Proyecto Ciencia de Datos

Castigliony Giuliana

22/06/23

Base de Datos de Properati para predecir precios de ventas de inmuebles

Datos

- Unificación de 9 datasets distintos para obtener una cantidad mayor de datos
- 550857 registros
- Información recopilada de Argentina del año 2013 hasta el año 2017.

Ventajas

- Útil para inversión inmobiliaria
- No se necesita conocimiento previo del mercado local
- Súper divertido el dominio.

Desventajas

- Presencia de una gran cantidad de datos absurdos o nulos debido a la posibilidad de que las personas ingresen información irrelevante o sin sentido.

Variables de la unión de los Datasets

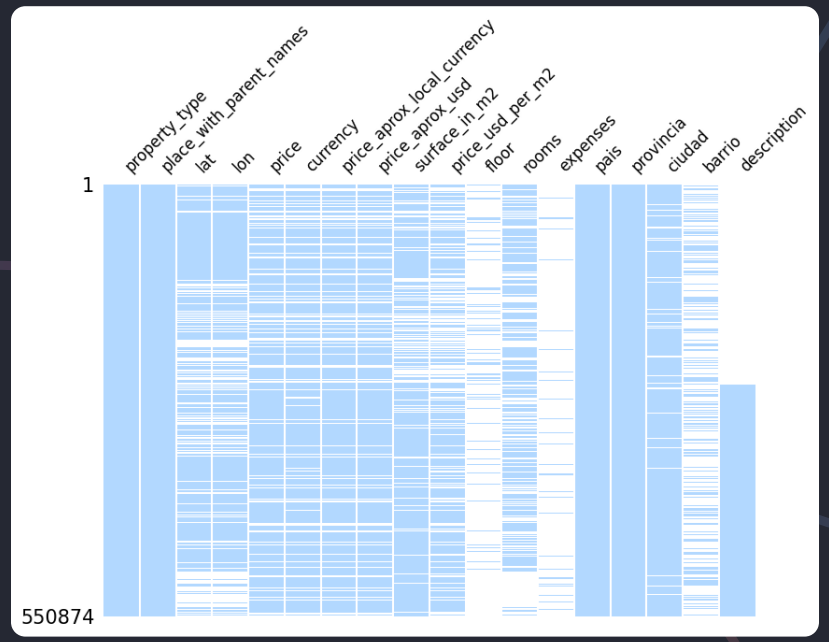
- **property_type:** Tipo de propiedad de las cuales puede ser 'house', 'apartment', 'PH', 'store'.
- **place_with_parent_names:** Dato concatenado que provee, Pais,Provincia,Ciudad, Barrio. el formato es x|x|x|x o x|x|x. donde Ciudad o Barrio varia.
- **lat:** Latitud geografía.
- **lon:** Longitud geográfica.
- **price:** Precio
- **currency:** Moneda en la cual se encuentra el precio
- **price_aprox_local_currency:** Precio aproximado en la moneda local.
- **price_aprox_usd:** Precio aproximado en dólares.
- **surface_in_m2:** Superficie en metros cuadrados.
- **price_usd_per_m2:** Precio en USD por metro cuadrado
- **floor:** Cantidad de pisos.
- **rooms:** Habitaciones.
- **expenses:** Expensas.
- **pais:** País fue sacado de "place_with_parent_names".
- **provincia:** Las provincias fueron sacadas de "place_with_parent_names".
- **ciudad:** Las ciudades fueron sacadas de "place_with_parent_names".
- **barrio:** Los barrios fueron sacados de "place_with_parent_names".
- **description:** Descripción dada por las personas que publican los anuncios.

