realizar este

método se tomaron en consideración medianas y números cercanos a las mismas de cada de las columnas a las que se aplicaban estas columnas y en algunos casos los cuartiles cuando el mínimo y la mediana tenían el mismo resultado como es el caso de la columna MásVnrArea (Ver Anexo 12) .

```
# imputacionde datos por Método de imputación aleatoria
import pandas as pd
import random as rd


# calcular el segundo cuartil de la columna 'MasVnrArea'
segundo_cuartil = df_ha['MasVnrArea'].quantile(0.50)
print("Segundo cuartil:", segundo_cuartil)

# Definir valores cercanos al segundo cuartil
valores_rd = [segundo_cuartil - 1, segundo_cuartil + 1]

# Función para seleccionar aleatoriamente un valor de la lista
def ran():
    return rd.choice(valores_rd)

# rellenar los valores nulos o ceros en la columna 'MasVnrArea'
df_ha['MasVnrArea'] = df_ha['MasVnrArea'].apply(lambda x: ran() if pd.isnull(x) or x == 0 else x)

# Mostrar el resultado
df_ha['MasVnrArea'].describe()

# guardar el DataFrame modificado en un archivo CSV
df_ha.to_csv('House-Prices-Advanced.csv', index=False)

df_ha['MasVnrArea'].describe()
```

Por otro lado para las columnas con variables categóricas se aplicó un enfoque similar utilizando la biblioteca numpy, método para crear variables aleatorias que incluyeran todas las categorías de cada columna. Este enfoque asegura una imputación efectiva y robusta de los valores nulos y valores en cero asegurando la eficacia de etapas posteriores. (Ver Anexo 13)

```python
import random as rd
import pandas as pd
import numpy as np

# imútacion de datos por el Método de imputación aleatoria
data = {
    'Alley': [np.nan, 'Grvl', 'Pave', np.nan, 'Grvl'],
    'MasVnrType': ['BrkFace', np.nan, 'Stone', 'BrkCmn', np.nan],
    'FireplaceQu': [np.nan, 'Gd', 'TA', 'Fa', 'Po'],
    'GarageFinish': ['Unf', 'RFn', np.nan, 'Fin', np.nan],
    'PoolQC': [np.nan, np.nan, 'Ex', np.nan, 'Gd'],
    'Fence': [np.nan, 'MnPrv', 'GdPrv', 'GdWo', np.nan],
    'MiscFeature': [np.nan, 'Shed', np.nan, 'Gar2', 'Othr']
}
df_hau = pd.DataFrame(data)

# Definir la función ran() para devolver un valor aleatorio basado en val
def ran(val):
    return rd.choice(val.index)

# Columnas categóricas con valores nulos
val_im_ob = ['Alley', 'MasVnrType', 'FireplaceQu', 'GarageFinish', 'PoolQC', 'Fence', 'MiscFeature']

# Imputar valores aleatorios en las columnas especificadas solo para valores nulos
for column in val_im_ob:
    val = df_ha[column].value_counts()  # Contar los valores únicos y sus frecuencias
    df_ha[column] = df_ha[column].apply(lambda x: ran(val) if pd.isnull(x) else x)

# Mostrar el DataFrame modificado
df_ha.head()

# Guardar el DataFrame modificado en un archivo CSV
df_ha.to_csv('House-Prices-Advanced.csv', index=False)
```

```python
import numpy as np
import pandas as pd

# imutacion de datos por el Método de imputación aleatoria
data = {
    'BsmtQual': ['TA', 'Gd', 'Ex', 'Fa', np.nan],
    'BsmtCond': ['TA', 'Gd', 'Fa', 'Po', np.nan],
    'BsmtExposure': ['No', 'Av', 'Gd', 'Mn', np.nan],
    'BsmtFinType1': ['GLQ', 'Unf', 'ALQ', 'Rec', 'BLQ'],
    'BsmtFinType2': ['Unf', 'Rec', 'LwQ', 'BLQ', 'ALQ'],
    'KitchenQual': ['TA', 'Gd', 'Ex', 'Fa', np.nan],
    'FireplaceQu': ['Gd', 'TA', 'Fa', 'Po', 'Ex'],
    'GarageCond': ['TA', 'Fa', 'Po', 'Gd', 'Ex'],
    'SaleType': ['WD', 'New', 'COD', 'ConLD', 'CWD'],
    'Utilities': ['AllPub', np.nan, np.nan, np.nan, np.nan],
    'MSZoning': ['RL', 'RM', 'FV', 'RH', 'C (all)'],
    'Exterior1st': ['VinylSd', 'MetalSd', 'HdBoard', 'Wd Sdng', 'Plywood'],
    'Exterior2nd': ['VinylSd', 'MetalSd', 'HdBoard', 'Wd Sdng', 'Plywood'],
    'Functional': ['Typ', 'Min2', 'Min1', 'Mod', 'Maj1'],
    'Neighborhood': ['CollgCr', 'Somerst', 'OldTown', 'NridgHt', 'Gilbert']
}

# Convertir el diccionario a un DataFrame
df_hau = pd.DataFrame(data)

# Lista de columnas categóricas
columnas_categoricas = ['BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1',
                        'BsmtFinType2', 'KitchenQual', 'FireplaceQu', 'GarageCond',
                        'SaleType', 'Utilities', 'MSZoning', 'Exterior1st', 'Exterior2nd',
                        'Functional', 'Neighborhood']

# Imputar valores nulos aleatoriamente con las categorías existentes
for col in columnas_categoricas:
    # Obtener las categorías existentes en la columna
    categorias_existentes = df_ha[col].dropna().unique()
    # Generar categorías aleatorias para seleccionar categorías
    random_index = np.random.randint(0, len(categorias_existentes), size=df_ha[col].isnull().sum())
    # Asignar valores aleatorios a los valores nulos
    df_ha.loc[df_ha[col].isnull(), col] = categorias_existentes[random_index]

# Mostrar el DataFrame resultan
df_ha.head()

df_ha.to_csv('House-Prices-Advanced.csv', index=False)
```

Eliminación de datos

Después de completar imputación de las columnas que contenían una mayor cantidad de datos, se decide que esta técnica de eliminación solo será aplicada a las columnas cuya cantidad de valores no supere la cantidad de una cifra por columnas. En síntesis una columna será una posible candidata para aplicar la eliminación de filas si la cantidad de datos nulos de dicha fina no sobrepase dígitos de una cifra. Además deberá comprobarse que no comprometa información importante para futuras etapas. Por lo que se toma la decisión de realizar esta técnica solamente un estas columnas del dataset House-Price-Advanced. FullBath, BedroomAbvGr, KitchenAbvGr ya que contaban con cantidades de nulos muy bajas lo que representa una perdida mínima de datos para el data set. (Ver Anexo 14)

```
import pandas as pd

#Eliminación de los datos nulos en las filaes de las columnas

columnas_d = ['FullBath', 'BedroomAbvGr', 'KitchenAbvGr']

df_ha = df_ha[~(df_ha[columnas_d] == 0).any(axis=1)]

# Mostrar el DataFrame modificado
df_ha.head(1)

df_ha.to_csv('House-Prices-Advanced.csv', index=False)
```

El objetivo principal de esta decisión conforme a la técnica de eliminación es evitar pérdidas de los datos del dataset House-Price-Advanced, lo que podría resultar en la reducción de la cantidad de información disponible que podría recolectarse en fases posteriores como la fase de modelado de los datos. Mantener una cantidad suficiente de datos es fundamental para asegurar que los modelos resultantes sean precisos y robustos los cuales son necesario para comentar las recomendaciones necesarias al negocio generadora de los datos. Garantizar que el dataset siga siendo representativo y útil para las etapas posteriores de análisis y modelado es crucial en esta etapa del proceso.

Resultados encontrados posteriores a la etapa de preparación de los datos

La aplicar las técnicas de imputación y eliminación de los datos fue posible observa ciertos cambios en el dataset House-Price-Advanced, por lo que a continuación se documentaran los cambios más visibles luego de realizar la preparación de los datos.

Cambios en los valores de las columnas

Al realizar técnicas de imputación y eliminación de datos con el método de imputación aleatoria se utilizaron números cercanos a la media

específicamente un número menor y un número mayor más cercanos a la media de cada una de las columnas a las que se le aplico esta técnica a sus valores nulos, también es importante recordar que a todos los datos de las columnas House-Price-Advanced que tuvieran datos en cero fueron remplazados por el numero 1 o 1.0 impedientemente de su tipo de variable. Por lo que se puede observar cambios en datos como el promedio los datos mínimos, la cantidad de filas y la desviación estándar de las columnas un ejemplo de eso son las columnas 1stFlrSF 2ndFlrSF los cuales experimentaron cambios en los valores mencionados anteriormente ya que su valor mínimo paso de ser cero a uno además de sufrir cambios mínimos en la cantidad de filas la desviación estándar y el promedio de estas columnas. Esta información se revelo usando el comando .describe() aplicado al todo el dataset. (Ver anexo 15)

Después

| | 1stFlrSF | 2ndFlrSF |
|---|---|---|
| count | 1452.000000 | 1452.000000 |
| mean | 1154.841598 | 328.112259 |
| std | 392.472074 | 420.566271 |
| min | 407.000000 | 1.000000 |
| 25% | 872.000000 | 1.000000 |
| 50% | 1079.500000 | 1.000000 |
| 75% | 1382.250000 | 676.000000 |
| max | 5095.000000 | 1862.000000 |

Antes

| | 1stFlrSF | 2ndFlrSF |
|---|---|---|
| count | 1459.000000 | 1459.000000 |
| mean | 1156.534613 | 325.967786 |
| std | 398.165820 | 420.610226 |
| min | 407.000000 | 0.000000 |
| 25% | 873.500000 | 0.000000 |
| 50% | 1079.000000 | 0.000000 |
| 75% | 1382.500000 | 676.000000 |
| max | 5095.000000 | 1862.000000 |

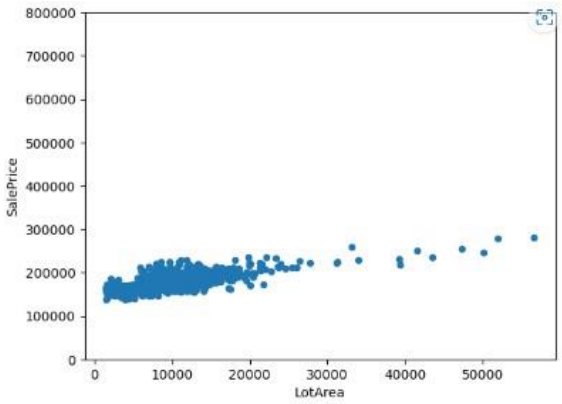
Cabios en la distribución de los datos al compararlos con la columna SalePrice

De igual manera de que los datos cambiaron datos como el promedio los datos mínimos, la cantidad de filas y la desviación estándar de las columnas se puede observar con facilidad que han ocurrido cambios significativos en la distribución de los datos y graficas tipo plot.scatter  al compararlas con columnas la relación o distribución con la columna SalePrice.

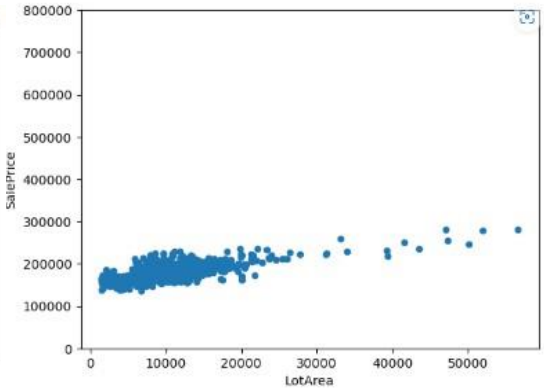
Es posible que los datos de se hayan cambiado al aplicar las técnicas de eliminación o imputación de los datos aplicada en la fase de preparación de los datos se presume que los datos lejanos al promedio de las columnas al hacer las comparaciones hayan desparecido de su posición original a dos a dos posibles opciones. Es probable que al cambiar o imputar datos en la gráfica se hayan los datos se hayan recolocado en una posición más cercana al promedio o pudieron ser eliminados al eliminar alguna filas del dataset House-PriceAdvanced las datos fueron eliminados y se pudo apreciar de mejor manera el faltante de algunos datos al realizar graficas tipo scatter plots  o gráficas de dispersión.

Estos cambio en la distribución de los datos del dataset House-Price-Advanced resulta ser importante para el entendimientos del impactos de las fase de preparación de los datos. Esto se puede visualizar la relación representada en la gráfica de LotArea y SalePrice donde se pueden comparar la comparación antes y después de realizar la fase de preparación de los datos. Se pueden visualizar cambios en la distribución de los datos lejanos al promedio de los datos o los más lejanos a los datos más agrupados estos datos cercanos sufrieron cambien en su distribución además de que algunos datos fueron eliminados. (Ver Anexo 16)

Después




Antes




Cambios en los tipos de variables de algunas variables

Para un mejor tratamiento de los datos algunas columnas cambiaron su tipo de variables algunas variables fueron cambiadas esto para un mejor manejo de estas a realizar la preparación de los datos y posteriormente la fase de modelados de esto se puede ver las columnas de GarageCars GarageYrBlt. Esto con los comandos .astype

## Fase 4: Modelado de los datos

La cuarta fase del Método CRISP-DM se le da el nombre de Modelado de datos. Dicha fase de modelado de datos es aplicar diferentes técnicas y algoritmos para encontrar información relevante que ayude a realizar conclusiones y recomendaciones necesarias para cumplir con los objetivos del negocio dueña o generadora de los datos. En esta fase no sigue una misma técnica o visión en todos los proyectos de CRISP-DM dependerá de las necesidades de la empresa o negocio generador de los datos, por lo que se debe evaluar cuales son las técnicas o algoritmos que mejor se adapten con el correcto cumplimiento y desarrollo de

los objetivos planteados en el proyecto así que es probable que sea necesario que se deba volver a etapas anteriores ya mejorar características del dataset.

