Prueba de conocimiento

Sebastian Saavedra

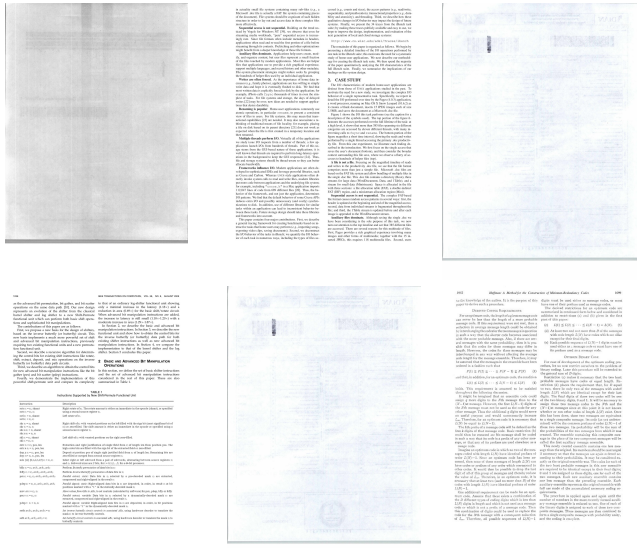
Junio 2024

Análisis de Imágenes

Prueba

Desarrollar un proceso que permita clasificar imágenes en diferentes carpetas, con el criterio de si las imágenes contienen o no información.

Información Original

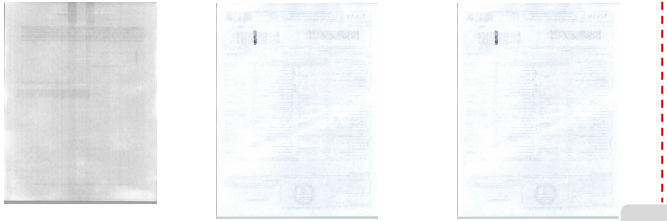


247 imágenes

Imágenes con información

100 imágenes

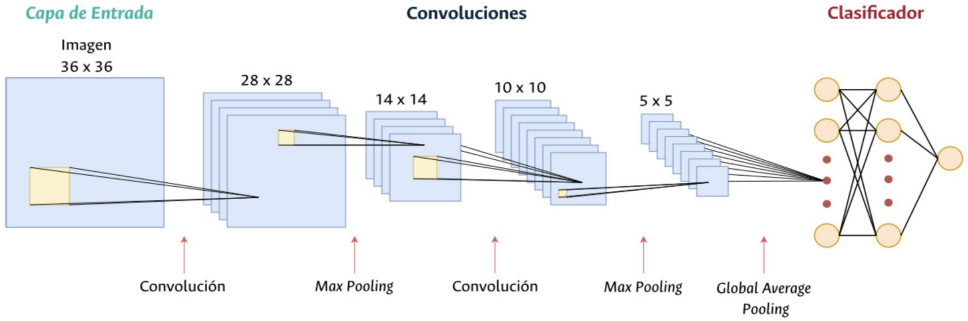
Imágenes en Blanco



147 imágenes

Metodología propuesta

Red Neuronal Convolucional



Red neuronal utilizada en el procesamiento de imágenes, incluidas tareas de clasificación.

Se entrena un red que clasificará las imágenes en “con información” y “sin información.

Data Augmented

Técnica que permite extender el conjunto de datos, generando transformaciones (cambios en traslación, rotación, intensidad, entre otros) sobre las imágenes originales.

Se selecciona esta técnica para tener información suficiente en el entrenamiento de la re.

Resultados

Red Neuronal convolucional

División de las imágenes en:

Entrenamiento: 148 imágenes
Validación: 49 imágenes
Prueba: 50 imágenes

Características red neuronal:

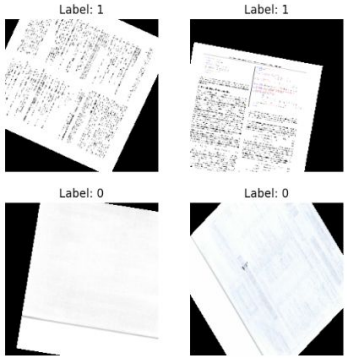
- 1 Capa de entrada
- 5 Bloques con capas convolucionales con función de activación *relu* y capas de pooling.
- 1 Capa Flatten
- 2 capas densas de 256 y 512 neuronas con función de activación *relu*.
- 1 Capa de regularización con un dropout de 0.2.
- 1 Capa de salida con 1 neurona con función de activación *sigmoid*.