Recuperación de datos faltantes

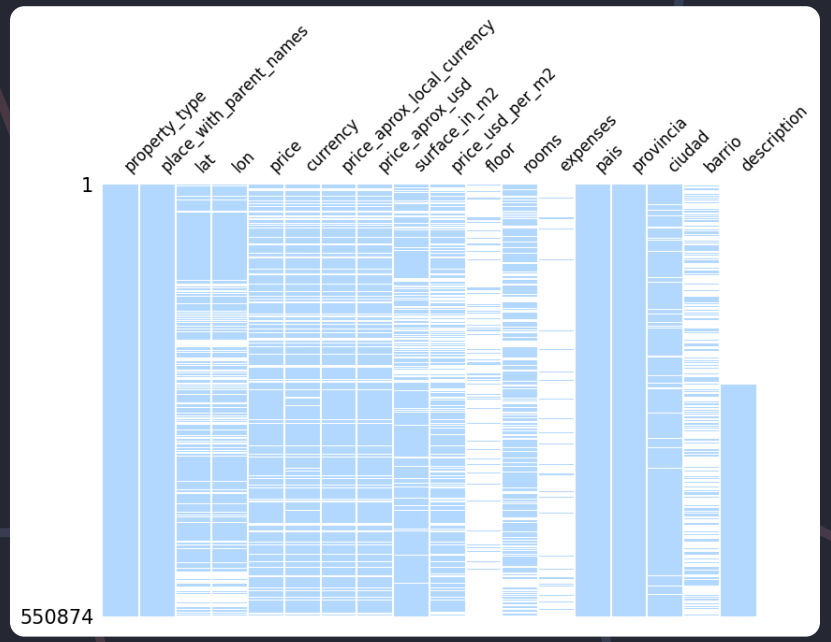
Utilizamos expresiones regulares para encontrar números y palabras clave relevantes en el texto de la variable "description". De esta manera, podemos extraer información importante relacionada con nuestro objetivo, a continuación se muestran la cantidad de datos recuperados:

- Se pudo rescatar de la característica "surface_in_m2" 30772 registros.
- Se pudo rescatar de la característica "rooms" 42505 registros.

Comparación de Datos nulos



Antes de la recuperación de datos



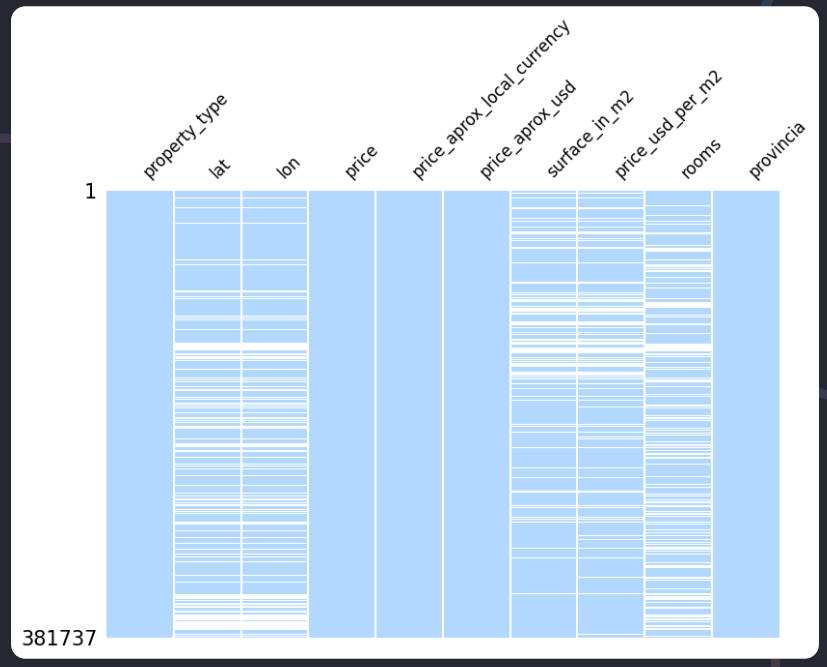
Después de la recuperación de datos

Definimos las clases que vamos a utilizar para comenzar a borrar datos:

1. currency : 'USD'
2. property_type: 'house' 'apartment' 'PH'
3. pais : 'Argentina'

Borramos las variables que no vamos a usar por falta de datos o porque no aportan información importante

- 'place_with_parent_names',
- 'currency',
- 'floor',
- 'expenses',
- 'pais',
- 'ciudad',
- 'barrio',
- 'description'



Actualización Dataset

A esta altura tenemos de 381737 datos de los cuales gran parte siguen siendo nulos como se ve en el gráfico, por lo tanto eliminamos todos los datos nulos y nos quedamos con un dataset de 177998 datos.

