Plan Overview

A Data Management Plan created using DMPTool

Title: Predictor de Precios de Vivienda en Colombia

Creator: Daniela Castillo

Affiliation: Universidad de Los Andes (uniandes.edu.co)

Principal Investigator: Andres Santiago Serna Tangarife, Katherine Russi Parra, William Morales, Daniela Castillo Tellez

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

El objetivo de este proyecto es desarrollar un modelo de aprendizaje automático para predecir los precios de vivienda en Colombia. Se utilizará un conjunto de datos de propiedades que posee información del precio, la ubicación, características y condición de las viviendas.

El modelo se entrenará utilizando un conjunto de datos de entrenamiento y se evaluará utilizando un conjunto de datos de prueba. Se utilizarán diferentes algoritmos de aprendizaje automático para evaluar su rendimiento.

Los resultados serán validados con los precios reales para evaluar la precisión del modelo. El modelo se puede utilizar para ayudar a los compradores y vendedores de propiedad raíz a tomar decisiones sobre el mercado inmobiliario.

Start date: 10-16-2023

End date: 12-02-2023

Last modified: 10-19-2023

Predictor de Precios de Vivienda en Colombia

Data Collection

What data will you collect or create?

Para iniciar el proyecto, se recolectaron datos de 12853 inmuebles de diferentes departamentos y ciudades de Colombia. El formato de los datos es un archivo de Excel donde se cuenta con toda la información de las propiedades y sus respectivos avalúos.

Una vez finalizado el proyecto, se continuará con la recolección de datos periódica para futuros re entrenamientos del modelo y para validar la capacidad predictiva del modelo actual con el fin de identificar la necesidad de re entrenamiento.

Dado que el volumen de datos no es muy grande, estos se podrán continuar almacenando en excel para su posterior procesamiento.

How will the data be collected or created?

La recolección de datos se realizó a partir de un proveedor de datos con la información de diferentes proyectos de vivienda en Colombia cuya información se complementó utilizando técnicas de scraping de diferentes páginas web de venta de inmuebles en Colombia.

Una vez finalizado el proyecto se continuará con la recolección de dichos datos sobre los cuales se aplicarán predicciones re entrenamientos del modelo. Estos datos serán versionados utilizando la herramienta DVC con estructura de carpetas por año y mes de descarga de los datos.

Sobre los datos obtenidos, se aplicará un pipeline de limpieza y preparación de datos que se encargará de asegurar calidad y estructurar los datos como se requieren para las predicciones y entrenamiento de los modelos.

Documentation and Metadata

What documentation and metadata will accompany the data?

Se cuenta con la documentación de metadatos de algunos campos de la base de datos. Al finalizar el proyecto, se complementará la documentación de los metadatos para todos los campos,

especialmente los requeridos por los modelos. La documentación de datos deberá tener:

Nombre de la variable
Descripción: Descripción de la variable.
Tipo de dato
Transformaciones: Las transformaciones realizadas para llegar a dicha variable si es una variable calculada.
Responsable
Fecha de creación
Fecha de modificación

Ethics and Legal Compliance

How will you manage any ethical issues?

La información no contiene datos confidenciales, no posee información relacionada con personas o datos que requieran algún tratamiento especial referente a cuestiones éticas.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

Teniendo en cuenta que los datos son públicos, su manejo no requiere derechos de autor ni derecho de propiedad intelectual.

Storage and Backup

How will the data be stored and backed up during the research?

Se cuenta con suficiente capacidad de almacenamiento de los datos, los cuales serán respaldados en aws en buckets de s3 utilizando versionamiento de datos con DVC

How will you manage access and security?

Los datos y despliegue de la solución serán almacenados en la nube AWS, desde las capacidades que ofrece este proveedor se controlará la seguridad de accesos tanto a los datos como a la aplicación. Se

utilizarán las herramientas propias de Backup de AWS S3, para realizar copias de seguridad o restauración de la misma en caso de ser necesario.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Actualmente no existen obligaciones contractuales, legales o regulatorias que exijan depuración o almacenamiento de los datos utilizados en el proyecto. Sin embargo, se considerará preservar al menos un año de información histórica, con el fin de contar con datos para futuros re entrenamientos del modelo.

What is the long-term preservation plan for the dataset?

De acuerdo al presupuesto con que se cuente una vez finalizado el proyecto, se re evaluará el conservar mayor información histórica que sea útil para la visualización de los tableros y validaciones de capacidad predictiva del modelo actual.

Data Sharing

How will you share the data?

Los datos estarán disponibles para los usuarios a través de un tablero de control desplegado en la nube. En este tablero, los usuarios podrán realizar un análisis descriptivo inicial basado en la información disponible. Podrán explorar las principales características de los predios en una primera sección. Además, habrá una sección dedicada para acceder a las predicciones generadas por el modelo, permitiendo así un seguimiento detallado de la capacidad predictiva del mismo.

Are any restrictions on data sharing required?

Dado el caso de que algún campo tenga información sensible esta será encriptada para mayor seguridad. Es posible que los datos se utilicen exclusivamente durante la duración del proyecto. Sin embargo, es esencial evaluar si ciertos componentes de los datos deben mantenerse exclusivamente durante un período específico.

Responsibilities and Resources

Who will be responsible for data management?

El proceso a desarrollar será supervisado por el grupo de estudiantes presentes de la Maestria de Analitica de Datos de la Universidad de los Andes incluidos en el presente plan

What resources will you require to deliver your plan?

Dentro del equipo de trabajo necesitamos habilidades y conocimientos en uso de herramientas específicas de análisis de datos, software de manipulación de datos, despliegue de soluciones analíticas junto a un conocimiento experto de la temática.

Se debe evaluar si se necesita hardware o software especializado para análisis de datos o modelado de aprendizaje automático que no esté disponible.

Considerar si los repositorios de datos o el despliegue de la solución aplicarán cargos por el almacenamiento y la preservación a largo plazo de los conjuntos de datos utilizados en el proyecto.