Data Augmented

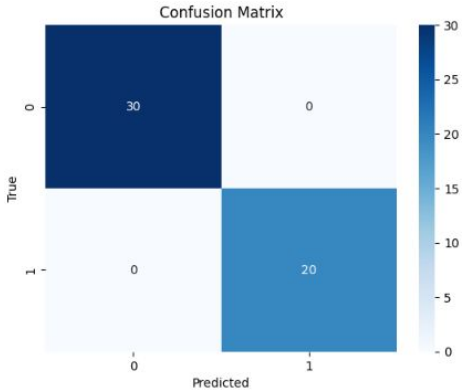
Se generan nuevas imágenes con las siguientes transformación:

- Rotación.
- Traslación (Vertical y horizontal).
- Aumento.
- Inversión.
- Rellenado de espacios.
- Reescalado

Ejemplo



Resultado

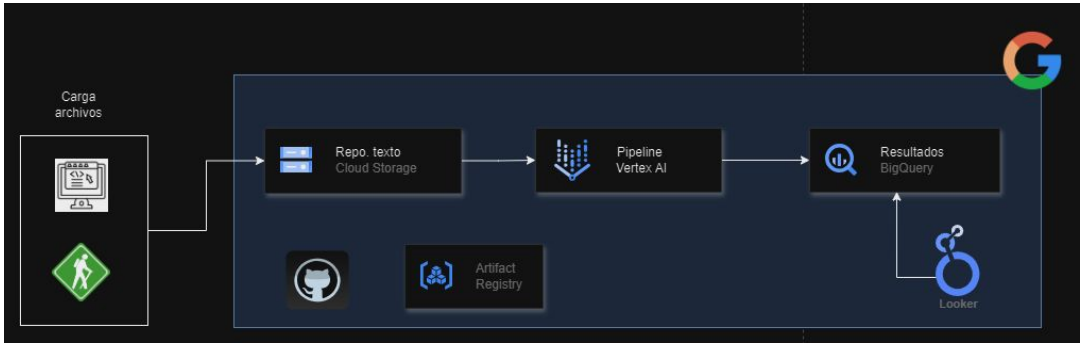


En el conjunto de prueba se logra una clasificación perfecta. Obteniendo las métricas:

Accuracy: 1.0
Precision: 1.0
F1-score: 1.0

Arquitecturas de referencia

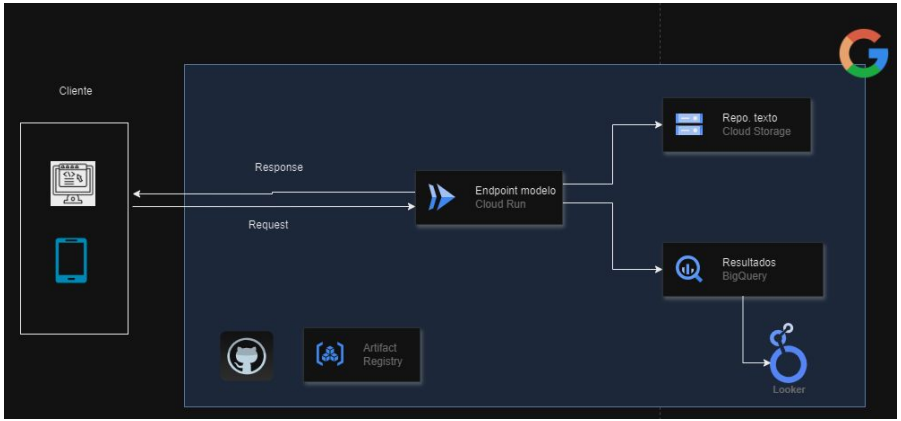
Arquitectura Batch



Las Arquitecturas son planteadas en GCP dado que el lago de Datos de Davivienda se encuentra en esta nube. Sin embargo, la solución puede ser desplegada en cualquier infraestructura (AWS, Azure, On premise. etc)

Para ambas arquitecturas propuestas se debe considerar:

- Esquema de Seguridad.
- Esquema de autenticación.
- Validación de Volumetrías



Arquitectura Real time

Retos, Consideraciones y Conclusiones

Principales retos

- Cantidad reducida en el conjunto de imágenes, sin embargo para el entrenamiento se aborda esto con Data Augmented.
- Se recomienda la evaluación del modelo con nuevas imagenes, en un conjunto de datos de mayor tamaño.