Según, (Niño, 2016) ,En esta fase se seleccionan y aplican diferentes técnicas (algoritmos) de modelado, calibrando sus parámetros para conseguir sus valores óptimos. Para un mismo problema de minería de datos tenemos diferentes técnicas susceptibles de ser usadas y, dado que cada una de ellas puede tener requisitos diferentes en la forma en que deben presentarse los datos de entrada, es probable que sea necesario realizar ciclos adicionales de "preparación de los datos".

Aunque esta fase no tenga pasos claros a seguir para ser desarrollada es importante mencionar la existencia de las diferentes tareas a realizar en este en esta fase de modelado de datos que son necesarias para desarrollar algoritmos, aplicar modelos adecuados y realizar las conclusiones correctas a estas. Por lo que a continuación se mencionan algunas de las tareas con las cuales es posible desarrollar un exitoso modelado de datos.

Selección de la técnica de modelado

En esta tarea se investiga y elige la técnica más adecuada para llevar a cabo la fase de modelado de datos en el dataset del proyecto que se desarrolla si bien en fases anteriores se debe investigar y hacer alusión a las posibles técnicas que se utilizara en el proyecto que se desarrolla en la fase de modelado es donde se debe investiga definir y documentar la técnica utilizada además de registrar los hallazgos encontrados al realizar dicha técnica.

Según, (Niño, 2016), Aunque ya desde el principio del proyecto, en la fase de comprensión del negocio, se realiza una selección preliminar del tipo de técnica a emplear, en este caso la tarea se centra en poner "nombre y apellidos" a la técnica, de entre las diferentes opciones de configuración, versionado, etc. que puede presentar. Además, hay que tener en cuenta que muchas técnicas de modelado funcionan bajo la premisa de unas asunciones específicas sobre los datos.

Aunque existan muchas técnicas de modelado de datos estas se pueden dividir en tres grandes grupos. Estos tres grupos son el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo.

- Aprendizaje supervisado: este tipo de algoritmo o técnica se basa en ofrecer "emparejados" ala algoritmo se le brinda tanto los datos de entrada como las salidas que debería obtener como resultado. Este tipo de técnica se realiza a base de ejemplos o datos con etiquetas por ejempló fotos con algún tipo de descripción como las cosas que aparecen en la imagen. Según, (Algoritmia, 2019)," Los algoritmos "aprenden" de datos que se le ofrecen emparejados (se le da al algoritmo tanto las entradas como las salidas que tendría que obtener)"
- Aprendizaje no supervisado: este tipo de técnica o algoritmo es capaz de aprender solamente a través de los datos de entrada esto quiere decir que no es necesario no entregarle al código con el que

se realiza el código los resultados que debería obtener ni asignar etiquetas a los datos.

- Según, (Algoritmia, 2019), Los algoritmos consiguen obtener conocimiento únicamente de los datos que se proporcionan como entrada. A diferencia del aprendizaje supervisado, no se dispone de datos etiquetados, y no se le enseña al sistema qué resultados o salidas queremos obtener (son desconocidos).
- Aprendizaje por refuerzo: este tipo de algoritmos aprende utilizado el método de retroalimentación, decir que aprende gracias a que se relaciona con el usuario cada acierto y error hace que algoritmo de respuestas más precisas realizando cada vez mejor las tareas que se les pide.
- Según, (Algoritmia, 2019), Los algoritmos aprenden y mejoran en su respuesta a partir de la experiencia, usando un proceso de retroalimentación. El sistema aprende del mundo que le rodea y de los errores que comete hasta que encuentra la mejor manera de realizar una tarea.

Como se investiga y documenta existen una gran cantidad de modelos que se pueden seccionar en los tres grandes grupos mencionados por lo que se debe elegir según las necesidades del proyecto.

Diseño de los test

Luego de elegir que tipos de técnicas o modelos se elegirán para el desarrollo de la fase de modelado de datos, es importante diseñar los estándares que medirán que tan exitoso y calidad ya sea en el algoritmo y las columnas de nuestro dataset que se usarán para realizar la etapa de modelado de datos.

Según, (Niño, 2016), Antes de ponernos a generar un modelo, debemos diseñar el procedimiento según el cual se va a medir la calidad y validez del modelo. Esto abarca la métrica concreta de error que se va a emplear, o la descripción del plan para entrenar y evaluar los modelos, incluyendo el diseño de la separación entre datos de entrenamiento, de testeo y de validación.

Con esta información se puede afirmar que la creación de los test es de suma importancia para un desarrollo correcto de cada uno de los métodos, técnicas y algoritmos por lo que es de suma importancia recalcar cada una de ellas.

Medir la calidad y validez del modelo: realizar test para medir la calidad de los modelos que utilizamos en la fase de modelados es de suma importancia pues es posible asegura que el modelo con el que se trabaja funcione bien con los datos de entrenamiento si no también que es capaz de hacer predicciones precisas con los datos nuevos.

Según (Codificandobits, 2022), La idea de construir un modelo de Machine Learning no es sólo que lo haga bien con los sets de entrenamiento y validación. La idea es que además lo haga bien con datos que nunca antes haya

visto, pues de esta manera veremos qué tan robusto es y cómo se comportará con nuevos datos. Esto se conoce como generalización.

Métrica de error: es importante decidir una manera clara y específica para medir el porcentaje de error esto con el objetivo de buscar alternativas para mejorar los métodos con los que se trabaja y mejorar el algoritmo o algoritmos con el que se decidido realizar la fase de modelado de datos. Para esto se utiliza una métrica llamada MSE.

Según ,(Shalldb, 2023), La interpretación del MSE puede aportar información valiosa sobre el rendimiento de un modelo. Un MSE más bajo indica un mejor ajuste, ya que significa que las predicciones del modelo se acercan más a los valores reales. Por otro lado, un MSE más alto indica que las predicciones del modelo están más alejadas de los valores reales, lo que indica un ajuste deficiente.

Construcción del modelo

La construcción del modelo tareas una de las partes más importantes de la fase de modelado en un proyecto de CRISP-DM durante el proceso machine learning. Durante esta fase construcción del algoritmo, se desempeñan varias tareas siendo la más importante el desarrollo y ejecución del algoritmo que se seleccionó previamente. Se selecciona un algoritmo adecuado basado en el problema que se debe resolver se debe ejecutar un dataset previamente preparado, permitiendo que algoritmo aprenda de los datos adecuadamente y que los resultados que arroga sean adecuados.

Esta tarea es una de las más importantes por lo tanto se debe documentar con detalle cada decisión y hallazgo que ocurra durante esta fase. Esto debe incluir cualquier ajuste realizado y los motivos por los cuales se realizó este ajuste ya sea al algoritmo o al mismo dataset.

También  es impórtate realizar una justificación clara de la elección de los modelos y columnas que se utilizaron en el dataset, de ser necesario documentar nuevas técnicas aplicadas para la preparación y limpieza de los datos

Además, es crucial proporcionar una explicación detallada del modelo resultante y el impacto de sus predicciones. Para garantizar la transparencia y confiabilidad del modelo, es fundamental evaluar y documentar su interpretabilidad, es decir, su facilidad de operación y las predicciones que realiza. Además, se debe identificar y discutir cualquier problema en la interpretación del modelo.

En resumen, crear un modelo implica desarrollar un algoritmo y proporcionar una justificación y documentación rigurosas para cada paso. Esto garantiza que el modelo producido sea robusto, interpretable y reproducible, y proporciona una imagen clara y comprensible de la información revelada a través del algoritmo utilizado en la fase de modelado.

Evaluación del modelo

Es la etapa final llamada evaluación del modelo es de igual importancia que la anteriores en el proceso de modelado de datos. Al realizar esta tarea el objetivo principal es evaluar la calidad del modelo o algoritmo que se desarrolló,

en síntesis se analiza qué tan preciso es el algoritmo que se diseñó y qué tan bien predice o acierta una condición específica, evaluándola en un rango determinado. Para esto, se emplean diversas métricas de evaluación que se documentan a continuación

- Accuracy: este tipo de métricas mide las predicciones que fueron correctas sobre el total de predicciones que se realizaron en el modelo. Este tipo de métrica es más eficiente y más utilizado cuando las clases que se evalúan se encuentran de manera balanceadas en los datos. Según, (The Machine Learners, 2023) ,"Es recomendable utilizar esta métrica en problemás en los que los datos están balanceados, es decir, que haya misma cantidad de valores de cada etiqueta (en este caso mismo número de 1s y 0s)."
- Predicción: este tipo de métricas es utilizada para saber qué porcentaje de valores se han calificado por el algoritmos pueden ser calificados como positivos son efectivamente positivos. Según, (The Machine Learners, 2023), "La métrica de precisión es utilizada para poder saber qué porcentaje de valores que se han clasificado como positivos son realmente positivos."
- Recall: esta métrica indica la proporción de positiva que fueron acertadas correctamente por el modelo. Esta métrica es útil cuando el objetivo en minimizar los datos falsos negativos. Según, (The Machine Learners, 2023),"La métrica de recall, también conocida

como el ratio de verdaderos positivos, es utilizada para saber cuántos valores positivos son correctamente clasificados.”

- F1-score: esta métrica es un punto medio o una unión de precisión y recall, proporciona un balance entre ambas métricas, esta métrica suele ser útil cuando hay clases desbalanceadas. Según, (The Machine Learners, 2023),” Esta es una métrica muy utilizada en problemás en los que el conjunto de datos a analizar está desbalanceado. Esta métrica combina el precisión y el recall, para obtener un valor mucho más objetivo.”
- Matriz de confusión: este tipo de métricas muestra una tabla de las cantidades de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Este tipo de métrica se vuelve útil para entender dónde está cometiendo errores el modelo que se está utilizando.
- Según (Madrigal, 2023), La matriz de confusión es una herramienta comúnmente utilizada en el aprendizaje automático para evaluar los modelos de clasificación. Se trata de una tabla que muestra la cantidad de veces que el modelo clasificó correcta incorrectamente una muestra en relación con su clase real.

  La tabla está compuesta por cuatro elementos principales: verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN).

Importancia del modelado de datos

Como se puede observar la fase de modelado en el modelo CRISP-DM es fundamental para la resolución de los problemás u objetivos planteados al principio del proyecto, pues en esta fase de modelado de datos es donde el científico d datos en este caso los estudiantes del curso de Data Science se ponen aplican los conocimientos aprendidos en clase y en investigaciones para la construcción de modelos para de esta manera poder lograr los objetivos de este según proyecto además de poder retomar objetivos pasados del primer proyecto por lo que esta fase se vuelve crucial para el cumplimiento de los objetivos y encontrar una solución al problema. De igual manera que se vuelve crucial para los administradores y científicos de datos entender los resultados de los algoritmos con el fin de administrar de manera más eficiente el negocio

La importancia de la fase de modelado se basa en su capacidad de transformar datos sin procesar en información valiosa y procesable. Los modelos creados durante esta etapa son cruciales para predecir resultados, identificar patrones y generar información que pueda usarse para tomar decisiones informadas.

Según, (Conectapyme, 2023), El modelado de datos es un proceso esencial en la creación de estructuras y diseños que permiten organizar y gestionar eficientemente la información de las organizaciones. En el ámbito empresarial, el modelado de datos se refiere a la representación visual y lógica de cómo se relacionan los datos entre sí, lo que proporciona un marco de referencia para entender y administrar de manera efectiva la información en una base de datos.

Luego de documentar la investigación para realizar la fase de datos de manera correcta y de la manera más informada posible se lleva a cabo el modelado de datos orientado a las tareas o sub fases que se mencionaron y explicaron anteriormente.

Selección de la técnica de modelado

Como se menciona anteriormente en esta fase se selecciona el tipo de técnica o modelado que se utilizara para llevar a cabo la fase modelado de datos. Por lo que a continuación se documenta el tipo de técnica de modelado que se utilizara en el proyecto.

Dado que las especificaciones del proyecto requieren el uso de modelado de datos no supervisada, se decide elegir algoritmos y técnica orientados a esta metodología. La principal ventaja de este tipo de técnica es que su ejecución no necesita intervención para realizar el modelado o entrenamiento de los datos esto garantiza que se obtendrán respuestas claras y precisas, libres de posibles errores causados por la intervención humana. Según, (IBM, 2024),"El aprendizaje no supervisado, también conocido como machine learning no supervisado, utiliza algoritmos de machine learning para analizar y agrupar conjuntos de datos no etiquetados. Estos algoritmos descubren patrones ocultos o agrupaciones de datos sin necesidad de intervención humana"

Además, el uso de modelos de datos no supervisados facilita la identificación de patrones complejos y relaciones entre variables que pueden pasar desapercibidas en los métodos supervisados.

