Instrucciones Proyecto_Final

El presente proyecto une datos de construcciones residenciales del sector privado en el año 2022 con datos de la encuesta nacional de hogares del mismo año. La variable objetivo a predecir es 'Tenencia de Vivienda' la cual indica si un hogar o no es dueño de la vivienda en que residen o esta es alquilada.

Dado que los datos de la ENAHO se recogen a nivel de region geográfica y los de construcción están a nivel de cantón dos bases de datos intermediarias que contienen codificaciones fueron utilizadas para poder hacer la unión final entre estas dos bases de datos

En resumen las bases de datos utilizadas son las siguientes:

Base_Anonimizada2022.csv (ENAHO)

BdBasePublica.csv (datos construcción)

division_territorial_por_region.csv (codificación de las regiones en Costa Rica)

SEN_GEOGRAFICO_1.csv (codificación cantones Costa Rica)

Descripción de las funciones utilizadas:

construcciones_region_df_func:

Esta función toma el dataset de construcción, cantón y región selecciona las variables de interés y hace un pre-procesamiento de los tres datasets antes de unir por los respectivos códigos que identifican a cada cantón y a cada región.

Finalmente los datos son agrupados para tener los promedios de las variables de construcción para cada región geográfica de Costa Rica.

enaho_func:

Esta función se encarga del pre-procesamiento de la base de datos de enaho. Primero se seleccionan las variables de interés, posteriormente se convierte la variable 'Tenencia de Vivienda' en una variable binaria y por ultimo se crea una lógica que permite agrupar los datos a nivel de hogar para aquellas variables que están a nivel individual.

unir_datos:

Esta función se encarga de unir los dos datasets generados por las dos funciones anteriores. Este será el dataset sobre el cual se entrenen los modelos.

guardar_a_postgres:

La función toma los tres datasets generados por las funciones anteriores y los manda a escribir a la base de datos.

Modelos Entrenados

Regresión logística:

El primer modelo entrenado aplica una regresión logística para predecir la variable de interés ‘Tenencia de Vivienda’

Este modelo genera un Área Bajo la Curva del ROC de 0.67.

Árbol de Decisión:

El segundo modelo es un árbol de decisión. Dicho modelo genera una Area Bajo la curva del ROC de 0.57.

Análisis de Resultados de Modelos.

Como puede observarse los mejores resultados son generados por el modelo de regresión logística sin embargo estos son deficientes. Una de las razones para esto es que se opto por imputar los valores nulos de los variables con 0 ya que son variables numéricas que representan magnitudes. Esto puede remplazarse por el promedio de cada variable y ver si esto mejora los resultados.

Una de las razones por las cuales el modelo de regresión logística tiene un mayor porcentaje de aciertos que el árbol de decisión puede deberse a que muchas de las variables predictoras muestran una relación lineal con la variable objetivo. En dichas circunstancias los modelos de regresión logística suelen tener una mejor tasa de acierto que los arboles de decisión.