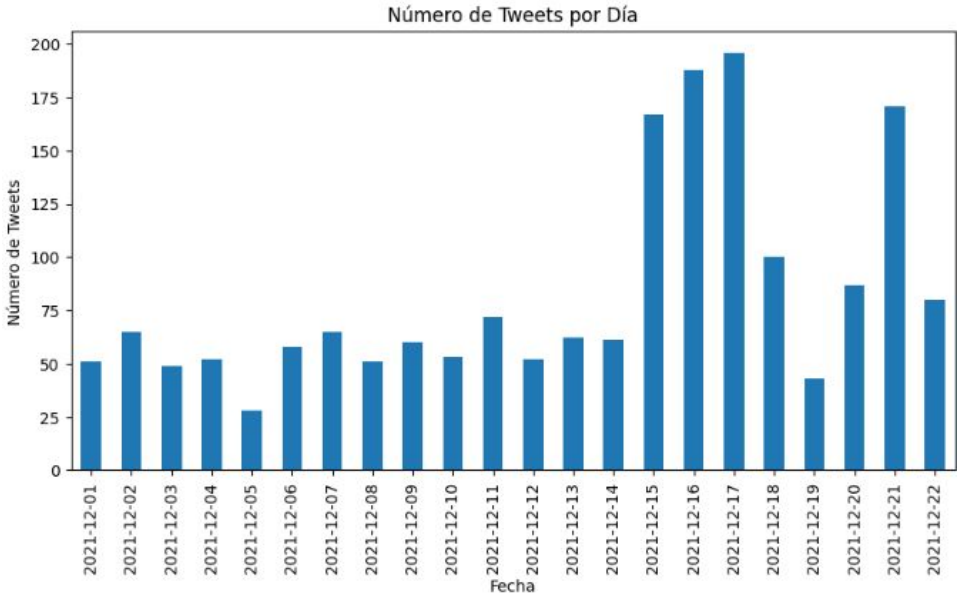
Conclusiones y consideraciones

- Con el proceso desarrollado se cumple el objetivo de poder clasificar imágenes con información e imágenes en blanco.
- La combinación de Data Augmented y la Red Neuronal Convolucional da un desempeño perfecto en la tarea de clasificación.
- Se debe monitorear el proceso para garantizar su correcto funcionamiento.
- En caso de deterioro se recomienda el reentrenamiento del modelo con imágenes adicionales.
- Se recomienda la implementación del modelo en tiempo real, dado que puede aportar en la recepción de documentos cargados por el cliente en múltiples procesos.
- El código desarrollado puede ser reutilizado en su mayoría, únicamente se deben considerar los repositorios de información a utilizar.

Análisis de Tweets

Prueba - Exploración de datos

Análisis de 1.811 Tweets queen la que se menciona a la organización Davivienda.



Los tweets se encuentran entre el 1 de Diciembre de 2021 y el 22 de Diciembre del 2021, lo cual es valioso de conocer, dado que la cantidad de transacciones realizadas por los clientes en este periodo de tiempo es mayor a comparación del resto del año debido a la temporada navideña

Los tweets contenidos en la base de datos son únicos por usuario, es decir se tiene información de 1.811 usuarios.

El conjunto de datos contiene 11 variables, incluido el texto del tweet y el resto de su medata (fecha de publicación, usuario, likes, Retweets, entre otros), sin embargo, para este análisis. únicamente se utilizará la variable que contiene el texto del Tweet.

Metodología propuesta

Estandarización de información

Procesamiento de la información contenida en los Tweets con el fin de poderla estandarizar (eliminación de acentos, etc).

Se realiza el Análisis sobre la variable *Embedded_text* del conjunto de datos, ya que aquí se contienen el texto de los Tweets.