Por lo que se reafirma que la técnica que se utilizará en la fase de modelado de datos será aprendizaje no supervisado como regresión logística, K-Means y matriz de correlación, por especificaciones del proyecto y las ventajas que este tiene.

Diseño de los test

Para evaluar el éxito de o presión con el que cada uno de los modelos clasifica o predice cierta características del dataset House-Price-Advanced en este caso que característica o caracterizas son las que hace que una casa suba de precio. Por lo que se procede a desarrollar un test con el que evaluaremos la calidad de los algoritmos con los que se desarrolla las fase de modelado de datos este test se tomara en cuenta cuando se realicen los algoritmos.

1- Nombramiento del modelo y título de la grafica

   Se nombra el tipo de modelo con su categoría además de nombrar el título de la gráfica.

2- Descripción de las columnas usadas en el modelo

   Se documenta el nombre de o las columnas que se utiliza para llevar a cabo el modelo

3- Definición del objetivo y justificación que se planea alcanzar con el algoritmo. Se plantea el objetivo que se quiere cumplir con los resultados que arroje el algoritmo, se justifica el uso del modelo

4- Selección de métricas

Se selecciona y describe las métricas con las que se mida el éxito, además de documentar el porcentaje de éxito del modelo

5- Evaluación y análisis del modelo

Se documentará los hallazgos encontrados al analizar los resultados encontrados en los algoritmos que se realicen para cumplir con los objetivos plateado en el proyecto

Este test sea de ayuda para documentar los resultados de los algoritmos y otra información importante en etapas posteriores por lo que será fundamental para el desarrollo para la fase de modelado de datos.

Construcción del modelo

En esta fase de documentar la construcción de los algoritmos que se mencionaron en la etapa de selección de la técnica de modelado, se planea documentar los tipos de bibliotecas además de justificar el uso del algoritmo que se construyó y con qué fin o que información de desea revelar. Por lo que a continuación se documenta los hallazgos encontrados en la construcción en los modelos de matriz correlacional, K-Means, agrupación jerárquica y regresión logística y lineal

Matriz Correlacional

Una matriz correlacional se puede conceptualizar como la creación de una tabla donde se muestran los coeficientes de correlación entre las columnas de un dataset en este caso las columnas de dataset House-Price-Advanced por lo que el

objetivo principal de la construcción de este algoritmo será descubrir que columnas o características están más relacionadas con la columna SalePrice. Funciona de manera que se grafica en forma de “cuadricula” que se pinta de colores depende en valor de correlación entre las columnas por lo que se vuelve sencillo examinar los resultados. Es posible realizar este tipo de modelo con tipos de variables numéricas como números entero o decimales “int,float” aunque es posible hacerlo con variables tipo objetos aunque su visualización es algo más complicada. Aun así se opta por la matriz orientada a tipos de variables numéricas.

Matriz Correlacional en análisis de variables numéricas

Para construir este tipo de algoritmos se necesitan varias bibliotecas en el lenguaje Python se necesitan las siguientes bibliotecas para crear una matriz de correlación;

- Numpy: se utiliza principalmente para cálculos numérico, también es utilizada para dar soporte a la creación de algoritmos de matrices multidimensionales (arreglos) y matemáticas avanzadas.
- Pandas: es una biblioteca altamente utilizada en el análisis de y manipulación de datos.

- Matplotlib.pyplot: principalmente se utiliza para crear visualización o gráficas, permite crear gráficos de líneas, barras, dispersión, histogramas y más.
- Seaborn: es una biblioteca similar a la biblioteca matplotlib.pyplot pero proporciona una interfaz más sencilla para crear gráficos se utiliza para explorar relaciones distribuciones de los datos del dataset.

El procedimiento para crear este tipo de matriz es la siguiente. Luego de importar las bibliotecas se utiliza la biblioteca numpy para seleccionar las variables de tipo numéricas con el comando “incluye[np.number]”.

A continuación con el método “.corr()” calcula la matriz de correlación para todas las columnas donde los valores de correlación varían entre -1 y 1 donde 1 representa la correlación perfecta -1 indica una correlación negativa perfecta y 0 indica que no hay correlación entre las columnas que se intentan relacionar entre sí.

Después se utiliza las bibliotecas matplotlib.pyplot y seaborn para graficar la matriz con estas bibliotecas se calcula la forma y las medidas.

K-Means

El método no supervisado K-Means se define como un método de agrupación que divide un conjunto de datos en grupos basándose en características que compartan entre sí quiere decir que agrupa los datos sin

necesidad de etiquetas. Este método se desarrolló con el fin de visualizar el comportamiento de los datos de ciertas columnas seleccionarlas sistemáticamente gracias al análisis de la matriz que se explica más adelante. este método funciona a través de diferentes fases o pasos.

En la primeras van de la mano, se selecciona la cantidad de centroides esto se realiza a través de un código para evitar que se escoja una cantidad de centroides excesiva o escasa lo que podría no arrojar datos correctos, esta fase se concreta de la siguiente manera; con la biblioteca "sklearn.cluster KMeans" para normaliza los datos de la copia previamente hecha para no dañar los datos de del dataset original con "preprocessing.normalize", después con el método KMeans se crea un modelo para luego grafiar el modelo con el método KElbowVisualizer y show después se entrena en modelo con el método .fit esto para realiza la técnica del método codo que funciona para saber la cantidad adecuada de clústeres para el modelo de K-Means.

En la tercera se actualizan los centroides para mejorar la precisión del modelo este es gracias al método ".fit" mencionado anterior mente luego en la cuarta fase con el comando "visualizer.fit(x1)" se repite el proceso o el entrenamiento de los datos hasta que alcance un criterio de convergencia o bien que los centroides de los clústeres no se muevan.

En la cuarta se fase se grafica el modelo con las bibliotecas plotly.express y pandas se hace una gráfica de clústeres en formato 3D, con el método px.scatter_3d se decide los valores en este caso columnas que se ubicaran en X,

Y, y Z , también el color y las dimensiones de la gráfica. Por último se utiliza ".show()" para mostrar la gráfica.

A continuación se documentan las bibliotecas utilizadas para desarrollar este modelo

- Pandas: es una biblioteca altamente utilizada en el análisis de y manipulación de datos.
- matplotlib.pyplot: esta es una biblioteca de trazado estándar en Python que proporciona funciones para crear visualizaciones y gráficos estáticos. Es altamente personalizable y permite crear una amplia variedad de gráficos como histogramás, diagramás de dispersión, líneas, barras, entre otros.
- seaborn: Seaborn es una biblioteca de visualización de datos basada en matplotlib. Proporciona una interfaz de alto nivel para crear gráficos estadísticos atractivos e informativos. Se utiliza comúnmente para visualizar fácilmente relaciones estadísticas complejas.
- plotly: Plotly es una biblioteca de visualización interactiva que le permite crear gráficos interactivos y dinámicos. Le permite crear gráficos complejos, como cuadros 3D, mapas y visualizaciones de datos interactivas. pio proporciona herramientas para administrar archivos y configuraciones de gráficos de Plotly.

- plotly.express): Esta es una interfaz de alto nivel para Plotly que facilita la creación rápida de gráficos complejos. Le permite crear visualizaciones interactivas con una sintaxis simple y eficiente
- os: Es un módulo que proporciona funciones para interactuar con el sistema operativo subyacente. Se utiliza para manipular rutas de archivos, administrar directorios y otras operaciones relacionadas con el sistema de archivos.
- warnings: este módulo controla las advertencias emitidas por Python. warns.filterwarnings("ignore") se utiliza aquí para ignorar las advertencias que pueden aparecer durante la ejecución del código, lo que puede resultar útil para mantener limpios los resultados de salida.
- sklearn.cluster.KMeans: este es un algoritmo de agrupación no supervisado que agrupa datos en grupos según las similitudes entre las observaciones. Busca dividir los datos en grupos homogéneos, maximizando la similitud dentro de cada grupo y minimizando la similitud entre grupos.
- sklearn.preprocessing: proporciona funciones para estandarizar, normalizar y transformar datos antes de aplicar algoritmos de aprendizaje automático. Ayuda a mejorar el rendimiento y la eficiencia del modelo.
- yellowbrick.cluster.KElbowVisualizer: Yellowbrick es una biblioteca para visualización de aprendizaje automático en Python.

KElbowVisualizer es una herramienta específica de Yellowbrick que le permite determinar el número óptimo de conglomerados (k) para el algoritmo KMeans visualizando la inercia (suma de cuadrados de distancias dentro del conglomerado) frente al número de conglomerados.

Agrpacion Jerarquica

La agrupación jerárquica es un método de agrupación que crea una jerarquía de grupos. Puede ser aglomeraría (de abajo hacia arriba) o una división (de arriba hacia abajo). Este modelo se crea con el mismo objetivo que le modelo K-means, visualizar el comportamiento de los datos de ciertas columnas pero también como se podían dividir las jerarquías de los datos según las columnas, así que se explicara el proceso de la creación de este algoritmo.

Como primer paso se importará bibliotecas necesarias para realizar el modelo:

- numpy, pandas, visualización
- matplotlib.pyplot, agrupación jerárquica
- scipy.cluster.hierarchy para dendrogram y linkage preprocesamiento
- StandardScaler de sklearn.preprocessing y

- AgglomerativeClustering de sklearn.cluster.

preprocesamiento

Después con la biblioteca StandardScaler se estandarizan las columnas que se analizara en el modelo para después con el método scaler.fit_transform calcular la desviación estándar de cada columna con la que se trabaja.

Por consiguiente se utiliza linkage del módulo scipy.cluster.hierarchy con el que se realizara un enlace jerárquico utilizando los datos que anteriormente se estandarizaron con “scaled”, también se utiliza el método “Ward” para minimiza la varianza de los clústeres fusionados (Ward se refiere a método de enlace (linkage method) utilizado para calcular la distancia entre clústeres durante el proceso de agrupación)

Por último con las bibliotecas matplotlib.pyplot y dendrogram se utilizan para graficar el modelo de agrupación jerárquica fijando las dimensiones de la gráfica los respectivos títulos y ejes además de dibujas el damerograma utilizando los resultados del enlace jerárquico Z.

Regresión Logística y Lineal

Aunque ambos modelos no entran en la categoría de modelo no supervisado es imposible admitir que estos métodos no sean de interés o que no proporciones información relevante por lo que se realizaron estos modelos con el objetivo de encontrar algún tipo de información relevante

sobre la correlación o indicados de las características relevantes de las casa que tuvieran precios altos

Regresión logística

En primera instancia se importan las bibliotecas necesarias para realizar el modelo

- pandas: Para manipulación y análisis de datos.
- numpy: Para operaciones numéricas.
- matplotlib.pyplot: Para visualización de datos.
- LogisticRegression: Clase de regresión logística de scikit-learn.
- train_test_split: Para dividir los datos en conjuntos de entrenamiento y prueba.

Después se crea una copia del dataset que se utilizara para realizar el modelo esto con el objetivo de no cambiar los datos originales esto con el método ".copy()".

Luego se toma el valor de la mediana de la columnas SalePrice como índice que decidirá si una cas tiene un costo bajo o alto con el método .median()

Seguidamente con la biblioteca matplotlib.pyplot se escribe el código necesario para graficar el modelo de regresión logística, donde se deciden valores como las dimensiones de la gráfica, colores de los datos,

las columnas que representaran a los ejes X y también de los títulos que llevara la columna, por último plt.axhline dibuja una línea horizontal en la mediana del precio de venta.

Por consiguiente se preparan los datos para la representación logística donde x será refrentada por la columna OverallCond y el precio es te caso la mediad por la columna HighPrice creada a partir del valor de la mediana de la columna SalePrice.

Por último de dividen y entrenan los datos con porcentajes de ochenta por ciento de entrenamiento y veinte por ciento de prueba esto con la biblioteca “sklearn.model_ train_test_split”. Por otra parte se entrena el modelo con los siguiente modelos:

- LogisticRegression(): Inicializa el modelo de regresión logística.
- model.fit(): entrena el modelo con los datos de entrenamiento.
- model.predict(): Predice etiquetas para datos de prueba.
- precision_score(): calcula la precisión del modelo.
- confusion_matrix(): calcula el valor de la matriz de confusión para evaluar el rendimiento del modelo.

Por último se visualizan los resultados del modelo como accuracy, y matriz de confusión

Regresión lineal

En primera instancia se importan las bibliotecas necesarias para hacer el modelo.

- pandas, numpy, matplotlib, seaborn: Bibliotecas estándar para manipulación de datos y visualización.
- LinearRegression: Clase de regresión lineal de scikit-learn.
- train_test_split: Para dividir los datos en conjuntos de entrenamiento y prueba.

Se crea una copia del dataset con el método ".copy()" con la se trabajará con el objetivo de no dañar los datos originales. Por otra parte con el método pd.cut() convierte la variable con la que se busca encontrar una correlación BedroomAbvGr con SalePrice en categorías binarias, por consiguiente con el método pd.get_dummies() convierte las variables categóricas en datos dummy.

Después del procesamiento de los datos se preparan los datos para realizar la regresión lineal, donde el eje X será representado por las categorías de las columna BedroomAbvGr y el eje Y será representado por la columna SalePrice.