Limpieza de datos irreales

- En la variable 'price_usd_per_m2', se eliminaron valores superiores a 10,000 y valores inferiores a 100.
- En la variable 'surface_in_m2', se eliminaron valores superiores a 1,000 y valores inferiores a 10.
- En la variable 'rooms', se reemplazaron los valores 0 por 1 y se eliminaron valores superiores a 15 y valores inferiores a 1(habían cantidad de ambientes negativos).
- En las variables 'lat' y 'lon', se eliminaron valores de latitud que estaban fuera del rango aceptable (-10 a -54.7) y valores de longitud que estaban fuera del rango aceptable (-80 a -50).
- También se eliminaron puntos específicos de 'lat' y 'lon' que no eliminaba la restricción anterior.

Gráfico de latitud y longitud sin la limpieza de los datos irreales

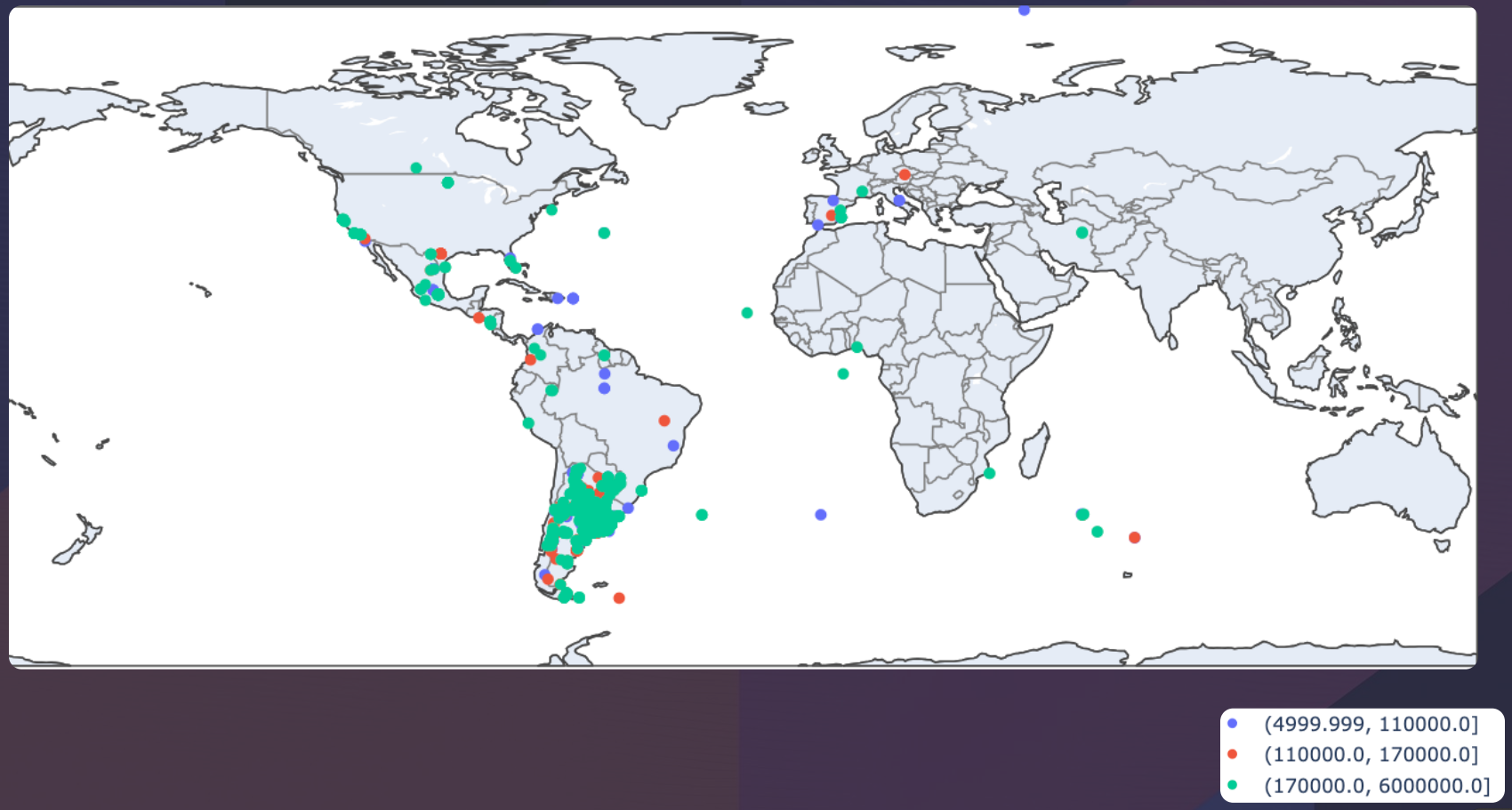
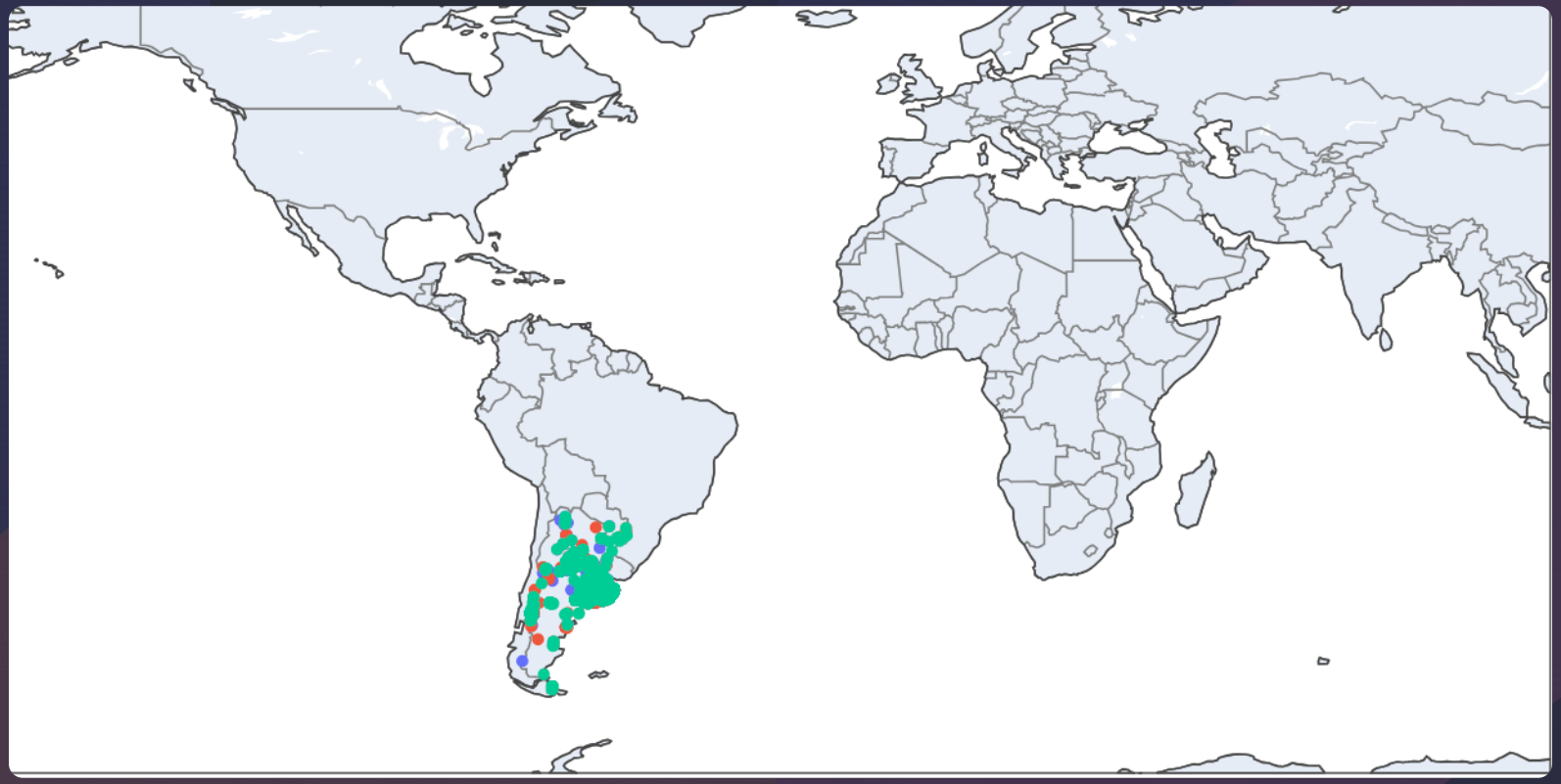
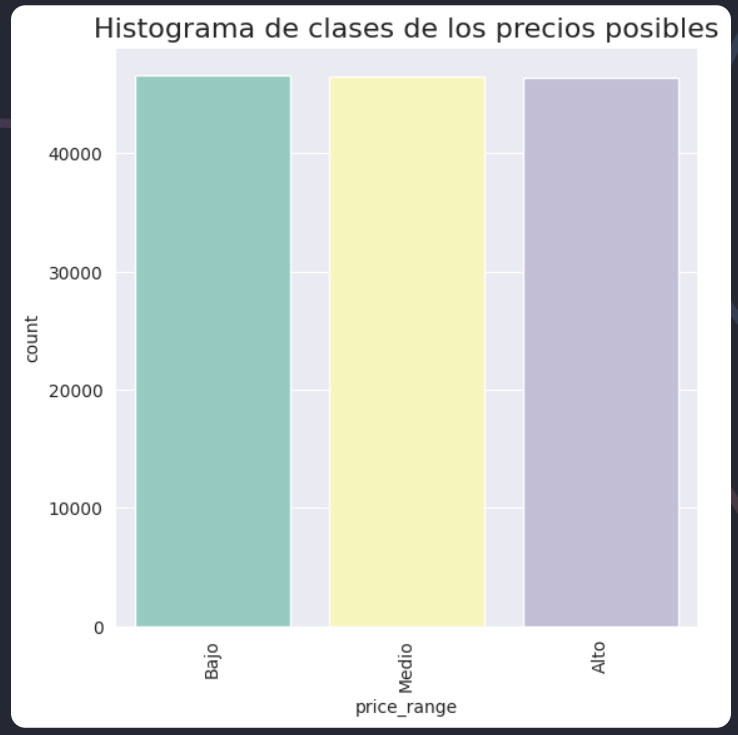