Nube de palabras

Representación visual de texto de palabras individuales (o conjuntos de palabras), donde el tamaño de las palabras representa su importancia en el texto analizado.

Se grafican nubes de palabras con unigramas y bigramas para caracterizar cada intención encontrada.

LLM de industria

Los Large Language Model son modelos diseñados para ender y entrenados con grandes conjuntos de información para entender y generar texto.

Se propone el uso de un LLM con el fin de detectar la intención de los tweets y obtener un corto resumen del mismo.

Resultados

Estandarización de información

Estandarización de información utilizando:

- Eliminación de acentos.
- Eliminación de caracteres diferentes a caracteres alfanuméricos.
- Conversión del texto a minúscula.
- Eliminación de espacios duplicados

Ejemplo original:

“La confianza se afectó. El indicador de confianza Davivienda tuvo una leve caída en Noviembre”

Procesado:

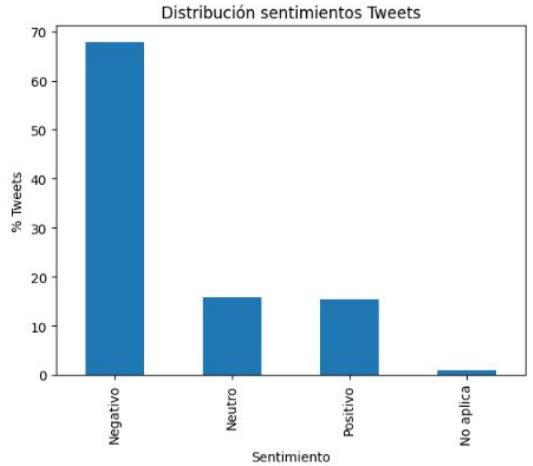
“la confianza se afecto el indicador de confianza davivienda tuvo una leve caida en noviembre”

LLM - Gemini Pro

Se utiliza el modelo de Gemini Pro con las APIs de GCP para caracterizar los sentimientos de cada Tweet.

El 70% de los Tweets son clasificados como Negativos, entendiendo que la mayoría pertenecen intenciones de reportar un fallo o hacer notar una inconformidad referente a un servicio.

Algunos Tweets no son posibles de caracterizar por el corto contenido o contenidos referentes a contenido peligroso, contenido sexual explícito o contenido de acoso



Resultados

Nube de palabras - Negativo

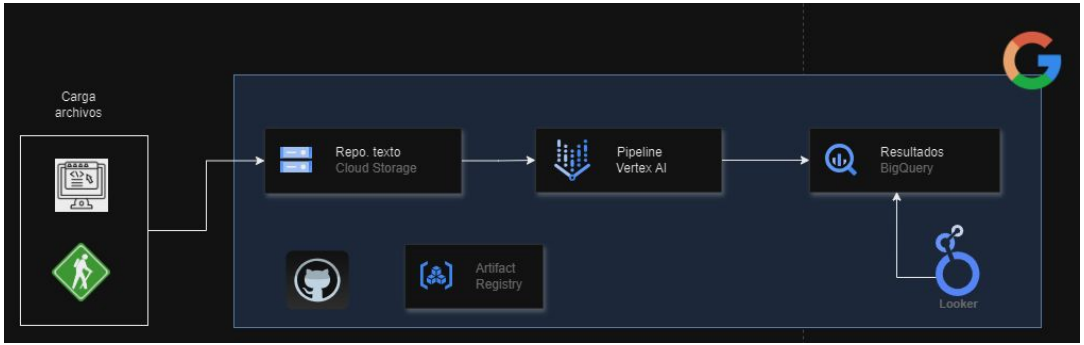
Los comentarios más recurrentes de los usuarios están relacionamos a malas experiencia, en torno a la atención y servicio, así como pérdida de dinero y problemas de la app.

Nube de palabras Positivo

Los comentarios son referentes a usuarios que tuvieron una experiencia favorable, ya que se les brindo una buena atención en los servicios solicitados. Vale la pena recordar que estos usuarios son la minoría en los Tweets analizados.