El siguiente paso es el entrenamiento de los datos con los métodos o comandos LinearRegression(): que inicializa un modelo de regresión lineal. y con model.fit(): que entrena el modelo con los datos de

entrenamiento. También se realiza una fase de predicción y evaluación del modelo, donde se utilizó el comando model.predict() para predecir los precios de veta de las casa con los datos de prueba y model.score para calcular el coeficiente de determinación (R^2) del modelo.

Por último se utilizan la biblioteca matplotlib para graficar el modelo de regresión lineal donde se deciden valores como la dimisión de la gráfica, los ejes la línea de regresión y la colocación de títulos tanto a las gráficas como los ejes.

Evaluación

Como se documentó esta etapa o una de las tareas a realizar es la evaluación de los modelos, esta evaluación se basa en el nivel de predicción con la que cuenta cada uno de los modelos que se construyó y en información relevante para alcanzar los objetivos propuestos en el proyecto.

Por lo que se utilizara el conjunto de test de prueba que se creó previamente en etapas anteriores con el que se planea evaluar el nivel de precisión de cada uno de los modelos creados y documentar información relevantes que se encontró al analizar los algoritmos y las gráficas que se crearon a partir de la etapa o tarea de construcción de los modelos.

Por lo que se procede a evaluar cada uno de los algoritmos que se construyeron donde se describirá el objetivo de cada proyecto, se justificará el porqué de uso en el proyecto y como el algoritmo ayudara al

cumplimiento de los objetivos planteados en el proyecto, además de evaluar el porcentaje de éxito que obtuvo el modelo con la métrica adecuada de los modelos creados también documentar información importante para el cumplimiento de los proyectos encontrada al analizar cada uno de los modelos, por último describir las columnas utilizadas en cada modelo.

1- Nombramiento del modelo y título de la grafica

   Matriz de correlación modelo no supervisado

   Nombre de la gráfica: Matriz de Correlación en Análisis de Variables

   Numéricas

2- Descripción de las columnas usadas en el modelo

   Para modelado de datos de matriz de correlación se opta por seleccionar las variables tipo numéricas que se documentan a continuación:

   Id SalePrice MSSubClass LotFrontage LotArea

   OverallQual OverallCond YearBuilt

   YearRemodAdd MásVnrArea BsmtFinSF1

   BsmtFinSF2 BsmtUnfSF TotalBsmtSF 1stFlrSF 2ndFlrSF LowQualFinSF

   GrLivArea BsmtFullBath BsmtHalfBath FullBath

   HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces

   GarageYrBlt GarageCars GarageArea WoodDeckSF

   OpenPorchSF EnclosedPorch 3SsnPorch ScreenPorch PoolArea

   MiscVal MoSold YrSold

3- Definición del objetivo y justificación que se planea alcanzar con el algoritmo. El objetivo del uso de una matriz correlacional es visualizar las correlaciones que existen entre las columnas

específicamente la correlación de las columnas con la columna SalePrice

Se justifica el uso de este tipo de modelo de matriz de correlación por su fácil analisis además de que la información obtenida de este proyecto se presume que pueda ser de importancia para el cumplimiento del proyecto

4- Selección de métricas

El tipo de métrica que se utilizó para este modelo recibe el nombre de Correlación de Pearson, este tipo de métricas mide la correlación que tiene las columnas. Según (Probabilidad y Estadistica, 2021)“El coeficiente de correlación de Pearson, también llamado coeficiente de correlación lineal o simplemente coeficiente de correlación, es una medida estadística que indica la relación entre dos variables.” . donde valores menores a cinco representa un coeficiente de correlación bajo, números entre cinco y diez represa un coeficiente de correlación aceptable y un coeficiente de correlación mayor o igual a diez se percibe como alto y preocupante.

A continuación se documentan los resultados de la métrica Correlación de

Pearson

| | | | | | |
|---|---|---|---|---|---|
| 0 | SalePrice | 5 | 14 | 1stFlrSF | 4 |
| 1 | Id | 2 | 15 | 2ndFlrSF | 4 |
| 2 | MSSubClass | 1 | 16 | LowQualFinSF | 5 |
| 3 | LotFrontage | 1 | 17 | GrLivArea | 6 |
| 4 | LotArea | 7 | 18 | BsmtFullBath | 1 |
| 5 | OverallQual | 3 | 19 | BsmtHalfBath | 2 |
| 6 | OverallCond | 1 | 20 | FullBath | 2 |
| 7 | YearBuilt | 4 | 21 | HalfBath | 1 |
| 8 | YearRemodAdd | 2 | 22 | BedroomAbvGr | 2 |
| 9 | MasVnrArea | 1 | 23 | KitchenAbvGr | 1 |
| 10 | BsmtFinSF1 | 1 | 24 | TotRmsAbvGrd | 4 |
| 11 | BsmtFinSF2 | 1 | 25 | Fireplaces | 1 |
| 12 | BsmtUnfSF | 1 | 26 | GarageYrBlt | 3 |
| 13 | TotalBsmtSF | 1 | 27 | GarageCars | 4 |
| 28 | GarageArea | 4 | | | |
| 29 | WoodDeckSF | 1 | | | |
| 30 | OpenPorchSF | 1 | | | |
| 31 | EnclosedPorch | 1 | | | |
| 32 | 3SsnPorch | 1 | | | |
| 33 | ScreenPorch | 1 | | | |
| 34 | PoolArea | 2 | | | |
| 35 | MiscVal | 1 | | | |
| 36 | MoSold | 9 | | | |
| 37 | YrSold | 2 | | | |

5- Evaluación y análisis del modelo

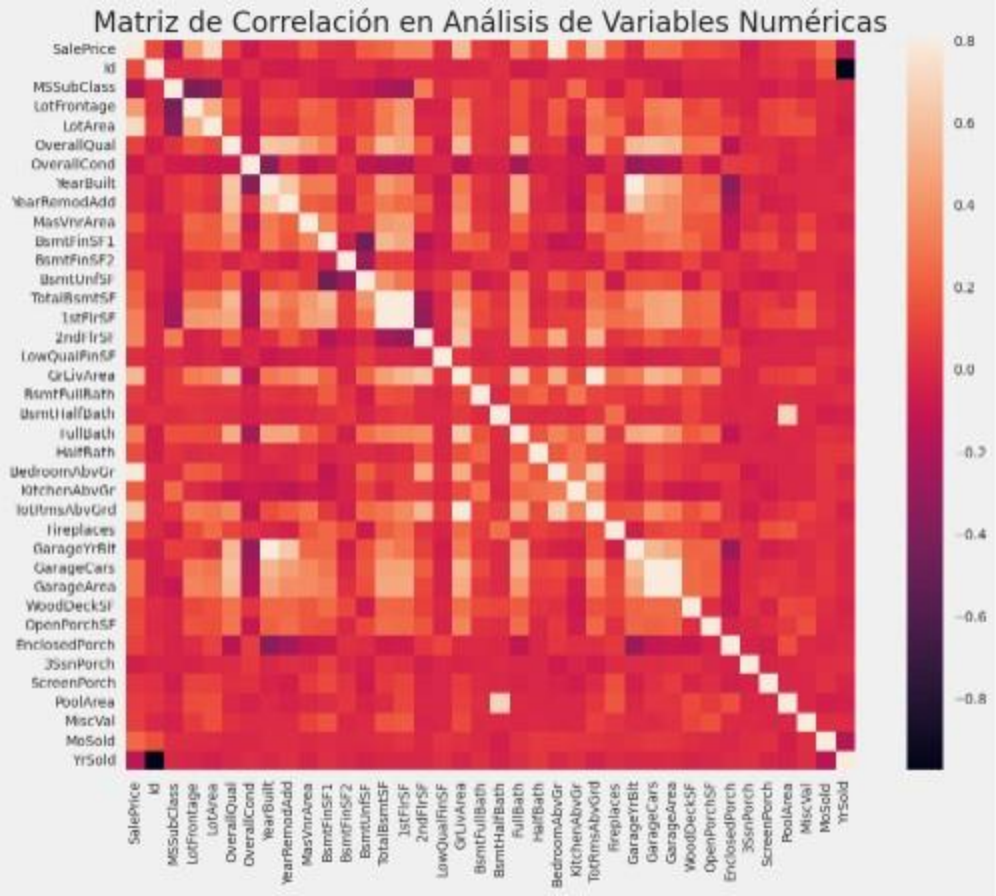
Como se puede observar los valores VIF más acertados son los de las columnas SalePrice, LotArea LowQualFinSF, GrLivArea y MoSold con lo que se puede deducir que estas columnas son las que más se correlacionan con el dataset.

Aun así para cumplir con el objetivo del modelo y los objetivos de los proyectos se debe examinar las columnas con mayor correlación con la columna SalePrice.

Esto se puede visualizar en la gráfica donde se puede observar que las columnas que comparte una mayor relación con la columna SalePrice son LotArea con 0.710941, 1stFlrSF con 0.321629, 2ndFlrSF con 0.337798, GrLivArea con 0.558863, BedroomAbvGr con 0.786504, TotRmsAbvGrd con 0.622890, GarageCars con 0.263258 y GarageArea con 0.254372.

Este tipo de métrica es diferente llamada coeficiente de correlación que indica que una que este más relacionada con números cercanos a uno en cambio entre menos relacionada este una columna con otra el coeficiente de correlación tiende a ser cero o números cercanos a cero

Con lo que se puede afirmar que las casas que esta representadas a travas de los datos de este data House-Price-Advances tuvieron un mejor precio si tenían buenas características; una cantidad de pies cuadrados del área del lote aceptable, tener primera y segunda planta con cantidad de pies cuadrados aceptable, una cantidad de habitaciones y por último que la casa tenga una zona de garaje (Ver Anexo 17)



1. Nombramiento del modelo y título de la grafica

   Modelo no Supervisado K-Means

   Nombre del grafico: Modelo No Supervisado K-Means Análisis del Comportamiento de los Datos

   Descripción de las columnas usadas en el modelo

   Para el modelo no supervisado K-Means se utilizan las columnas: BedroomAbvGr,

   SalePrice, MSSubClass, GrLivArea y LotArea

2. Definición del objetivo y justificación que se planea alcanzar con el algoritmo. Como objetivo principal del uso del modelo es observar el comportamiento de los datos de las columnas previamente seleccionadas sistemáticamente gracias al análisis del modelo de matriz de correlación

   Se justifica el uso de este modelo por su fácil análisis de manera gráfica y cumplir con las normas establecidas en el proyecto sobre el uso de modelos no supervisados

3. Selección de métricas

   Los modelos de K-Means no suelen ser medidos con métricas de éxito para observar su precisión pues K-mean se suele utilizar para visualizar la distribución de los datos. Aunque se utilizó el método de codo o Elbow Method para determinar la cantidad de centroides necesarios para desarrollar el método correctamente  se entrenó el modelo

para reajustar los centroides en una ubicación adecuada con ".preprocessing.normalize" para normalizar los datos y ".fit" para que los datos se entrenaran revelando que la cantidad de centroides y clústeres que necesitaba el modelo son tres.

4. Evaluación y análisis del modelo

Al construir el código y observar la gráfica se puede deducir cierta situaciones que de documentaran a continuación.

Al parecer las columnas BedroomAbvGr y SalePrice con las características que mayor tiene influencia en el agrupamiento de los datos, se presume que esto pase gracias a la correlación existente entre las columnas BedroomAbvGr y SalePrice,

Se puede observar en la gráfica que las columnas MSSubClass y BedroomAbvGr tiene una relación lineal con SalePice ya que los datos de casas con mayor número de cuartos tiendes a subir de precio, también las casas con una clase de construcción que oscila entre los
90 y 190 el precio de las casa suele ser más alto que las casa con modelos de construcción de 85 a 20 el precio no es tan elevado que los móldelos de concreción más altos.

Aunque no es totalmente asertivo afirmar que el modelo es lineal pues al analizar la gráfica es posible darse cuenta que no todos los datos siguen la misma tendencia explicada anteriormente. Por lo que se decide cambiar la columna MSSubClass por LotArea que es una columna que ya se comprobó que si tiene una rígida relación con la columna Sale Price.

Al visualizar nuevamente los datos surge un cambio de clústeres pues el código indica que ahora es necesario tener cuatro claustros para realizar el modelo K-means adecuadamente.

Otro cambio que se pudo observar es que la distribución del modelo es estrictamente lineal además de estar segmentados por el número de cuartos que tenían las casa. se presume que esto se debe a que en este segundo modelo se trabajó con columnas que tenían un alto coeficiente de relación entre SalePrice y o tener un coeficiente de relación tal alto entre las demás columnas del modelo es decir LotArea y BedroomAbvGr no tiene un coeficiente de relacional alto.

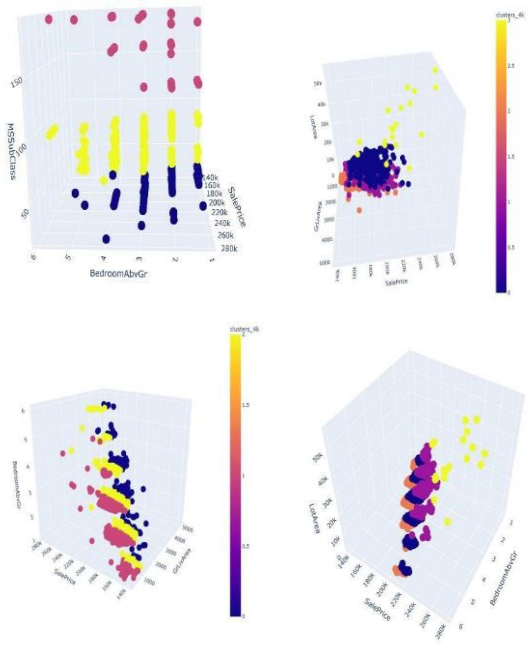
Por lo que se opta realizar un nuevo modelo donde esta vez se analicen las columnas GrLivArea, SalePrice, LotArea, donde se observaron los siguientes cambios.

