Trabajo Práctico 1 : Propiedades en Venta

Introducción

En este trabajo práctico se propone que cada grupo de alumnos se enfrente a un problema real de ciencia de datos, que trabaje en cada una de las etapas del proceso y que pueda resolverlo aplicando los contenidos que vemos en la materia.
Vamos a utilizar el conjunto de datos provisto por la empresa Properati correspondiente a anuncios de propiedades en venta de la República Argentina publicados durante el año 2021.
La información fue extraída desde BigQuery (producto de Google Cloud para consultar grandes volúmenes de datos) donde la empresa disponibilizaba sus datasets con avisos de propiedades y publicados en Properati en todo Latinoamérica desde 2015.
El objetivo principal del trabajo será aplicar técnicas de análisis exploratorio, preprocesamiento de datos, agrupamiento, clasificación y regresión. En la sección enunciado se detallan los objetivos particulares.

Modalidad de entrega

Notebook

El trabajo debe ser realizado en una notebook de python, se espera que la misma contenga **todos** los resultados de la ejecución los cuales siempre deben ser **reproducibles**. La notebook debe respetar la siguiente nomenclatura : 7506R_TP1_GRUPOXX_ENTREGA
En el caso que sea estrictamente necesario entregar más de una notebook las mismas deben contar con una numeración correlativa manteniendo un orden lógico entre ellas (7506R_TP1_GRUPOXX_ENTREGA_**N1**, 7506R_TP1_GRUPOXX_ENTREGA_**N2**, etc)

Las secciones del trabajo deben estar claramente diferenciadas en la notebook utilizando celdas de markdown. Se debe incluir una sección principal con el título del trabajo, el número de grupo y el nombre de todos los integrantes.

Todo análisis realizado debe estar acompañado de su respectiva explicación y toda decisión tomada debe estar debidamente justificada. Cualquier hipótesis que sea considerada en el desarrollo del trabajo práctico debe ser detallada y debe estar informada en la entrega. Cualquier criterio que se utilice basado en fuentes externas (papers, bibliografía, etc.) debe estar correctamente referenciado en el trabajo.

Visualizaciones

Todos los gráficos que se incorporen deben tener su correspondiente título, leyenda, nombres en los ejes, unidades de medidas, y cualquier referencia que se considere necesaria. Es importante que tengan presente que los gráficos son una herramienta que facilita entender el problema, por lo tanto, deben ser comprensibles por quien los vaya a leer.

Preprocesamiento

A partir de las tareas de preprocesamiento, y de las diferentes estrategias que se planteen, es posible que se generen nuevos datasets sobre los cuales se entrenarán los modelos. Todo conjunto de datos creado debe ser almacenado y debe estar disponible en la entrega para ser utilizado por el equipo docente.

Modelos

Todos los modelos entrenados tanto para clasificación como para regresión deben ser guardados en un archivo (joblib / pickle) y deben estar disponibles en la entrega para ser utilizado por el equipo docente.

Reportes

Todos los reportes solicitados deben estar en formato pdf y deben tener la siguiente nomenclatura 7506R_TP1_GRUPOXX_CHPX_REPORTE. Además deben seguir los templates proporcionados por la cátedra.

Repositorio

Cada grupo deberá crear su propio repositorio en github: 7506R-1C2024-GRUPOXX
En dicho repositorio deberá estar disponible la notebook, los modelos entrenados, los conjuntos de datos utilizados para el entrenamiento y cualquier archivo que sea necesario para la correcta ejecución del trabajo.

Fechas de entrega

Check Point: 11/04 para esta fecha se espera que como mínimo tengan avanzadas las tareas de análisis exploratorio, preprocesamiento y agrupamiento. Deberán entregar un reporte de avance tomando como referencia los templates correspondientes. El código deberá estar disponible en el repositorio.

Check Point: 02/05 para esta fecha se espera que como mínimo tengan finalizadas las tareas de análisis exploratorio, preprocesamiento , agrupamiento y avanzadas las tareas clasificación y regresión. Deberán entregar un reporte de avance tomando como referencia los templates correspondientes.El código deberá estar disponible en el repositorio.

Entrega: 16/05 para esta fecha se espera que todos los grupos hayan resuelto en su totalidad el trabajo práctico cumpliendo con todas las consignas y condiciones de entrega. **Esta fecha es obligatoria.** Deberán entregar un reporte final tomando como referencia los templates correspondientes. El código deberá estar disponible en el repositorio.