Arquitecturas de referencia

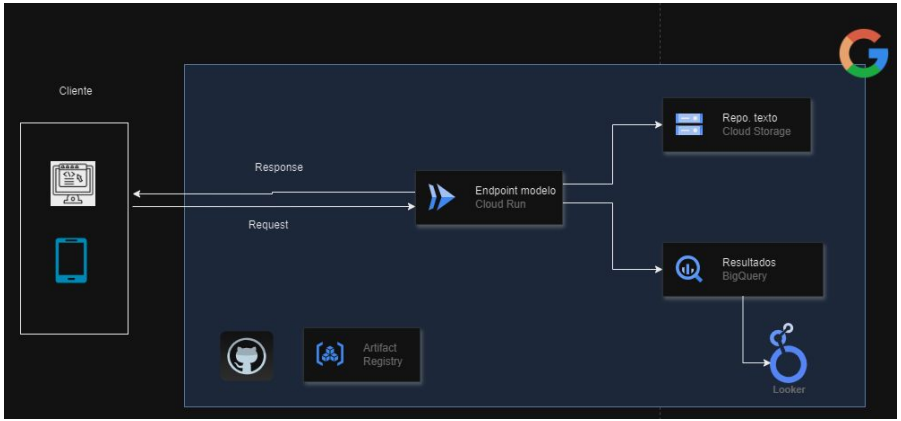
Arquitectura Batch



Las Arquitecturas son planteadas en GCP dado que el lago de Datos de Davivienda se encuentra en esta nube. Sin embargo, la solución puede ser desplegada en cualquier infraestructura (AWS, Azure, On premise. etc)

Para ambas arquitecturas propuestas se debe considerar:

- Esquema de Seguridad.
- Esquema de autenticación.
- Validación de Volumetrías



Arquitectura Real time

Retos, Consideraciones y Conclusiones

Principales retos

- La cantidad de Tweets del conjunto de datos es limitada para lograr una caracterización más profunda.
- El API de Gemini tiene una cuota de 5 llamados por minuto, fue requerido un ticket a soporte para la modificación de la misma.
- Algunos Tweets contienen contenido explícito o poca información, por lo cual no fueron sujetos a análisis.
- En los Tweets se incluye información comercial del banco, si se requiere conocer la percepción del cliente se debe generar un filtro al obtener dicha información.

Conclusiones y consideraciones

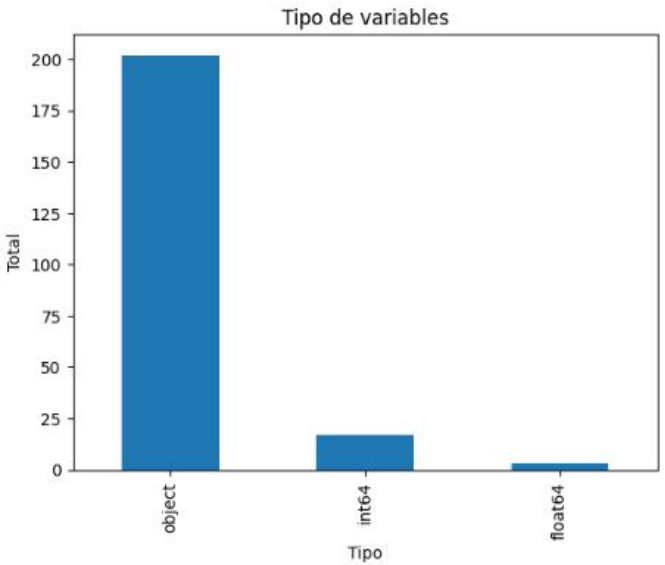
- La combinación de estandarización de información y Gemini pro, propone una solución potente y fácil de implementar, al contar con un modelo pre entrenado y disponible para el uso.
- Aproximadamente el 70% de los Tweets son comentarios negativos, lo cual es de esperar porque este tipo de redes sociales son utilizadas para reportar fallos o inconformidades con algún servicio.
- Se recomienda la obtención de una mayor cantidad de Tweets para detallar la caracterización de sus intenciones, con el fin de poder abordar las principales falencias reportadas por los usuarios.
- De ser requerida la implementación del desarrollo en un proceso masivo, se recomienda validar y ajustar las cuotas de Gemini Pro.
- La implementación del proceso se puede desarrollar en cualquier nube, únicamente se debe garantizar la comunicación al API de Gemini Pro.