Se sigue necesitando la misma cantidad de clústeres es decir cuatro clústeres por esta vez no existe otra columna más que SalePrice que secciones los datos, además los datos siguen una dispersión lineal donde mayor cantidad de cuartos o pies cuadrados de área vivible mayor sea el precio, se presume que este

cambio en el modelo se da gracias a que las columnas con las que se construyó nuevamente el modelo están relacionadas entre sí.

Algo similar sucede al realizar el modelo con las columnas GrLivArea, SalePrice, BedroomAbvGr solo que en este caso solo se necesita tres clústeres y las columnas BedroomAbvGr y SalePrice son las que segmentan la división.

Lo que se puede concluir con el análisis del modelo es que la columna SalePrice tiene un alto coeficiente de relación con las columnas, pero pocas columnas con las que se le relacionan tienen relación entre sí. También que se puede crear cierta categorías dependiendo de la distribución de los datos de casas relacionado al precio bajo, precio moderado, precio medio alto y precio alto. (Ver Anexo 18)




1- Nombramiento del modelo y título de la grafica

   Modelo supervisado Regresión Logística

Nombre del Grafico: Regresión Logística de SalePrice vs OverallCond

2- Descripción de las columnas usadas en el modelo

Las columnas que se utilizaron para realizar el modelo de regresión logista fueron OverallCond y SalePrice

3- Definición del objetivo y justificación que se planea alcanzar con el algoritmo.

Como objetivo principal del uso de este modelo es examinar otras columnas que puedan tener una posible relación con la columna SalePrice

Se justifica el uso de esta columna por su fácil visualización a través de las gráficas además de la información relevante que este pueda aportar con en cumplimiento de los objetivos del proyecto

4- Selección de métricas

Este modelo será evaluado con los métodos de matriz de confusión y accuracy

Los cuales se documentará los resultados a continuación

Exactitud del modelo (accuracy): 0.51

Es un valor aceptable pero no precisamente bueno ya que ronda apenas el cincuenta por ciento de exactitud o precisión del modelo

Matriz de confusión:

[[82  9]

 [79 11]]

Lo que significa que 11 casos el modelo predijo correctamente positiva, 9 casos el modelo predijo incorrectamente la clase positiva, 82 casos el modelo predijo correctamente la clase negativa y 79 casos modelo predijo incorrectamente la clase negativa.

5- Evaluación y análisis del modelo

Como se pudo observar el modelo sufrió varios fallos en la predicción de los datos, se presume que esto se debe a la poca relación que tiene esta columna con la columna SalePrice. También se puede observar que no existe alguna relación entre las columnas que se estudian pues existen cantidades similares de casas con la misma calificación en la columna OverallCond donde unas se encuentran en la categoría de precio bajo y otras de la misma categoría en precio alto.

Las casa que se vendieron a un mayor precio se encuentran en la categoría 5, lo mismo pasa con las que se vendieron en menor precio también existen varios valores a típicos en la gráfica donde suponen que casa con condición de general de las casa en 1 alcanza pecio elevado de 280000 dólares. Con lo que se puede concluir que la característica OverallCond o condición general no decide el precio de una casa, si no que existen otras características que influyen en la calidad general de la casa. (Ver Anexo 19)


1- Nombramiento del modelo y título de la grafica

Modelo no supervisado Agrupación jerárquica

Nombres de las gráficas: Dendrograma de Agrupación Jerárquica, Agrupación Jerárquica

2- Descripción de las columnas usadas en el modelo

Para este modelo se planea utilizar las columnas SalePrice y LotArea

3- Definición del objetivo y justificación que se planea alcanzar con el algoritmo. Como objetivo principal se tiene indagar en la relación que tiene las características del precio y el área del lote de una casa.

Se justifica el uso de este método por su visualización fácil de las jerarquías que surgen a partir de comprar dos variables o columnas en este caso SalePrice y LotArea

4- Selección de métricas

Para este modelo se utilizaron diferentes métricas que se documentaran y explicaran a continuación.

Silhouette Score: 0.44

Es un puntaje bajo pero aceptable ya que es más cercano a cero que a uno lo que nos da indicios que el modelo no pudo predecir correctamente los datos

Calinski-Harabasz Index: 1105.32

Es un puntaje considerablemente bueno tomando en cuenta los resultados de los otros modelos

Davies-Bouldin Index: 0.73

En un puntaje considerablemente bueno, este indica la separación y compacidad de las jerarquías o clústeres

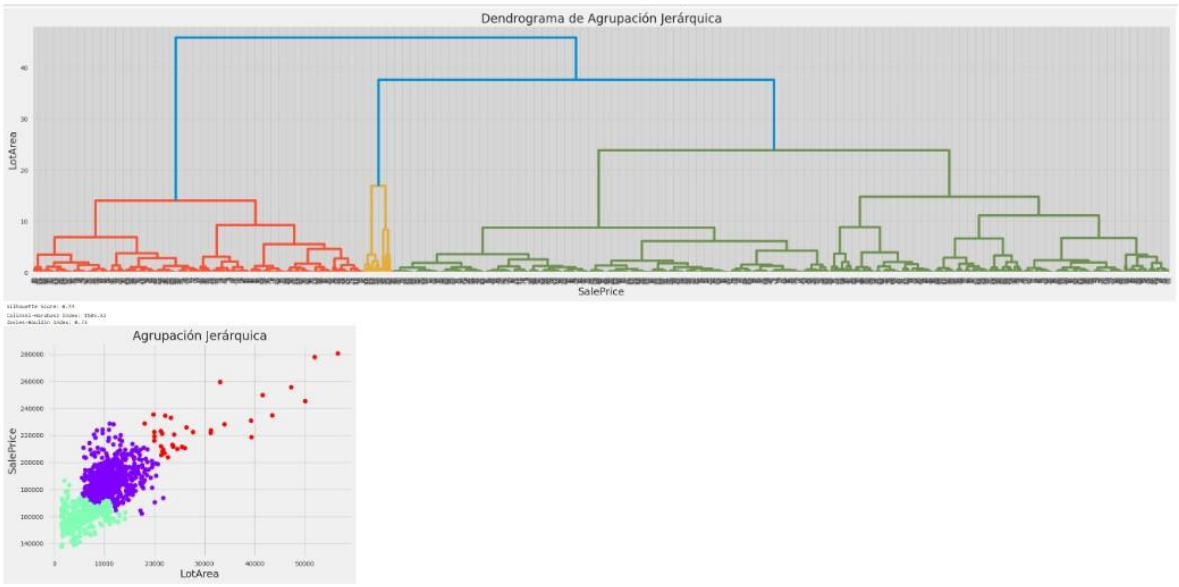
5- Evaluación y análisis del modelo

Como se puede observar en los resultados de las métricas existe una relación con SalePrice ya que valores como la dispersión, separación y compacidad de los clústeres es aceptable, lo que reafirma que existe una relación entre las variables SalePrice y LotArea.

Visualizando las grafica se puede ver que la comparación de ambas columnas SalePrice y LotArea genera una gran cantidad de jerarquías pero se pueden visualizar que existen cuatro jerarquías principales las mismas que se habían visualizado que el modelo K-Means al utilizar las columnas SalePrice, LotArea y GrLivArea. Otro otra parte en el gráfico de dispersión se puede ver una fuerte relación entre las dos columnas ya los datos tienden a tener una

dispersión lineal lo que reafirma la idea de que el precio de una casa está altamente influido por el área total del lote.

En loque se puede concluir que existe una relación entre el precio y el área del lote de una casa. (Ver Anexo 20)




1) Nombramiento del modelo y título de la gráfica.

   Modelo supervisado regresión lineal

   Nombre de la gráfica: Regresión Lineal: Precio de Venta Predicho vs Real por  Cantidad de Dormitorios

2) Descripción de las columnas las en el modelo

Para este modelo se utiliza las columnas BedroomAbvGr y SalePrice

3) Definición del objetivo y justificación que se planea alcanzar con el algoritmo.

Como objetivo principal se tiene indagar en la relación que tiene las características del precio y cantidad de dormitorios de una casa.

Se justifica el uso de este método por su visualización fácil de las que surgen a partir de comprar dos variables o columnas en este caso SalePrice y BedroomAbvGr

4) Selección de métricas

Para este modelo se opta por utilizar la métrica de coeficiente de determinación. A continuación se documenta y comenta los resultados.

R^2: 0.65

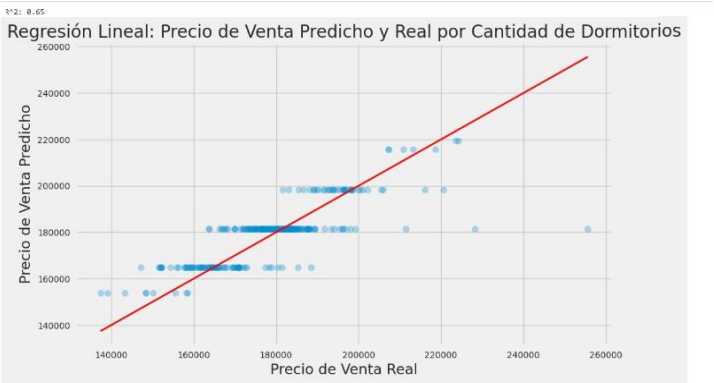
Es un resultado relativamente bueno ya que el modelo logra predecir el 65 % de los datos.

5) Evaluación y análisis del modelo

Como se puede observar en este modelo se puede apreciar la relación de las columnas BedroomAbvGr y SalePrice ya que existe una relación lineal que corresponde a una alza en el precio de las casa dependiendo la cantidad de cuartos que tenga la casa es decir a mayor número de cuartos el precio de la casa subirá.

Por otra parte el modelo logro predecir una gran cantidad de datos lo que refleja la relación de las columnas estudiadas en el modelo y un entrenamientos adecuado de los datos.

Se puede concluir que existe una relación entre las columnas de las BedroomAbvGr y SalePrice esto quiere decir que las casa que contengan una mayor cantidad de cuartos el precio tendera a subir (Ver Anexo 21)




1) Nombramiento del modelo y título de la grafica

   Modelo Supervisado Regresión Lineal

   Nombre de las Gráficas: BedroomAbvGr y SalePrice

   LotArea y SalePrice

2) Descripción de las columnas usadas en el modelo

   Las columnas utilizadas en este modelo fueron BedroomAbvGr, LotArea y SalePrice

3) Definición del objetivo y justificación que se planea alcanzar con el algoritmo.

   Como objetivo principal de este modelo será visualizar el comportamiento de los datos, además de buscar información relevante para cumplir con los objetivos del proyecto.

   Se justifica el uso de este modelo por su facilidad y versatilidad de uso además por la facilidad de interpretación de los datos.

   Se plantea el objetivo que se quiere cumplir con los resultados que arroje el algoritmo, se justifica el uso del modelo

4) Selección de métricas

Este modelo fue evaluado con la métrica de coeficiente de determinación. Este modelo se rige por la puntuación de 0 a 1 donde los valores más cercanos a uno significaran que el modelo es más eficiente y con mejor precisión.

A continuación se muestran los resultados y comentarios del modelo de Regresión Lineal

R^2: 0.95

El modelo predijo los datos de manera exitosa en su mayoría teniendo un porcentaje de fallo del 5% lo que demuestra que el modelo es eficiente.

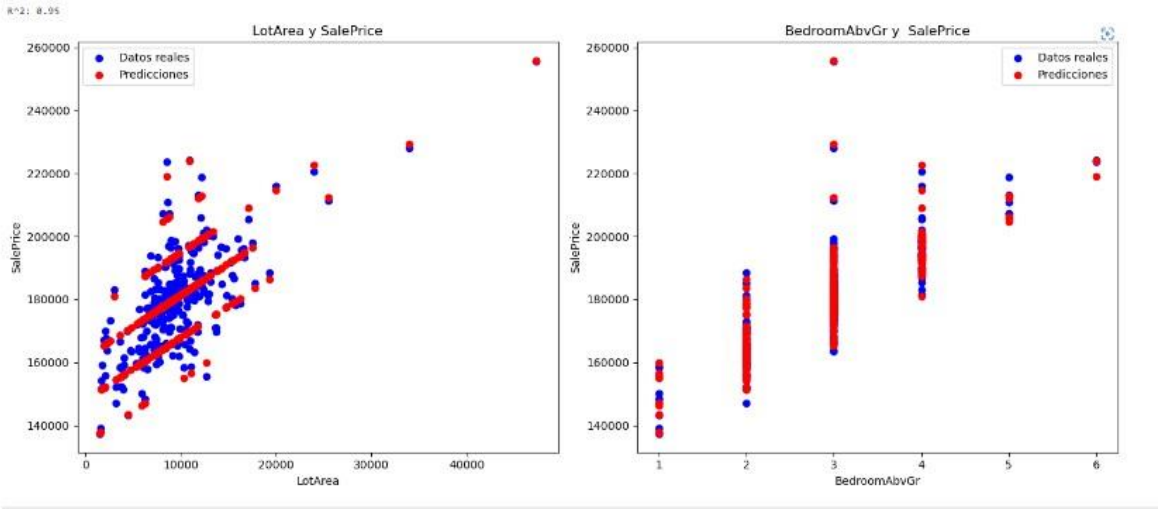
5) Evaluación y análisis del modelo

Se observa que ambas columnas tanto BedroomAbvGr como LotArea tiene una relación lineal con el precio de las casa, por lo que se puede afirmar que las características más influyentes a la hora de que un precisión de una propiedad o casa suban en la cantidad de pies cuadrados de área el lote y la cantidad de cuartos con la que cuente la casa.