Re entrega: 20/06 aquellos grupos que deban realizar correcciones contarán con esta fecha para volver a entregar el trabajo práctico. Esta es la última oportunidad para aprobarlo.

Enunciado

El conjunto de datos a utilizar **properati_argentina_2021** se encuentra disponible en el siguiente enlace, para este trabajo se plantean los siguientes objetivos generales:

> *Análisis Exploratorio y Preprocesamiento de Datos***:** el objetivo será explorar y preparar los datos para poder trabajar con los distintos modelos de aprendizaje automático.
>
> *Agrupamiento*: el objetivo será analizar si es posible agrupar los datos en función de algún criterio, identificando a qué obedece el mismo.
>
> *Clasificación:* el objetivo será clasificar cada anuncio en tres categorías relacionadas al precio de venta (alto, medio y bajo).
>
> *Regresión:* el objetivo será predecir el precio de venta en dólares de una propiedad tipo vivienda ubicada en Capital Federal.

A continuación se detallan las etapas que deben ser desarrolladas en el trabajo:

1. Análisis Exploratorio y Preprocesamiento de Datos

El primer paso consiste en la selección de los datos que se van a utilizar, se deben filtrar únicamente los anuncios de propiedades de tipo vivienda (Casa, PH y Departamento) ubicados en Capital Federal cuyo tipo de operación sea venta y su precio se encuentre en dólares (USD). Se debe separar un conjunto de entrenamiento (80%) y un conjunto de test (20%).

a) Exploración Inicial : analizar cada variable, considerando los siguientes aspectos

- Variables Cuantitativas: calcular medidas de resumen: media, mediana,moda,etc
- Variables Cualitativas: mostrar cantidad de valores posibles, y frecuencias de cada uno.
- Determinar variables irrelevantes en el análisis
- Realizar un análisis gráfico de las distribuciones de las variables más relevantes
- Analizar las correlaciones existentes entre las variables.

b) Visualización de los datos: en esta sección se espera que puedan realizar una primera aproximación a los datos apoyándose en visualizaciones, por ejemplo: gráficos de dispersión entre variables, histogramas, heatmaps, exploración de las columnas y cualquier otro gráfico adicional que se considere útil justificando su utilización.

c) Datos Faltantes : analizar la presencia de datos faltantes en el dataset

- Realizar análisis de datos faltantes a nivel de columna. Graficar para cada variable el porcentaje de datos faltantes con respecto al total del dataset.
- Realizar un análisis de datos faltantes a nivel de fila. Calcular el porcentaje de datos faltantes de cada registro. Realizar un gráfico que permita conocer la proporción de faltantes por fila en el dataset.
- Revisar los datos faltantes o mal ingresados y tomar una decisión sobre estos: reemplazo de valores, eliminación de registros incompletos, etc.
- En caso de realizar imputaciones comparar las distribuciones de cada atributo reparado con la distribución anterior a la imputación de los datos faltantes.

d) Valores atípicos : analizar la existencia de valores atípicos

- Detectar valores atípicos en los datos tanto en forma univariada como multivariada. Realizar gráficos que permitan visualizar los valores atípicos.
- Explicar qué características poseen los datos atípicos detectados.
- Decidir el tratamiento a aplicar sobre los mismos.
- Analizar la relación entre el precio de venta y los metros de superficie ¿hay valores atípicos que no se detectaron previamente?

Nota : Los ítems a, b, c y d son los mínimos requeridos para esta etapa, cada grupo puede crear nuevas variables que resulten derivadas de los atributos existentes o que resulten de incorporar nuevas fuentes de datos.

2. Agrupamiento

En este punto se busca analizar si es posible agrupar los datos en función de algún criterio. Para esta tarea utilizar el algoritmo K-Means y se deberán realizar los siguientes puntos:

a. Analizar la tendencia al *clustering* del dataset.
b. Estimar la cantidad apropiada de grupos que se deben formar.
c. Evaluar la calidad de los grupos formados realizando un análisis de *Silhouette*.
d. Realizar un análisis de cada grupo intentando entender en función de qué características fueron formados.
e. Graficar sobre un mapa de CABA los avisos coloreados según el grupo al que pertenecen.
f. Repetir el análisis anterior, utilizando sólo 3 grupos.