Gráfico de latitud y longitud con la limpieza de los datos irreales





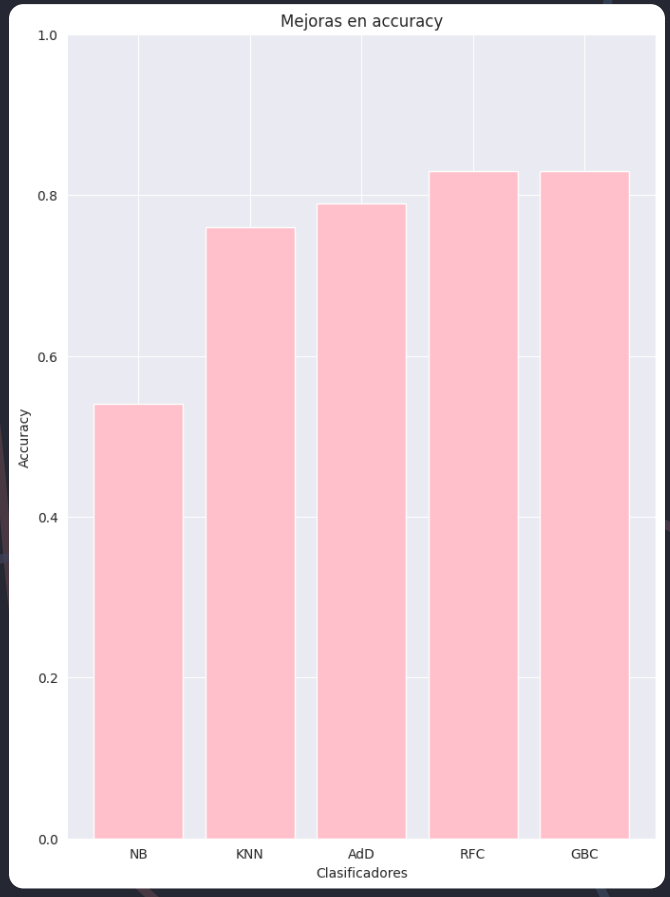
Selección de Variables para clasificar

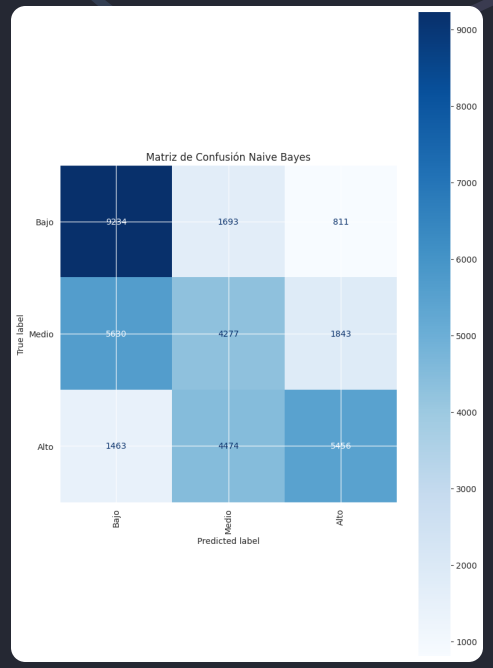
El dataset utilizado para clasificar contiene 139523 registros y se seleccionaron las siguientes variables:

- 'property_type'
- 'lat',
- 'lon'
- 'surface_in_m2'
- 'rooms'
- 'provincia'

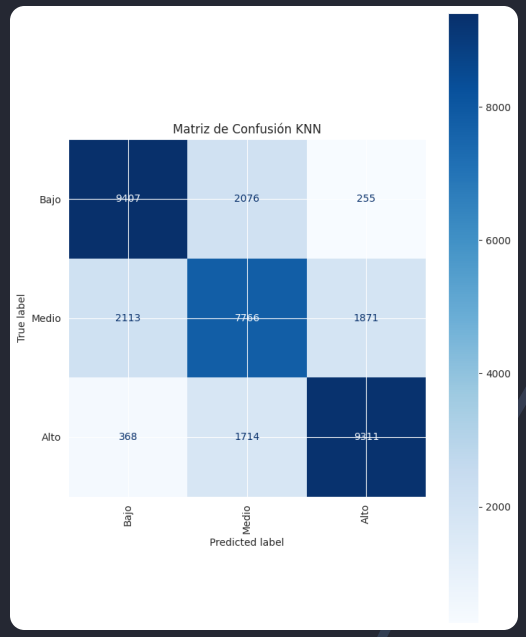
Resultados

Modelos	
Naive Bayes	0.54
K-Nearest Neighbors	0.76
Árboles de Decisión	0.79
Random Forests	0.83
Gradient Boosting	0.83