Se logra observar que la columna BedroomAbvGr tiene una mayor correlación con la Columna SalePrice, esto se puede evidenciar en los resultados del modelo ya que las comparaciones de las columnas las columnas BedroomAbvGr y SalePrice tienden a tener una mayor relación lineal que las columnas LotArea y SalePrice esto también se logra apreciar en la matriz de correlación (LotArea 0.710941

BedroomAbvGr 0.786504).

Se observa que el modelo logra predecir de manera más eficiente los datos correspondientes a las columnas BedroomAbvGr y SalePrice lo que se presume que pueda suceder por su mayor correlación de igual manera LotArea y SalePrice tiene una relación lineal. (Ver Anexo 22)




1- Modelo y Título de la Gráfica:

Modelo: Regresión Lineal

Título de la Gráfica:

BedroomAbvGr y SalePrice, GrLivArea y SalePrice, OverallQual y SalePrice

2- Columnas Utilizadas en el Modelo:

Columnas independientes: BedroomAbvGr, GrLivArea, OverallQual

Columna dependiente: 'SalePrice'

3- Objetivo y Justificación del Modelo:

Objetivo: Predecir SalePrice basado en el número de habitaciones (BedroomAbvGr y el área habitable GrLivArea.

Observar la dispersión de los datos utilizando las columnas BedroomAbvGr, GrLivArea, OverallQual SalePrice

Justificación: Estas variables pueden influir en el precio de venta de una propiedad.

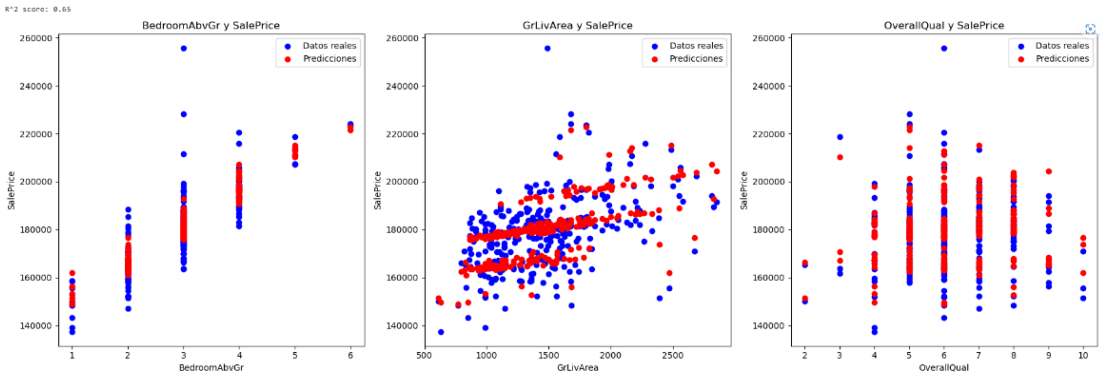
4- Métricas Seleccionadas:

Para este modelo se utilizó R^2 Coeficiente de Determinación: 0.65 en síntesis 65% de éxito

Evaluación y Análisis del Modelo:

El coeficiente de determinación (R^2) sugiere que el modelo explica aproximadamente el 65% de la variabilidad en 'SalePrice'.

Con esto se puede afirmar que las características más influyentes en la decisiones del precio de una casa en el negocio dueño de los datos del estado de Iowa es la cantidad de pies cuadrados del lote de la propiedad y la cantidad de cuartos con la que cuenta la casa. esto se puede comprobar por los resultados de este tiempo de modelos en comparación con el modelo BedroomAbvGr y SalePrice, LotArea y SalePrice . (Ver Anexo 23)



Método CRISP-DM

## Fase 5: Evaluación

Esta fase se basa en evaluar el resultado del análisis del modelos que se construyeron analizaron y valuados en la fase anterior de modelado de datos por lo que esta fase se tomará como una continuación de la sub fase evolución del fase 3 de modelado de datos por lo que lo se tomará los test evaluados en la fase anterior como referencia de esta fase de evaluación.

A continuación se desarrollarán las fases de evaluación de los resultados, revisión del proceso y Decisión sobre Siguientes Pasos

Evaluación de los Resultados

La tarea de evaluación de los resultados se basa en, medir el modelo y el grado en que el modelo cumple con los objetivos comerciales y determinar si existen razones relacionadas con el negocio para que el modelo falle. Esto es similar al paso de evaluación anterior, que se centró en la precisión y generalidad del modelo.

Según (El blog de Mikel Niño, 2016), Así como los pasos previos ligados a la evaluación se centraban en la precisión y la generalidad del modelo, en este caso la tarea se centraría en medir el grado en el que el modelo cumple los objetivos de negocio y detectar si hay alguna razón ligada al negocio por la que el modelo es deficiente. Se puede plantear también la evaluación del modelo dentro de su aplicación real, si el tiempo y presupuesto lo permiten.

Cada modelo implementado en el proyecto de minería de datos sobre el conjunto de datos House Price Advanced cumplió con los objetivos establecidos tanto en el proyecto

No1 orientado a la minería de datos como el proyecto No2 orientado al machine learning. Al evaluar y analizar los resultados de cada modelo, se puede determinar que cumple con los objetivos del proyecto y el propósito para el que fue utilizado.

Los coeficientes de correlación de la métrica Pearson el cual revelo niveles de correlación significativamente estables, se utilizó para revelar las relaciones cruciales entre las variables numéricas y los precios de venta por ejemplo las columnas LotArea, GrLivArea, BedroomAbvGr, 1stFlrSF, 2ndFlrSF y TotRmsAbvGrd. La matriz de correlación proporcionó información importante sobre las características determinantes de los precios de la vivienda.

Por otra parte, el modelo K-means ayuda a comprender mejor las preferencias del mercado inmobiliario al dividir los datos en grupos basados en similitudes y facilitar estrategias de marketing personalizadas. Y la distribución de los datos conforme a las columnas que se evaluaban en los modelo encontramos relaciones lineales entre algunas columnas como es el caso del modelo donde se evalúan las columnas BedroomAbvGr, SalePrice, LotArea

También la regresión lineal y logística demostró el comportamiento y relación de los datos al ser evaluadas con diferentes combinaciones de columnas encontrando casos en las que se observa una alta relación con las columnas como es el caso de las columnas BedroomAbvGr, SalePrice, LotArea donde el porcentaje de “éxito” fue alto alcanzado un 95% en la métrica R^2 .

También es el caso del modelo de Regresión Logística: Probabilidad de Precio Alto según la Cantidad de Dormitorio donde se evaluó el comportamiento de los datos y la

relación entre las columnas  BedroomAbvGr, SalePrice donde se encuentran resultados aceptables del 77% en la métrica de accuracy .

Por otra parte se encontraron casos en que los modelos no fueron eficientes a realizar el cálculo necesario, estos se relacionan con la falta de relación de las columnas que se evalúan como es el caso del modelo de Regresión Logística de SalePrice y OverallCond que tuvo un resultado de éxito del 51% en la métrica de accuracy

Revisión del proceso

Para determinar si hay factores importantes que se han pasado por alto y analizar aspectos de aseguramiento de la calidad de los modelos, se hace necesaria una revisión más exhaustiva de cuál ha sido el trabajo de minería de datos y los pasos seguidos.

Según (El blog de Mikel Niño, 2016), “Se debe realizar una revisión más exhaustiva de lo que ha sido el trabajo de minería de datos y los pasos seguidos (si han sido eficaces y eficientes, si admiten mejoras, si podrían haberse planteado con una aproximación diferente), para determinar si hay factores importantes que se han pasado por alto y analizar aspectos de aseguramiento de la calidad de los modelos.”.

Metodología CRISP-DM Preparación de datos para la base de datos de evaluación del valor de la vivienda En la fase de evaluación de resultados, el logro de los objetivos se demuestra de manera efectiva mediante la implementación cuidadosa de varias técnicas. Un ejemplo notable es la aplicación de cálculos de datos aleatorios a columnas complejas como 'MasVnrArea' y 'GarageYrBlt', donde los valores nulos se reemplazan de manera sólida y representativa. Esta técnica no sólo aseguró que los datos faltantes se manejaran

correctamente, sino que también ayudó a mantener la integridad y la realidad del conjunto de datos, haciéndolo adecuado para análisis posteriores.

Además, se utilizó la técnica de eliminar filas selectivamente en columnas como 'Baño completo', BedroomAbvGr' y ' KitchenAbvGr, donde los datos menos importantes nos permitieron aumentar la calidad de los datos. Análisis Este enfoque estratégico no sólo facilitó la preparación de datos para la fase de modelado, sino que también sentó las bases para generar conocimientos profundos relacionados con los valores de los activos en este caso los bienes inmuebles del estado de Iowa.

El uso efectivo y eficiente de estas técnicas no solo cumplió con los objetivos específicos del proyecto, sino que también logró la representación de patrones y relaciones ocultos dentro de conjuntos de datos, proporcionando así una base sólida para crear recomendaciones estratégicas dirigidas al sector inmobiliario. Estos hallazgos son fundamentales para optimizar la gestión de la información empresarial y facilitar decisiones comerciales informadas en un mercado dinámico como el inmobiliario de Iowa, donde la precisión y relevancia de los datos son fundamentales para el éxito a largo plazo.

Decisión sobre siguientes pasos

Con base en la evaluación de los resultados y la revisión del proceso, se deciden los siguientes pasos: pasar a la fase de implementación para operacionalizar el modelo, realizar una nueva iteración de la fase anterior, iniciar. Nuevos proyectos de minería de datos, etc.

Según (El blog de Mikel Niño, 2016), “Según las conclusiones de la evaluación de los resultados y de la revisión del proceso, se toma una decisión sobre los siguientes pasos a afrontar: pasar a la fase de despliegue para poner el modelo en operación, hacer nuevas iteraciones de las fases anteriores, iniciar nuevos proyectos de minería de datos, etc.”.

Decidir los próximos pasos en un proyecto de minería de datos y aprendizaje automático es esencial para garantizar su éxito continuo y la maximización del valor comercial.

Despliegue del modelo: si los modelos cumplen con los objetivos comerciales y técnicos, la implementación de producción pasa al uso operativo.

Iteración y mejora: Si se identifican áreas de mejora durante la evaluación de resultados y revisión del proceso, se recomiendan iteraciones adicionales de la fase de minería de datos para refinar los modelos.

Método CRISP-DM

## Fase 5: Despliegue

La creación del modelo y la aprobación de este no son la conclusión del proyecto. El proceso de minería de datos debe organizar y presentar el conocimiento adquirido de una manera que sea útil para las empresas. Esto no solo requiere que los modelos se integren en los procesos de toma de decisiones de la organización, sino que también requiere que los clientes participen en las etapas propias de implementación del modelo.

Según, (Niño, 2016) , La creación del modelo y su evaluación positiva no significa el final del proyecto. Se debe organizar el conocimiento adquirido gracias al proceso de minería de datos y presentarlo de una manera que sea utilizable en el contexto de negocio. Esto implica la integración de los modelos dentro de los procesos de toma de decisiones de la organización, además de requerir la implicación del cliente en los propios pasos de puesta en operación del modelo.

Esta estrategia se realiza en un escenario hipotético en el caso que existieran los equipos de personal y dispositivos para implementarlo

Estrategia de Despliegue

1. Planificación del Despliegue

a. Identificación de Objetivos y Requisitos

Objetivos: Asegurar que los modelos cumplen con los objetivos comerciales establecidos y proporcionan valor al negocio.

Requisitos: para llevar a cabo la fase de despliegue siendo guiado por esta estrategia se necesita infraestructura y persona. En el área de personal se necesita como mínimo los equipos de, Equipo de Desarrollo de Software, Equipo de Data Science, Equipo de Soporte Técnico, Equipo de Gestión de Proyectos, Equipo de Seguridad.

b. Asignación de Responsabilidades

- Equipo de desarrollo de software: responsable de configurar el entorno de producción e integrar el modelo con los sistemas existentes.
- Equipo de Data Science: Responsable de ajustar y refinar el modelo en base a los resultados de las pruebas de validación.
- Equipo de soporte técnico: Responsable del soporte continuo y la resolución de problemas relacionados con los modelos en producción.
- Equipo de Gestión del Proyecto: Coordinará el plan de despliegue y garantizará el cumplimiento del cronograma establecido.
- Equipo de Seguridad: Garantiza que el modelo y su entorno cumplan con los estándares de seguridad establecidos.

c. Cronograma de Despliegue

- Configurar el entorno de producción: esta fase puede tardar aproximadamente 2 semanas. Esto contemplando preparar el entorno, instalar el software necesario y configurar la infraestructura.