Costos Vivienda

Prueba - Exploración de datos

Estimación del precio de avalúo de vivienda a partir de características de vivienda dadas.

Distribución tipo variables



Número Registros

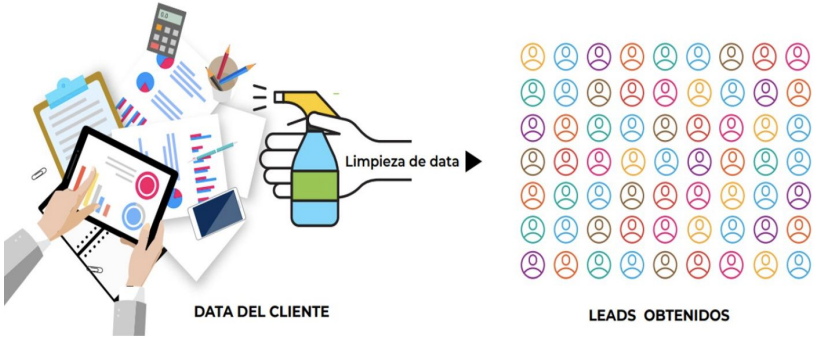
Entrenamiento	11.571
Prueba	417

Falencias en la información

- Valores Faltantes
- Inconsistencia entre naturaleza y tipo de variable.
- Datos atípicos
- Variables sin información relevante

Metodología propuesta

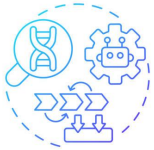
Limpieza de información



Se implementará un proceso de limpieza de información, en donde se debe evaluar completitud, coherencia, variabilidad, entre otros.

El proceso se implementa para los conjuntos de entrenamiento y test, donde se realizará eliminación de variables y registros según corresponda.

Modelo de regresión



- XGBoosT
- Red Neuronal profunda
- Support Vector machine
- Árboles de decisión de Regresión

Teniendo en cuenta que la variable respuesta es continua, se propone el ajuste de múltiples modelos de regresión, en donde se seleccionará el mejor con la métrica MAPE.

Para el alcance de este proceso, se realiza el entrenamiento de 4 modelos (XGBoosT, Red Neuronal profunda, Support Vector machine y Árboles de decisión de Regresión). Para cada modelo se utilizan diferentes hiperparametros para seleccionar el que tenga un mejor ajuste.

Resultados

Limpieza de información

Se realiza una limpieza de información so con los siguientes puntos:

- Eliminación de variables sin variabilidad.
- Eliminación de variables con mucho texto (no es el alcance del ejercicio)
- Eliminación de registros atípicos (formatos, outliers)
- Estandarización de información (normalización de texto).
- Conversión de formatos de variables (Float, object)
- Agrupación categorías con tamaño menor al 5% de la base

El tratamiento de información se hizo en conjunto para el conjunto de entrenamiento y pruebas.

Después de la limpieza de datos se tienen los siguientes tamaños:

Entrenamiento: 10.739 Registros.
Pruebas: 356 Registros.

Ambos conjuntos cuentan con 83 variables

+

Modelos de Regresión

Se crean variables Dummys a partir de variables categóricas.

Se crea una función para el cálculo del MAPE, ya que también será utilizada para la optimización de los modelos

Se realiza el entrenamiento de 4 tipos de modelos diferentes:

- Red Neuronal profunda.
- XGBoosT
- Árboles de decisión de Regresión
- Support Vector machine

El entrenamiento se genera con múltiples hiper parámetros para cada modelo.

