Entregable 1

Integrantes:

Cristian Andres Henao Londoño

Jhon Janer Torres Restrepo

Asignatura:

Introducción a la Inteligencia Artificial

Docente:

Raul Ramos Pollan

1. Este problema de predicción es relevante para la detección temprana de fraudes en aperturas de cuentas bancarias, lo que puede ayudar a las instituciones financieras a tomar medidas preventivas para proteger sus activos y la integridad de sus servicios. Además, la consideración de la equidad en el modelo es fundamental para garantizar que las decisiones de predicción sean justas y no discriminatorias en función de atributos protegidos.
El objetivo principal es desarrollar un modelo estadístico de predicción que pueda predecir si una solicitud de apertura de cuenta bancaria es fraudulenta o legítima utilizando el conjunto de datos BAF. Este problema de predicción implica clasificar las solicitudes de apertura de cuentas en dos categorías: fraudulentas o legítimas, basándose en las características proporcionadas en el conjunto de datos.

2. Vamos a usar el dataset de kaggle https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022, que contiene 1 millón de instancias y 32 columnas entre las cuales encontramos las siguientes como las mas relevantes para el trabajo de prediccion :

- **fraud_bool:** Esta columna parece ser la variable objetivo o etiqueta que indica si una solicitud de apertura de cuenta fue fraudulenta o no (1 para fraudulenta, 0 para legítima). Es la variable que deseas predecir.

- **income:** El ingreso del solicitante puede ser relevante para la detección de fraudes, ya que un ingreso inusualmente alto o bajo podría ser una señal de actividad sospechosa.

- **name_email_similarity:** Esta columna podría indicar la similitud entre el nombre del solicitante y su dirección de correo electrónico. Podría ser útil para verificar la autenticidad del solicitante.

- **prev_address_months_count:** El número de meses desde la dirección previa del solicitante podría ser relevante para detectar cambios frecuentes de dirección, lo que podría ser una señal de actividad fraudulenta.

- **current_address_months_count:** Similar a la columna anterior, el número de meses en la dirección actual del solicitante puede ser relevante para detectar cambios frecuentes de dirección.

- **customer_age:** La edad del solicitante podría ser útil para evaluar la coherencia con la información proporcionada en la solicitud.

- **days_since_request:** El número de días desde la solicitud podría ser relevante para detectar solicitudes antiguas o inusuales.

- **intended_balcon_amount:** El monto deseado para la cuenta bancaria podría ser relevante, ya que solicitudes de montos inusuales podrían ser sospechosas.

- **payment_type:** El tipo de pago utilizado en la solicitud podría proporcionar información sobre la autenticidad de la solicitud.

- **zip_count_4w:** El recuento de códigos postales en las últimas 4 semanas podría ayudar a detectar múltiples solicitudes provenientes de diferentes ubicaciones.

- **has_other_cards:** Indica si el solicitante tiene otras tarjetas de crédito. Esto podría ser relevante para evaluar la capacidad de pago y la autenticidad del solicitante.

- **proposed_credit_limit:** El límite de crédito propuesto podría ser útil para evaluar la coherencia con la información proporcionada en la solicitud.

Estas son algunas de las columnas que podrían ser relevantes para la predicción de fraudes en tu conjunto de datos.Sin embargo mediante el proyecto avance se pueden descartar algunas variables e incluir otras que se puedan considerar relevantes.

3. Métricas de Desempeño de Machine Learning:
Precisión (Accuracy): Esta métrica te indicará la proporción de solicitudes de apertura de cuentas que se han clasificado correctamente como fraudulentas o legítimas. Se busca una alta precisión, pero no debe ser la única métrica, ya que los conjuntos de datos desequilibrados pueden dar una falsa impresión de precisión.

Recall (Sensibilidad): El recall dirá cuántas de las solicitudes de apertura de cuentas fraudulentas se detectaron correctamente. Se busca obtener un alto recall para tener la seguridad de capturar la mayoría de los casos de fraude.

F1-Score: Esta métrica combina precisión y recall y es útil cuando se tiene un desequilibrio en las clases. Un alto F1-score indica un equilibrio entre la precisión y la capacidad para detectar fraudes.

Matriz de Confusión: Proporciona una visión detallada de cómo se están clasificando las solicitudes, incluyendo los falsos positivos y falsos negativos.

4. El objetivo es reducir significativamente la detección de fraudes mediante la implementación del modelo en un alto porcentaje que ronde ente el 60% y el 80%. Esto significa que se busca lograr una drástica disminución en la cantidad de fraudes detectados mediante la aplicación efectiva del modelo. La meta es que el modelo tenga un impacto sustancial en la reducción de casos fraudulentos, contribuyendo así a la protección de los activos y la integridad de los servicios financieros, al tiempo que minimiza el riesgo asociado con actividades fraudulentas en solicitudes de apertura de cuentas bancarias.