3. Clasificación

a) Construcción del target

Para esta tarea se debe crear una nueva variable ***tipo_precio*** que tendrá tres categorías: **alto, medio, bajo.** Esta nueva variable será nuestra clase en el problema de clasificación.
Para determinar cuándo el tipo_precio de una propiedad es alto, medio o bajo se deberá analizar el precio por metro cuadrado (pxm2). Se propone evaluar las siguientes alternativas para establecer los límites de cada categoría:

1. Dividir la variable *pxm2* en 3 intervalos con igual cantidad de observaciones.
2. Dividir la variable pxm2 en 3 intervalos, el primero con el 25% de las observaciones, el siguiente con el 50% y el último con el 25% de las observaciones restantes.
3. Trabajar la variable pxm2 relativa a cada tipo de propiedad y luego dividirla como en el punto anterior.

Se pide:

a. Mostrar la distribución del precio por metro cuadrado
b. Mostrar la distribución del precio por metro cuadrado por tipo de propiedad

c. Para cada una de las tres alternativas mostrar gráficamente la distribución de la nueva variable creada *tipo_precio*.
d. Seleccionar una de las alternativas, justificando la misma.
e. Comparar, si aplica, la alternativa seleccionada con la división en 3 grupos obtenida utilizando agrupamiento por K Means.
f. Mostrar en un mapa de CABA los avisos coloreados por *tipo_precio*

b) Entrenamiento y Predicción

Predecir el valor del atributo **tipo_precio**, excluyendo del entrenamiento la variable price, **pxm2** y cualquier otra que pueda contener información del precio de venta. Para todos los modelos se pide realizar las tareas de ingeniería de características necesarias para trabajar con cada algoritmo (*encoding*, normalización, etc).

Modelo 1 : Árbol de decisión

a. Construir un árbol de decisión y optimizar sus hiperparámetros mediante *k-fold Cross Validatio*n para obtener la mejor performance. ¿Cuántos *folds* utilizaron? ¿Qué métrica consideran adecuada para buscar los parámetros?
b. Graficar el árbol de decisión con mejor performance encontrado en el punto anterior. Si es muy extenso mostrar una porción representativa.
c. Analizar el árbol de decisión seleccionado describiendo los atributos elegidos, y decisiones evaluadas (explicar las primeras reglas obtenidas).
d. Evaluar la performance del árbol en el conjunto de evaluación, explicar todas las métricas y mostrar la matriz de confusión. Comparar con la performance de entrenamiento.

Modelo 2: Random Forest

a. Construir un clasificador RF y optimizar sus hiperparámetros mediante *k-fold Cross Validation* para obtener la mejor performance. ¿Cuántos *folds* utilizaron? ¿Qué métrica consideran adecuada para buscar los parámetros?
b. Analizar la importancia de los atributos
c. Mostrar la conformación final de uno de los árboles generados. Si es muy extenso mostrar una porción representativa y explicar las primeras reglas.
d. Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas y mostrar la matriz de confusión. Comparar con la performance de entrenamiento.

Modelo 3: a elección

a. En este punto se debe entrenar (mediante cross-validation) un modelo elegido por el grupo. Se debe evaluar su performance en entrenamiento y sobre el conjunto de evaluación, explicar todas las métricas y mostrar la matriz de confusión.

¿Qué modelo elegirían para clasificar el tipo de precio de las propiedades?

4. Regresión

En esta etapa se busca predecir el precio de la propiedad utilizando dos modelos diferentes. Para todos los modelos se pide realizar las tareas de ingeniería de características necesarias para trabajar con cada algoritmo (encoding, normalización, etc).

Modelo 1: KNN

a. Construir un modelo KNN para regresión y optimizar sus hiperparámetros mediante *k-fold Cross Validation* para obtener la mejor performance .¿Cuántos *folds* utilizaron? ¿Qué métrica consideran adecuada para buscar los parámetros?
b. Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

Modelo 2: XGBoost

a. Construir un modelo **XGBoost** y optimizar sus hiperparámetros mediante *k-fold Cross Validation* para obtener la mejor performance. ¿Cuántos *folds* utilizaron? ¿Qué métrica consideran adecuada para buscar los parámetros?
b. Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

Modelo 3: a elección

a. En este punto se debe entrenar (mediante *cross-validation*) un modelo elegido por el grupo. Se debe evaluar su performance en entrenamiento y sobre el conjunto de evaluación explicando todas las métricas.

¿Qué modelo elegirían para predecir el precio de venta de las propiedades?