- Pruebas de validación: Esto puede tardar de 2 a 4 semanas dependiendo de la cantidad y complejidad de las pruebas requeridas. Incluye pruebas unitarias, de integración, de sistemas y pruebas de resiliencia y recuperación ante desastres.
- Implementación: la implementación en sí puede demorar entre 1 y 2 semanas, dependiendo de la complejidad del proceso de integración e implementación.
- Monitoreo: Esta fase es continua, pero para establecer una estructura inicial y asegurar su efectividad, se pueden dedicar de 3 a 5 semanas de monitoreo para verificar que el sistema está funcionando correctamente, de no ser así se debe volver a etapas anteriores para encontrar una solución a la problemática.

2. Planificación de la Monitorización y Mantenimiento

a. Establecimiento de Métricas de Monitoreo

- Métricas clave: definir la precisión del modelo y las métricas de rendimiento, así como los indicadores clave de rendimiento (KPI) para monitorear su precisión.
- Equipo de Monitoreo: Seleccione el equipo y la plataforma que se utilizará para el monitoreo continuo del modelo en producción.

b. Procedimientos de Mantenimiento

- Mantenimiento preventivo: Se establece un programa de mantenimiento regular para garantizar que el modelo funcione de manera adecuada y eficiente.
- Esto incluye actualizar periódicamente el modelo con datos disponibles recientemente, revisar la infraestructura de procesamiento y almacenamiento de datos para garantizar la escalabilidad y optimizar continuamente el modelo en función de las tendencias observadas en el mercado inmobiliario.
- Mantenimiento correctivo: se define procedimientos para la resolución de problemas y el manejo de fallas del modelo.
- Esto engloba las tareas de identificar rápidamente problemas a través de alertas automatizadas, explore las causas subyacentes utilizando análisis de datos y técnicas interpretativas de aprendizaje automático, e implemente soluciones correctivas que pueden variar desde el ajuste de los hiperparámetros del modelo hasta el reentrenamiento con datos actualizados.

3. Informe Final del Proyecto

Resumen del Proyecto

El presente informe final del proyecto resume los pasos realizados y los resultados obtenidos en dos proyectos relacionados con minería de datos y machine learning, utilizando el dataset "House Price Advanced" para la venta de bienes raíces en Iowa. El objetivo principal fue analizar relaciones y patrones en los datos para mejorar la precisión

en la predicción de precios de vivienda y proporcionar recomendaciones estratégicas para el mercado inmobiliario.

## Metodología Utilizada

Se empleó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) para estructurar el proyecto en las siguientes fases:

- Comprensión del Negocio
- Comprensión de los Datos
- Preparación de los Datos
- Modelado
- Evaluación
- Despliegue

## Documentación del Proyecto

Definición de objetivos del proyecto.

- Aplicar un correcto desarrollo de Minería de Datos en el Data Set House Prices Advanced  Desarrollar correctamente el método CRISP-DM.
- Implementar técnicas necesarias en la minería de datos del Data Set House Prices Advanced.
- Realizar una estadística descriptiva y sumatoria del Data Set House Prices Advanced.
- Detectar algún tipo de características relacionadas al precio de las casa del Data Set House_Prices_Advanced

- El objetivo de este proyecto es investigar y aplicar diversas técnicas de modelos no supervisados en el dataset House-Price-Advanced. A partir de los resultados obtenidos, se generarán recomendaciones constructivas para el negocio propietario de los datos, con el fin de mejorar la gestión eficiente de la información y optimizar la toma de decisiones en entornos cambiantes. Realizar los cambios necesarios relacionados con la preparación de los datos del dataset House-Price-Advanced para de esta manera mejorar la calidad de los datos y garantizar que los datos estén listos para la fase de modelado de datos.
- Evaluar diferentes algoritmos de modelos no supervisados en el dataset mencionado. El propósito es descubrir patrones ocultos y relaciones entre variables para obtener información relevante para el negocio dueño de los datos
- Generar recomendaciones concretas para el negocio propietario de los datos, basadas en los resultados del modelado de los datos del dataset House-Price-Advanced. Estas recomendaciones se centrarán en la optimización de decisiones y la gestión eficiente de la información.
- Analizar y determinar las características más valoradas por los compradores de viviendas en Iowa, evaluando variables como ubicación, tamaño, número de dormitorios y baños, y características adicionales. El objetivo es comprender los factores que influyen en las decisiones de compra para que agentes y promotores adapten sus ofertas, y contribuyan a una mejor planificación urbanística.

Descripción de los datos y su origen.

- Se trata de dos data set House Price Advanced y Sale Price que contiene las características de un grupo de casa y el precio de las casa respectivamente.
- El Dr. Samuel Saldaña Valenzuela proporcionará los datos sets House Price Advanced y Sale Price en el formato "cvs" para este proyecto desarrollado mediante el método de CRISP-DM. Estos datos sets se ubicarán en el repositorio de la plataforma GitHub del Dr. Samuel Saldaña Valenzuela.

Estrategia de preparación de datos.

Preparación de Datos

- Imputación de valores nulos en columnas con el método imputación aleatoria
- Eliminación de filas en columnas

Modelado y Resultados

- Para la fase de modelado se implementación de modelos de matriz de correlación, regresión lineal y logística, K-means y agrupación jerárquica.
- Para la evolución de modelos utilizando métricas como R^2,  accuracy, matriz de confusión, y person.

Evaluación de los Resultados

- La evaluación de los resultados se centró en medir el grado en que cada modelo cumple con los objetivos comerciales establecidos y determinar áreas de mejora:

- Regresión lineal: Se demostró una alta correlación entre las columnas BedroomAbvGr, SalePrice y LotArea, alcanzando el 95% en la métrica R^2. Esto indica que el modelo explica bien la variabilidad de los precios de la vivienda en función de las características seleccionadas.
- Regresión Logística: Lograr un 77% en la métrica de precisión al evaluar la probabilidad de valor alto según el número de dormitorios. Aunque fue exitoso, se identificó un área de mejora con una precisión del 51 % para el costo de ventas y el costo de ventas.
- K-Meaning: Facilita una mejor comprensión de las preferencias del mercado inmobiliario al agrupar datos en segmentos similares, apoyando así estrategias de marketing personalizadas.

Revisión del Proceso

Una revisión profunda reveló lo siguiente:

- Preparación de datos: las técnicas utilizadas, como la imputación aleatoria de datos faltantes y la eliminación selectiva de filas, las cuales, mejoraron significativamente la calidad del conjunto de datos para el modelado para que estas fueran aptas para la fase de modelado de datos.
- Modelado de datos: los modelos que se construyeron y analizaron tú venido éxito en general, se identifican áreas de mejora, particularmente en la precisión de la regresión logística para algunos combinaciones de variable. Por ejemplo las columnas o variables SalePrice y OverallCond

Decisión sobre Siguientes Pasos

- Con base en la evaluación de resultados y revisión del proceso, se recomiendan los siguientes pasos:
- Despliegue del modelo: si los modelos cumplen con los objetivos comerciales y técnicos, la implementación de producción pasa al uso operativo.
- Iteración y mejora: Si se identifican áreas de mejora durante la evaluación de resultados y revisión del proceso, se recomiendan iteraciones adicionales de la fase de minería de datos para refinar los modelos.

Reunión de Cierre

Se recomienda una reunión de cierre para resumir los hallazgos y discutir los próximos pasos. Esta reunión deberá incluir a todos los stakeholders clave para asegurar una transición efectiva a la fase de despliegue y establecer un plan de mantenimiento y monitorización del modelo en producción.

Este informe final consolida todos los aspectos del proyecto, proporcionando una visión clara del trabajo realizado, los resultados obtenidos y los próximos pasos recomendados para maximizar el valor del proyecto en el mercado inmobiliario de Iowa.

Se recomienda una reunión de cierre para resumir los resultados y discutir los próximos pasos. Esta reunión debe incluir a todas las partes interesadas clave para garantizar una transición efectiva durante la fase de implementación y establecer un plan para mantener y monitorear los modelos en producción.

4. Revisión del Proyecto

a. Evaluación de Lecciones Aprendidas

Implementación de nuevos tipos de aprendizaje en futuros proyectos:

- Aprendizaje supervisado: Utilizar datos etiquetados para entrenar modelos de regresión y clasificación.
- Aprendizaje no supervisado: aplicar técnicas como la agrupación en clústeres para descubrir patrones en datos redundantes.
- Aprendizaje por refuerzo: Desarrollar modelos que aprendan a través de la interacción con el entorno y la retroalimentación sobre sus acciones.

Consideraciones adicionales:

- Mejora la preparación de datos para garantizar la calidad y coherencia de los conjuntos de datos.
- Establecer un proceso de evaluación continua en producción para refinar y mejorar los modelos en uso.
- Manténgase actualizado con investigaciones y experimentos para adoptar nuevas técnicas y algoritmos.

b. Evaluación de la Calidad del Proceso

Identificar mejoras y recomendaciones.

- Puntos de mejora: Se identificaron áreas donde los modelos podrían optimizarse, como incluir más variables predictivas o explorar mejores técnicas de modelado.

- Recomendaciones: para futuros proyectos, se recomienda explorar técnicas de aprendizaje profundo, mejorar la gestión de datos en tiempo real y fortalecer la integración de modelos en los sistemas existentes para una implementación más fluida.

5. Decisión sobre Siguientes Pasos

a. Implementación de Producción

- Despliegue del modelo: Si el modelo cumple con los objetivos comerciales y técnicos, se procede a continuar con el despliegue a producción.
- Monitoreo Continuo: Establecer un sistema de monitoreo continuo para asegurar el desempeño y efectividad del modelo en producción.

b. Iteración y Mejora

- Refinamiento del modelo: si se identifican áreas de mejora durante la evaluación de resultados, se realiza iteraciones adicionales para refinar los modelos.
- Pruebas adicionales: se ejecutan pruebas adicionales para garantizar que las mejoras sean efectivas y estén alineadas con los objetivos comerciales.

# Conclusiones

El proceso de imputación de los datos garantiza que el precio total de la propiedad avanzada no contenga valores cero o cero, lo que podría afectar negativamente a las etapas posteriores de análisis y modelado. Los modelos predictivos serán más precisos y sólidos y, al mismo tiempo, preservarán la integridad y la calidad de los datos, esenciales para hacer recomendaciones efectivas al propietario de la empresa. Además, al eliminar el hilo solo se eliminan las columnas que contienen algunos de estos datos, lo que reduce la pérdida de información importante.

Los precios de venta de viviendas con características superiores, como lotes más grandes y áreas más urbanizadas, son generalmente más altos. Los propietarios pueden centrarse en propiedades con estas características para maximizar el valor de las ventas y mejorar la gestión del inventario y el marketing. Además, esto garantiza que los datos estén preparados y sean de alta calidad para los pasos de modelado posteriores.

El análisis de modelos con modelos K-Means y matrices correlacionadas muestran que el precio de venta y el espacio habitable son los principales componentes de la segmentación de propiedades. En función principalmente de estas características, las viviendas se agrupan en categorías de precio bajo, moderado, medio-alto y alto. Esto permite a los propietarios identificar segmentos de mercado y adaptar las estrategias de ventas y marketing a diferentes categorías de propiedades, mejorando la eficiencia de las decisiones y la gestión de la información.

El proyecto se desarrolló utilizando el marco CRISP-DM, lo que permitió estructurar adecuadamente el proceso de minería de datos en el Conjunto de datos avanzados de precios de la vivienda. Esto garantizó una metodología sistemática desde la comprensión del negocio hasta la implementación de modelos predictivos, garantizando eficiencia y claridad en cada etapa del proyecto.

Las estadísticas descriptivas y exploratorias utilizadas para el conjunto de datos presentados sobre los precios de la vida brindan información útil sobre las características que afectan los precios de la vida. Entre estas características variables como dimensión, tamaño, calidad de materiales y características del barrio, tenemos una visión e información integral para entender la dinámica del mercado inmobiliario en el conjunto de datos analizados.

El uso de técnicas avanzadas de extracción de datos en el conjunto de datos de Precios Avanzados de la Vivienda permitió la creación de modelos predictivos precisos y la identificación de tendencias importantes relacionadas con los precios de la vivienda. Estos modelos no solo pudieron predecir con precisión los precios de los bienes raíces, sino que también proporcionaron información importante sobre las variables más importantes, como la ubicación, las características estructurales y las condiciones del mercado, facilitando así decisiones informadas sobre bienes raíces.

# Recomendaciones

Se recomienda continuar actualizando y mejorando el conjunto de datos avanzados sobre precios de propiedades con datos externos y variables como tendencias económicas, tasas de interés y datos demográficos. Esto aumentará la precisión de los modelos predictivos y proporcionará una mejor comprensión de los factores que influyen en los precios inmobiliarios.

Se recomienda implementar un sistema que monitoree y dé respuesta continuamente a los modelos predictivos que se han desarrollado. Esto implica evaluar periódicamente el rendimiento y la exactitud de los modelos y actualizarlos con nuevos datos. También es recomendable formar al personal implicado en el uso y comprensión de estas herramientas analíticas para maximizar su eficacia y aplicación en la toma de decisiones estratégicas.

Se recomienda capacitar al personal que muestra las casa para mejorar la atención al cliente para mostrar y resaltar las cualidades que más influyen al comprar una casa como área del lote área vivible, la cantidad de cuartos, de la propiedad con otras, área del primer y segundo piso además de la longitud del baño y el garaje.