=

Resultado

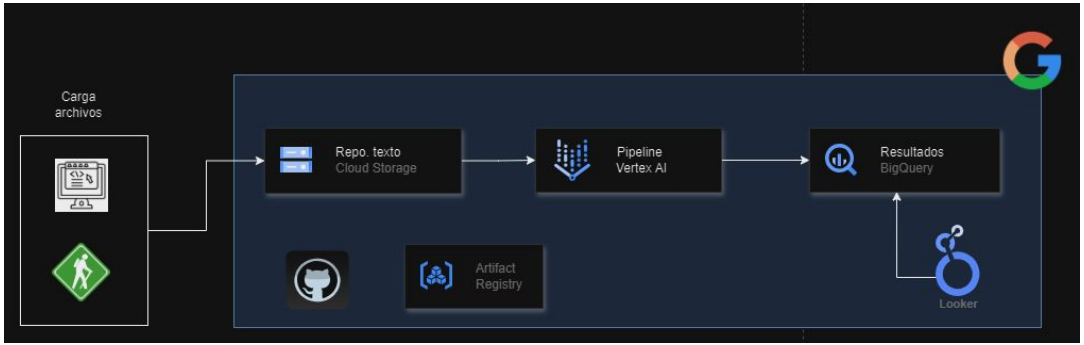
Según lo solicitado, la métrica para la comparación de modelos es el MAPE.

El modelo que mejor se ajusta a los datos y tiene un mejor desempeño es el modelo de Árboles de decisión de Regresión, dado que su MAPE es el menor, adicionalmente, este presenta un valor cercano a cero, lo cual indica que su ajuste es idóneo para este caso.

	Modelo	MAPE	MSE
0	Red Neuronal	81.184	3115252217251284123648
1	XGBoosT	342759.308	707480961170108317696
2	Árboles de decisión de Regresión	0.002	52811494924
3	Support Vector machine	550309.823	311482631693340088729 6

Arquitecturas de referencia

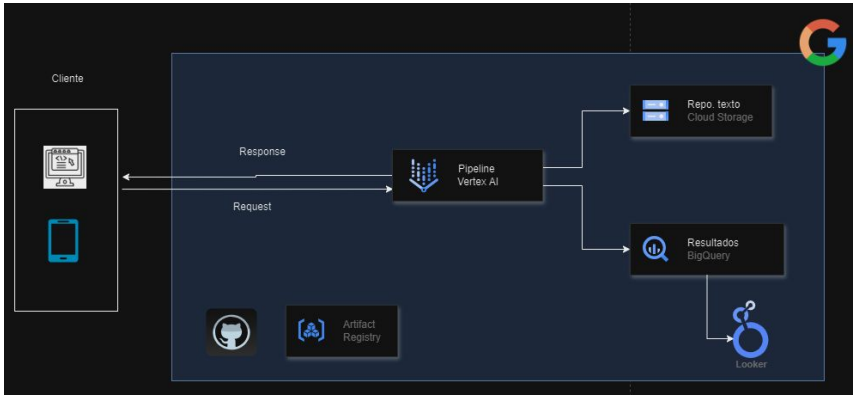
Arquitectura Batch



Las Arquitecturas son planteadas en GCP dado que el lago de Datos de Davivienda se encuentra en esta nube. Sin embargo, la solución puede ser desplegada en cualquier infraestructura (AWS, Azure, On premise. etc)

Para ambas arquitecturas propuestas se debe considerar:

- Esquema de Seguridad.
- Esquema de autenticación.
- Validación de Volumetrías



Arquitectura Real time

Retos, Consideraciones y Conclusiones

Principales retos

- El principal reto del ejercicio fue la calidad de la información, por la cual fue requerida la eliminación de variables y registros.
- La base de datos contaba con un tamaño limitado.
- Para la interpretación de valores nulos y vacíos, es requerido un diccionario que de un detalle de la variable.

Conclusiones y consideraciones

- Con las técnicas implementadas se logra tener una limpieza aceptable, sin embargo, se recomienda un proceso de estandarización en la obtención de la información para contar con completitud y coherencia desde un inicio.
- Se recomienda un entrenamiento constante del modelo con nueva información.
- Para evitar impactos externos sobre la variable objetivo, se recomienda realizar una conversión a número de salarios mínimos, teniendo en cuenta el valor de salario del año del avalúo.
- El modelo con mejor ajuste, es el modelo de Árboles de decisión de Regresión, el cual muestra un gran desempeño en el conjunto de entrenamiento, con una métrica de MAPE del 0.002.
- El comportamiento del modelo a futuro dependerá en gran medida a la calidad de la información, recordemos un buen modelo depende de buenos datos.
- La implementación del proceso se puede desarrollar en cualquier nube,

Gracias por su atención!