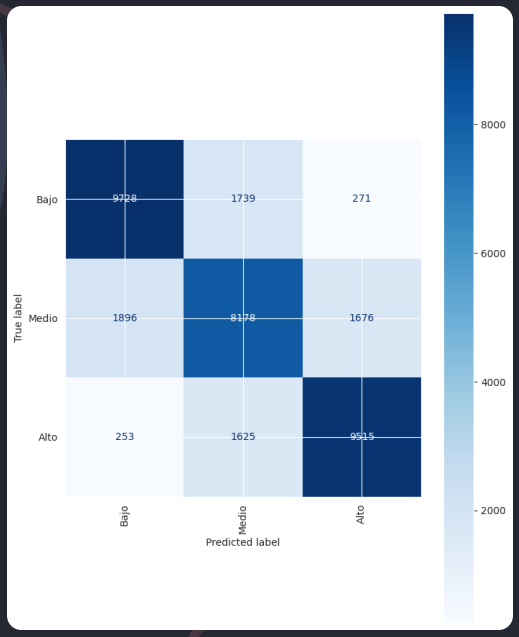




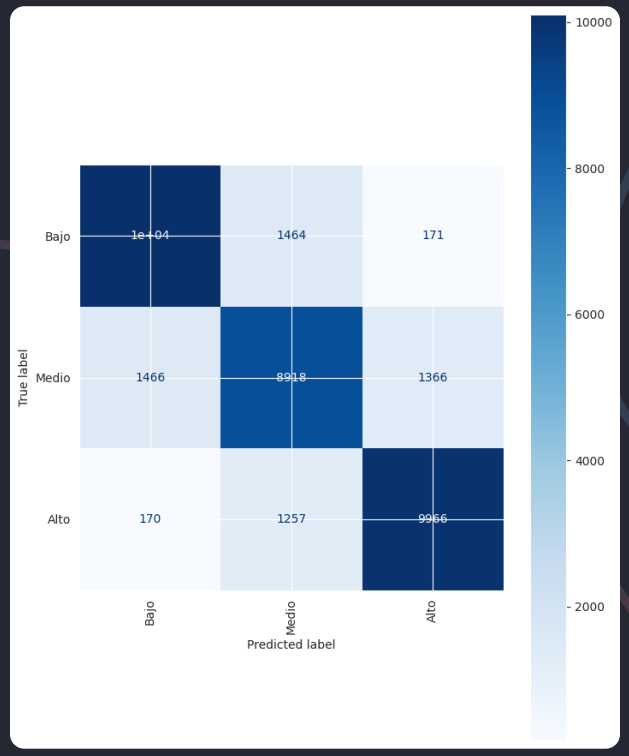
Naive Bayes



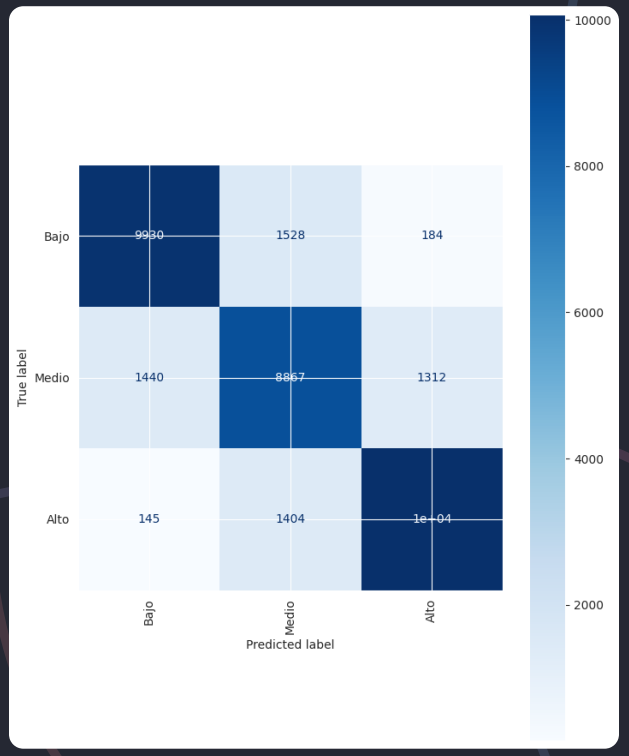
KNN



Árboles de decisión



Random Forests



Gradient Boosting

Fin