Se recomienda hacer una preparación general a las casa para su venta en primavera principalmente a las casa que su última remodelación se realizó entre los años 1950 1963 antes de la primavera esto porque es una de las épocas de más compra de bienes. Contratar de 5 a 10 equipos que den un manteniendo general por la posibilidad que la las casa tengan algún defecto que se pueda reparar fácilmente a las casa que se encuentren en el rango. Según (Thinkin World, 2021)“Como la mayoría de las industrias, los bienes raíces residenciales

tienen una estacionalidad. El número de casas vendidas en Estados Unidos durante la primavera es casi siempre mucho mayor que en cualquier otra época del año.”

Se recomienda dar mantenimiento a las áreas exteriores a la casa por ejemplo zonas verdes o piscina, el frente y alrededores de una casa es la primera impresión que un comprador posible comprador tiene al interactuar o ver una casa. por lo que dar un mantenimientos a estas áreas es de suma importancia ya que mejora el atractivo de la propiedad lo que significa una invención significativa para aumentar el interés de los posibles compradores por otra parte también se podría aumentar el precio de la propiedad al contar con estas características.

Se recomienda contratar un equipo de Home staging que vendría siendo un servicio orientado a la decoración y organización de casa. no se recomienda hacer esto con todas las casa disponibles para la venta sino dar este servicio a las casa que estén prontas a la visita de un posible comprados. Esto con el objetivo de atraer la atención de los posibles compradores de la propiedad ya que el interior estará decorado y organizado de manera que pueda mejorar la impresión ante el vendedor.

Se recomienda la investigación y uso de nuevas técnicas de Marketing y promoción del negocio. Esto con el objetivo de implementar nuevas formas de dar a conocer el negocio dueño de los datos y sobresalir entre otros negocios que se dediquen a la venta de casas. Se propone crear o dar mantenimiento al sitio web del negocio donde se puedan realizar trámites o consultorías en línea,

también promoción del negocio a través de redes sociales donde se pueda aumentar la popularidad del negocio.

# Bibliografía


Academia-lab.(13 de enero 2021). *Crisis financiera de 2008*. academia-lab.com https://academia-lab.com/enciclopedia/crisis-financiera-de-2008/

Canal A2 Capacitación: Excel (1 de mayo de 2023) *Aprende Python para ciencia de datos.* https://youtu.be/PpLtEo3TvFw?si=6dsZ3Us9Jl9ZoMAn

Climaytiempo.(26 de octubre del 2023). *El clima de Iowa y la mejor época para viajar.*

climaytiempo.es https://climaytiempo.es/estados-unidos/iowa/

Conectandoideas.(21 de abril del 2023). *Análisis de PCA: qué es para qué sirve y cómo aplicarlo en tus investigaciones* .conectandoideas.net. https://conectandoideas.net/analisis-dehttps://conectandoideas.net/analisis-de-pca/pca/

Eldiariony.(24 de mayo del 2024*). Los mejores lugares para compradores de vivienda en EE.UU., según Zillow.* Eldiariony.com https://eldiariony.com/2024/05/24/los-mejores-lugareshttps://eldiariony.com/2024/05/24/los-mejores-lugares-para-compradores-de-vivienda-en-ee-uu-segun-zillow/para-compradores-de-vivienda-en-ee-uu-segun-zillow/

excel-dashboards. (1 de mayo del 2021). *Tutorial de Excel: ¿Que es el formato CSV de Excel* .excel-dashboards.com. https://excel-dashboards.com/es/blogs/blog/excel-tutorial-what-ishttps://excel-dashboards.com/es/blogs/blog/excel-tutorial-what-is-excel-csv-formatexcel-csv-format

Ibm.(19 de marzo 1986). *Conceptos básicos de ayuda de CRISP-DM.* ibm.com. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Ibm.(19 de marzo del 1986). *¿Qué es la minería de datos?* Ibm.com https://www.ibm.com/es-es/topics/data-mining

Iebschool. (23de diciembre 2008). *Algoritmo k-means: ¿Qué es y cómo funciona?*

iebschool.com https://www.iebschool.com/blog/algoritmo-k-means-que-es-y-como-funcionahttps://www.iebschool.com/blog/algoritmo-k-means-que-es-y-como-funciona-big-data/big-data/

Leojimzdev.(8 de mayo del 2024). *Manejo correcto de los valores nulos en la estadística.* Leojimzdev.com https://leojimzdev.com/manejo-correcto-de-los-valores-nulos-en-lahttps://leojimzdev.com/manejo-correcto-de-los-valores-nulos-en-la-estadistica/ -Tecnicas_para_tratar_los_valores_nulos_en_la_estadisticaestadistica/#Tecnicas_para_tratar_los_valores_nulos_en_la_estadistica

Mikelnino.(16 de noviembre del 2019). *CRISP-DM: Fase de "Comprensión de los datos" (Data Understanding)* mikelnino.com https://www.mikelnino.com/2016/11/crisp-dmhttps://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-comprension-datos-data-understanding.html -:~:text=En%20esta%20fase%20el%20objetivo%20principal%20es%20poder,conocimiento%20que%20se%20puede%20extraer%20de%20los%20datosmetodologia-data-mining-comprension-datos-datahttps://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-comprension-datos-data-understanding.html -:~:text=En%20esta%20fase%20el%20objetivo%20principal%20es%20poder,conocimiento%20que%20se%20puede%20extraer%20de%20los%20datosunderstanding.html#:~:text=E

n%20esta%20fase%20el%20objetivo%20principal%20es%20poder

,conocimiento%20que%20se%20puede%20extraer%20de%20los%20datos.

Mikelnino.(20 de noviembre del 2019). *CRISP-DM: Fase de "Preparación de los datos"* mikelnino.com https://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mininghttps://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-preparacion-datos-data-preparation.htmlpreparacion-datos-data-preparation.html

Smartup.(11 de enero del 2011). *CRISP-DM: los 6 pasos del proceso de Data Mining.*

smartup.es. https://blog.smartup.es/crisp-dm-6-pasos-proceso-data-mining/

totvs.(14 de marzo del 2022). *Minería de datos: qué es, importancia y herramientas.*

totvs.com https://es.totvs.com/blog/gestion-de-negocios/mineria-de-datos-que-es-importanciahttps://es.totvs.com/blog/gestion-de-negocios/mineria-de-datos-que-es-importancia-y-herramientas/y-herramientas/

algoritmia8. (8 de Mayo del 2019). ¿Cómo aprenden los algoritmos? Aprendizaje supervisado, no supervisado y por refuerzo. algoritmia8.com https://algoritmia8.com/2020/09/15/como-aprenden-los-algoritmoshttps://algoritmia8.com/2020/09/15/como-aprenden-los-algoritmos-aprendizaje-supervisado-no-supervisado-y-por-refuerzo/ -:~:text=Gran%20parte%20de%20los%20algoritmos%20y%20t%C3%A9cnicas%20que,aprendizaje%20no%20supervisado%20y%20el%20aprendizaje%20por%20refaprendizaje-supervisado-no-supervisado-y-

porhttps://algoritmia8.com/2020/09/15/como-aprenden-los-algoritmos-aprendizaje-supervisado-no-supervisado-y-por-refuerzo/ -:~:text=Gran%20parte%20de%20los%20algoritmos%20y%20t%C3%A9cnicas%20que,aprendizaje%20no%20supervisado%20y%20el%20aprendizaje%20por%20refrefuerzo/#:~:text=Gran%20parte%20de%20los%20algoritmos%20y%20t%C3%A9cnicas%20que,aprendizaje%20no%20supervisado%20y%20el%20aprendizaje%20por%20ref

Codificandobits. (7 de Diciembre del 2022). Sets de entrenamiento, validación y prueba .codificandobits.com. https://www.codificandobits.com/curso/introduccion-machinehttps://www.codificandobits.com/curso/introduccion-machine-learning/8-sets-entrenamiento-validacion-prueba/ -:~:text=La%20idea%20de%20construir%20un%20modelo%20de%20Machine,con%20nuevos%20datos.%20Esto%20se%20conoce%20como%20generalizaci%C3%B3nlearning/8-sets-entrenamiento-validacionhttps://www.codificandobits.com/curso/introduccion-machine-learning/8-sets-entrenamiento-validacion-prueba/ -:~:text=La%20idea%20de%20construir%20un%20modelo%20de%20Machine,con%20nuevos%20datos.%20Esto%20se%20conoce%20como%20generalizaci%C3%B3nprueba/#:~:text=La%20idea%20de%20construir%20un%20modelo%20de%20Machine,con%20nuevos%20datos.%20Esto%20se%20conoce%20como%20generalizaci%C3%B3n.

Conectapyme.(7 de Julio del 2023). Modelado de Datos: Cómo estructurar tus datos para una mejor gestión. .conectapyme.com.

https://www.conectapyme.com/blog/modelado-de-datos-comohttps://www.conectapyme.com/blog/modelado-de-datos-como-estructurar-tus-datos-para-una-mejor-gestion/estructurar-tus-datos-para-una-mejor-gestion/

(Estadisticool, 2023)

Estadisticool.(11 de Octubre del 2023). Imputación de valores faltantes

(estadísticas): cómo imputar datos incompletos. estadisticool.com https://estadisticool.com/imputacion-de-valores-faltantes-estadisticas-comohttps://estadisticool.com/imputacion-de-valores-faltantes-estadisticas-como-imputar-datos-incompletos/imputar-datos-incompletos/

(IBM, 2024) Ibm. (16 de Marzo del 2024). ¿Qué es el aprendizaje no supervisado?. ibm.com. https://www.ibm.com/mx-es/topics/unsupervised-learning

(Madrigal, 2023) Growupcr. (20 de Eenero del 2023). Matriz de confusión: una herramienta para evaluar tus modelos de clasificación en Machine Learning. growupcr.com. https://www.growupcr.com/post/matriz-confusion

(Mikelnino, CRISP-DM: Fase de “Modelado” (Modeling), 2016) Mikelnino. (21 de Noviembre de 2016). CRISP-DM: Fase de “Modelado” (Modeling). mikelnino.com.

https://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mininghttps://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-modelado-modeling.html -:~:text=Consiste%20en%20la%20ejecuci%C3%B3n%20del%20algoritmo%20de%20modelado,que%20resulta%20y%20las%20dificultades%20para%20dicha%20interpretaci%C3%B3nmodeladohttps://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-modelado-modeling.html -:~:text=Consiste%20en%20la%20ejecuci%C3%B3n%20del%20algoritmo%20de%20modelado,que%20resulta%20y%20las%20dificultades%20para%20dicha%20interpretaci%C3%B3nmodeling.html#:~:text=Consiste%20en%20la%20ejecuci%C3%B3n%20del%20algoritmo%20de%20modelado,que%20resulta%20y%20las%20dificultades%20para%20dicha%20interpretaci%C3%B3n.

Probabilidadyestadistica.(14 de Agosto del 2021). Coeficiente de correlación de Pearson. probabilidadyestadistica.net. https://www.probabilidadyestadistica.net/coeficiente-de-correlacion-dehttps://www.probabilidadyestadistica.net/coeficiente-de-correlacion-de-pearson/ -%c2%bfque-es-el-coeficiente-de-correlacion-de-pearsonpearson/#%c2%bfque-es-el-coeficiente-de-correlacion-de-pearson

Thedataschools. (4 de Diciembre del 2022). Qué es la Limpieza de datos o data cleansing. thedataschools.com. https://thedataschools.com/quehttps://thedataschools.com/que-es/limpieza-de-datos-data-cleansing/es/limpieza-de-datos-data-cleansing/

Shallbd. (15 de Septiembre del 2023). El error cuadrático medio: Qué le dice y cómo interpretarlo. shallbd.com. https://shallbd.com/es/el-errorhttps://shallbd.com/es/el-error-cuadratico-medio-que-le-dice-y-como-interpretarlo/cuadratico-medio-que-le-dice-y-como-interpretarlo/

Themachinelearners. (3 de febrero del 2023). Métricas de Clasificación. themachinelearners.com. https://www.themachinelearners.com/metricas-dehttps://www.themachinelearners.com/metricas-de-clasificacion/ -

Accuracyclasificacion/#Accuracy thinkinworld.com. (2 diciembre del 2021) El mercado de bienes raíces en Estados Unidos como nunca antes visto en el invierno. thinkinworld.com. https://thinkinworld.com/realestate/el-mercado-de-bienes-raices-enhttps://thinkinworld.com/realestate/el-mercado-de-bienes-raices-en-estados-unidos-como-nunca-antes-visto-en-el-invierno/ -:~:text=El%20n%C3%BAmero%20de%20casas%20vendidas%20en%20Estados%20Unidos,y%20se%20mantiene%20fuerte%20durante%20todo%20el%20veranoestados-unidos-como-nunca-antes-visto-en-elhttps://thinkinworld.com/realestate/el-mercado-de-bienes-raices-en-estados-unidos-como-nunca-antes-visto-en-el-invierno/ -:~:text=El%20n%C3%BAmero%20de%20casas%20vendidas%20en%20Estados%20Unidos,y%20se%20mantiene%20fuerte%20durante%20todo%20el%20veranoinvierno/#:~:text=El%20n%C3%BAmero%20de%20casas%20vendidas%20en%20Estados%20Unidos,y%20se%20mantiene%20fuerte%20durante%20todo%20el%20verano.

-